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

Supplementary Material

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

Contents

1. Experiments on OSIE and MIT1003

To further validate the effectiveness of our proposed HAT in free-viewing scanpath prediction, we compare HAT to the previous state-of-the-art method in free-viewing scanpath prediction, Chen et al. [2], using the OSIE dataset [14] and the MIT1003 dataset [7]. Here we only report SS, cIG, cNSS and cAUC and do not use SemSS because free-viewing attention is bottom-up and does not rely on semantics. Moreover, OSIE and MIT1003 do not contain pixel-wise segmentation annotation which is required in SemSS. Tab. 1 and Tab. 2 consistently show that HAT surpasses Chen et al. [2] in all metrics by a large margin especially in cIG and cNSS on both free-viewing datasets. The results are consistent with our findings in Tab. 3 of the main text—HAT accurately predicts the scanpaths (reflected by SS), with well-calibrated confidence (as evidenced by the high cIG and cNSS). Additionally, we compare HAT to the best alternative overall, Chen et al. [2], by evaluating the models trained using COCO-FreeView on an *unseen* dataset MIT1003 in Tab. 3. The results show that HAT out-

	SS	cIG	cNSS	cAUC
Human consistency	0.380	-	-	
Chen *et al.* [2]	0.326	-1.526	2.288	0.920
HAT	**0.386**	**2.434**	**4.515**	**0.973**

Table 1. **Comparing free-viewing scanpath prediction algorithms on OSIE** (rows) using multiple scanpath metrics (columns). The best results are highlighted in bold.

	SS	cIG	cNSS	cAUC
Human consistency	0.363	-	-	-
Chen *et al.* [2]	0.260	0.042	1.408	0.927
HAT	**0.364**	**1.311**	**2.966**	**0.956**

Table 2. Comparing free-viewing scanpath prediction algorithms (rows) on **MIT1003 training set using 5-fold cross validation** using multiple scanpath metrics (columns). The best results are highlighted in bold.

	SS	cIG	cNSS	cAUC
Human consistency	0.363	-	-	-
Chen *et al*. [2]	0.210	-9.735	0.186	0.750
HAT	**0.251**	**1.052**	**2.577**	**0.951**

Table 3. **Generalization to an unseen dataset MIT1003**. Both models are trained on COCO-Freeview. The best results are in bold.

performs Chen et al. [2] in all metrics and with significant improvement in cIG, cNSS and cAUC. This suggests that Chen *et al*.'s model is prone to be overconfident, whereas HAT better calibrates the confidence in predicting free-viewing fixations and thus provides a more robust prediction of human attention with better generalizability to unseen datasets.

econd ro

2. Scene-to-scene Generalization.

To further demonstrate the generalization ability of HAT to unseen scenes, we re-partition the COCO-Search18 dataset [3] by **scenes**. The new partition contains two test sets: one test set shares the scenes as the training set and the other test set only contains unseen (new) scenes from the training set. We partition COCO-Search18 for each category independently (shown in Tab. 4). For instance, for the target-absent microwave search task, the training set only contains kitchen scenes while the unseen test set has a variety of other scenes including living rooms, dining rooms, bedrooms and outdoor scenes. To further ensure the unseen images do not exist in the training set, we remove the unseen images for all tasks from the training set because some tasks share the same image stimuli. In the new partition, the target-present set consists of 2170 training images, 568 testing images of unseen scenes, and 273 testing images of seen scenes. The target-absent set consists of 2299 training images, 378 testing images of unseen scenes, and 282 testing images of seen scenes.

Tab. 5 presents results of HAT in predicting the TP and TA search scanpaths under both seen and unseen novel scenes. We use human consistency as the baseline. For both tasks, the gap between human consistency and HAT in seen test set is smaller than that in the unseen test set, which is expected. Importantly, HAT's performance on the unseen scenes is on par with human consistency in TA setting although worse in TP setting. Note that the performance difference between seen and unseen human consistency of TA setting is due to the fact that human consistency on the TA test data with seen scenes in the new partition is low. These results suggest that HAT learns to extrapolate from scene to scene and generalize well on novel scenes in visual search scanpath prediction. The visualization of predicted scanpaths in Fig. 1 further reinforces our observations. By comparing HAT's predicted scanpaths with the ground-truth human scanpaths for unseen scenes in both TP and

	Target-present		Target-absent	
Target	Seen	Unseen	Seen	Unseen
Bottle	Others	Food	Others	Kitchen
Bowl	Others	Kitchen	Others	Kitchen
Car	Indoor	Outdoor	Vehicle	Others
Chair	Others	Kitchen	Indoor	Outdoor
Clock	Others	Building	Others	Office
Cup	Others	Office	Food	Others
Fork	-	-	Others	Bathroom
Keyboard	Office	Others	Others	Bedroom
Knife	Food	Others	Others	Bathroom
Laptop	Others	Living	Others	Living
Microwave	Kitchen	Others	Kitchen	Others
Mouse	Office	Others	Others	Office
Oven	-	-	Others	Living
Potted plant	Indoor	Outdoor	Others	Food
Sink	Others	Kitchen	Indoor	Outdoor
Stop sign	-	-	Others	Vehicle
Toilet	Indoor	Outdoor	Others	Bedroom
TV	Others	Office	Indoor	Outdoor

Table 4. **Data split for scene-to-scene training and testing.** COCO-Search-18 includes 12 scenes: outdoor, street, building, vehicle, food, eatery, kitchen, bathroom, bedroom, living-room (living), dining-room, office. **Others** in the table includes some of the scenes excluding the unseen scene. **Outdoor** in the table includes outdoor, street, building and vehicle. In target-present, fork, oven and stop sign are not splittable because they only contain one scene, so we remove them from testing.

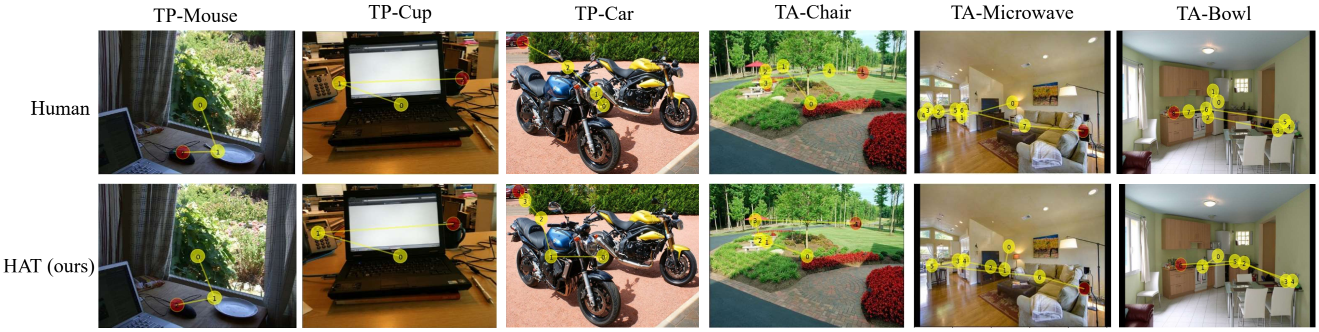


Figure 1. **Visual search scanpath visualization for unseen scenes predictions.** The first row is human consistency and the second row is test result on unseen scenes. The first three columns are trained for target-present tasks for mouse, cup and car search, the second three columns are trained for target-absent tasks for chair, microwave and bowl search.

		SS	cIG	cNSS	cAUC
TP	Human(Seen)	0.520	-	-	
	HAT(Seen)	0.499	2.074	5.032	0.976
	Human(Unseen)	0.546	-	-	
	HAT(Unseen)	0.481	2.454	4.537	0.973
TA	Human(Seen)	0.364	-	-	
	HAT(Seen)	0.368	1.586	2.852	0.955
	Human(Unseen)	0.416	-	-	
	HAT(Unseen)	0.416	2.075	3.115	0.962

Table 5. Quantitative result of scene-to-scene generalization on target-present and target-absent task. The first four columns are analysis of scene-to-scene target-present search, and the last four columns are analysis of scene-to-scene target-absent search. Each search contains human consistency and testing results on seen (first two columns of the search) and unseen scenes (last two columns of the search).

TA settings, we unveil HAT's robust generalization. For example, in a TP car search task, HAT trained on *indoor* scenes successfully located the car on the top-left corner in an *outdoor* scene much like humans do. Similarly, when addressing microwave searches under target-absent conditions—with the training set exclusively comprising kitchen scenes—HAT demonstrates significant generalization. This is evident in its proficient extension of predictive capabilities to living-room scenes, as showcased in the fifth column of Fig. 1. These findings underscore HAT's consistent and robust generalization across diverse scenes, emphasizing its reliable performance in a spectrum of visual search scenarios.

3. Individual Scanpath Recall

The importance of predicting personalized scanpath lies in the fact that each person's unique life experiences shape their individual mental representations of scenes, resulting in personalized perceptions. Therefore, testing the model's ability to generate diversified scanpaths is crucial to learn individual perceptions and to avoid potential biases. To this end, we compute *scanpath recall* to measure the extent of individual representation within the model's predictions. For a human scanpath in the stimuli, we consider it to be covered if its sequence score with at least one prediction is higher than the threshold τ. The ratio of covered human scanpaths to all human scanpaths is the recall of the stimuli. Tab. 6 presents the average recall and sequence score of HAT and Gazeformer in both target-present and target-absent search scanpath prediction. For each visual stimuli, we sample 10 scanpaths from the fixation density map and set τ to the human consistency of sequence score (i.e., $\tau = 0.5$ for TP and $\tau = 0.381$ for TA). We can see that HAT outperforms Gazeformer in both recall and sequence score by a large margin in target-absent scanpath prediction. This is consistent with our findings in Sec. 4.1 of the main text. Although Gazeformer has a slightly higher sequence score than HAT in TP setting, HAT outperforms Gazeformer significantly in scanpath recall. This implies that HAT better captures the entire scanpath distribution from multiple subjects whereas

	Target-present		Target-absent	
	Recall	SS	Recall	SS
Gazeformer [12]	0.563	**0.489**	0.428	0.357
HAT	**0.727**	0.453	**0.750**	**0.381**

Table 6. Recall and sequence score comparison between Gazeformer and HAT.

Pixel enc.	Pixel dec.	SemSS	SS	cIG	cNSS	cAUC
R50	MSD	0.382	0.402	1.686	3.103	0.961
R50	FPN	0.367	0.388	1.582	2.908	0.958
R101	MSD	0.372	0.397	1.598	2.998	0.961
Swin-B	MSD	0.382	0.405	1.645	3.103	0.962

Table 7. **Comparing different pixel encoder and pixel decoder** in HAT. The ablation experiments are done on the target-absent set of COCO-Search18.

heads	α	β	SemSS	SS	cIG	cNSS	cAUC
4	2	4	**0.382**	**0.402**	**1.686**	**3.103**	0.961
8	2	4	0.375	0.390	1.310	2.826	0.961
4	2	2	0.381	0.401	1.129	2.633	0.960
4	1	4	0.378	0.393	1.566	3.046	**0.962**

Table 8. Hyperparameters ablation using COCO-Search18 TA set.

	Target-present	Target-absent
Dense	**0.470**	**0.403**
Regression	0.452	0.330

Table 9. Comparison between HAT's dense prediction paradigm and Gazeformer's regression paradigm on COCO-Search18 using HAT's architecture.

Gazeformer tends to overfit to an "average person", thereby repeatedly sampling similar scanpaths given the same image input.

4. Additional Ablation Study

In this section, we provide further ablation on HAT. First we ablate the backbones of HAT. We perform the ablation experiments using the target-absent (TA) visual search fixation prediction task on the TA set of COCO-Search18. By default, HAT uses ResNet-50 [6] as the pixel encoder and MSD [17] as the pixel decoder. However, HAT is also compatible with other architectures. Hence, in Tab. 7, we evaluate HAT with different pixel encoders and decoders. Three pixel encoders: ResNet-50 (R50), ResNet-101 (R101) [6] and Swin Transformer [10] (we use the base model, Swin-B); and two pixel decoders: FPN [9] and MSD [17], are evaluated. One can observe that MSD is better than FPN as the pixel decoder and HAT performs the best when using R50 and Swin-B as the pixel encoder. Notice that the performance gap between different pixel encoders is small, suggesting that the performance of HAT is robust to the choice of different pixel encoder architectures. More importantly, all of these configurations of HAT significantly outperforms all baselines in Tab. 2 of the main text.

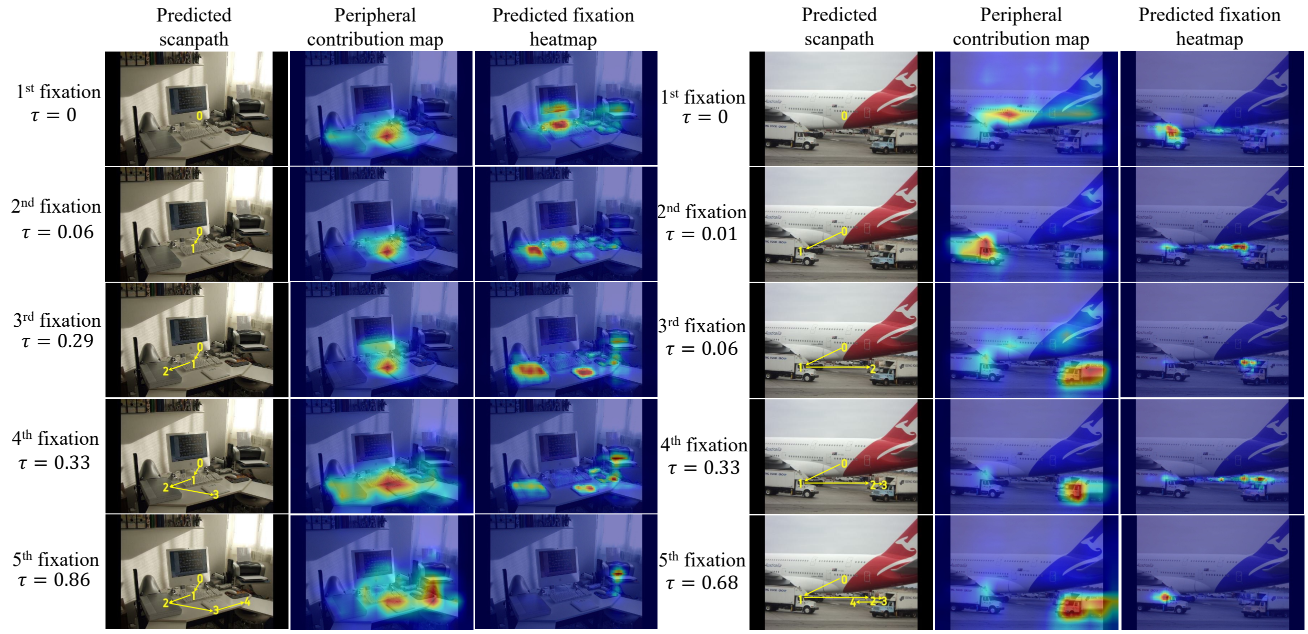
In Tab. 8, we also present HAT's results with varied hyperparameters: the number of attention heads in the transformer module of HAT, α and β of (**??**), demonstrating HAT's robustness w.r.t. difference choices of hyperparameters. Notably, the choice of $(4, 2, 4)$ in the three ablated hyperparameters achieves the best performance.

Tab. 9 compares HAT's DP task with Gazeformer's Reg task using HAT in TP and TA settings. The proposed DP outperforms Reg in both settings, especially in TA setting. This aligns with our findings in Tab. 1-3 of the main text which show that Gazeformer's Reg paradigm, assuming a Gaussian fixation distribution, is less effective for TA and FV scanpaths.

5. Additional Qualitative Analysis

5.1. Model interpretability

Peripheral contribution map visualization. In Sec. 4.2 of the main text, we showed that the peripheral contribution maps in HAT can be leveraged to interpret the model's behaviors using a target-present search example. We also observe a similar pattern in the target-absent (TA) setting (see Fig. 2). In Fig. 2a, we see that


Predicted scanpath
Peripheral contribution map
Predicted fixation heatmap
1st fixation $\tau = 0$
2nd fixation $\tau = 0.06$
3rd fixation $\tau = 0.29$
4th fixation $\tau = 0.33$
5th fixation $\tau = 0.86$
Predicted scanpath
Peripheral contribution map
Predicted fixation heatmap
1st fixation $\tau = 0$
2nd fixation $\tau = 0.01$
3rd fixation $\tau = 0.06$
4th fixation $\tau = 0.33$
5th fixation $\tau = 0.68$

(a) Target-absent laptop search

(b) Target-absent car search

Figure 2. Visualization of the **predicted scanpath, peripheral contribution map and fixation heatmap** (columns) of HAT for target-absent (a) laptop and (b) car visual search examples at every fixation (rows). We also include the predicted termination probability τ for each step on the left. The model terminates searching if $\tau > 0.5$.

in a TA laptop search task pixels of *table* and *keyboard* contribute significantly in predicting human fixations as tables and keyboards can provide spatial cues for the laptop. This reveals a unique factor that guides visual search attention—anchor objects [1]. In Fig. 2b, we see that TA car search fixations are attracted to truck pixels as trucks and cars are closely related concepts that are considered as distractors.

Peripheral vs foveal. We also *collectively* analyze the contribution of peripheral tokens and foveal tokens in predicting human attention control under the TP, TA and FV settings, separately. Fig. 3 visualizes the temporal change of contributions of all peripheral tokens collectively and the foveal token in predicting human attention averaged over all test images. We observe that the peripheral tokens contribute the most in predicting TP fixations across all fixations (forming the yellow column on the left). This is because in TP images there is a strong target signal available in the visual periphery to guide attention. Contrast this with FV fixations, where the contribution of the peripheral tokens diminishes over the temporal space and the only the current foveal token has a strong and consistent contribution (a clear red diagonal line). An interpretation of this pattern is that people have only a poor memory of what they viewed in previous fixations and their attention is controlled by salient pixels within a local neighborhood around the current fixation. Interestingly, for TA fixations we also observe a diminishing contribution of the peripheral tokens over the temporal space, but not as pronounced. Moreover, as more fixations are made, the contribution of recent fixations increases, approaching the pattern in FV. This suggests that the later fixations of a TA scanpath behave like a FV scanpath, which confirms a finding in [4]. Lastly, the bottom row visualizes the contribution of each individual peripheral token (averaged over the temporal axis), where we see peripheral tokens encode a strong center bias for FV fixations, whereas TA fixations show a weaker center bias and TP fixations show no obvious center bias at all, again as expected and confirming previous suggestion. This showcases the potential for HAT to make highly interpretable predictions of human attention control.

Target prior. A natural question arising from this observation is whether the peripheral tokens of TA and

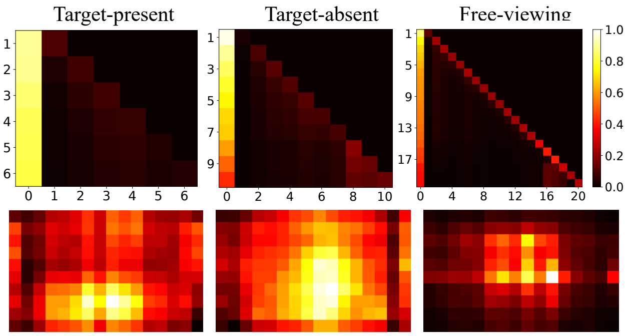


Figure 3. **Peripheral tokens vs foveal tokens** under TP, TA and FV settings (from left to right). The top three figures visualize the temporal change of the contribution of peripheral and foveal memory tokens in predicting human attention. Here the contribution is measured by the attention weight from the last cross-attention layer of the aggregation module in HAT. X-axis shows the token index, with 0 representing all peripheral tokens (by summing the attention weights of all peripheral tokens) and $i > 0$ being the i-th foveal token. Y-axis indicates temporal fixation step from first to max number of fixation steps allowed for each task. The bottom three figures show the spatial distribution of the attention weights of all peripheral tokens, averaged over the temporal axis. The brighter the color, the larger is the contribution.

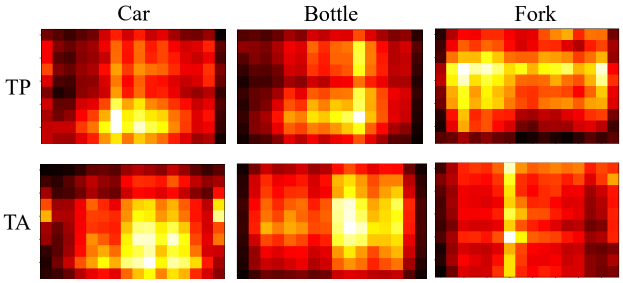


Figure 4. **Categorical peripheral contribution map of visual search fixations**. We show the contribution map of the peripheral tokens for two categories (rows): car and bottle, in target-present and target-absent settings (columns). We measure the contribution of each peripheral token by the attention weights from the last cross-attention layer of the aggregation module in HAT, averaged over the temporal axis of all testing data in COCO-Search18 [3]. The brighter the color, the larger the contribution.

TP fixations encode a *target prior*–spatial distribution of the possible target location. To answer this question, we visualize the category-wise peripheral contribution maps for TP and TA fixations by averaging the attention weights (on the peripheral tokens) of the last cross-attention layer over all testing fixations for each target category. As shown in Fig. 4, the category-specific peripheral contribution map does not provide a clear evidence of TA and TP peripheral contribution map being a target prior, but we find some target-specific pattern, e.g., the contribution is pronounced around the bottom horizontal area for "car" and around the vertical area for "bottle", which may represent the spatial prior of each category.

5.2. Failure cases analysis

Our analysis of failure cases offers insights for future research. A scanpath prediction would be taken as a failure case if its sequence score falls below 50% of the human consistency of its stimulus. Under this criterion, we find some common features of failure cases. For target-present, the ambiguity of the target object often leads to a decline in HAT performance. For instance, in the first row of Fig. 5, the laptop in the first case has a very similar color to the table and its surrounding objects. In the second case, the TV is indistinguishable even to an individual. Both scenarios present an ambiguous visual representation of the target, complicating the prediction of the scanpath during visual searches. For target-absent, it is hard for HAT to learn the perception pattern of human when the human scanpaths are very short. In free-viewing, from the visualization in the third row of Fig. 5, HAT only allocates a few fixations to text in the image, which is opposite to human perception. This discrepancy is attributed to the limitations of the image encoder and decoder in capturing text features.

5.3. Scanpath visualization

We further visualize additional scanpaths for human (ground truth), our HAT, Gazeformer [12], FFMs [16], Chen *et al*. [2], IRL [15], and a heuristic method (target detector for visual search and saliency heuristic for free viewing) in the TP, TA, and FV settings. Fig. 6 shows the TP scanpaths. In all examples, HAT shows superior performance in predicting the human fixation trajectory not only when humans correctly fixate on the target, but also when their attention is distracted by other visually similar objects. For example, in the last column of Fig. 6 when the task is to find a knife, HAT is the only model that correctly predicts the fixation on

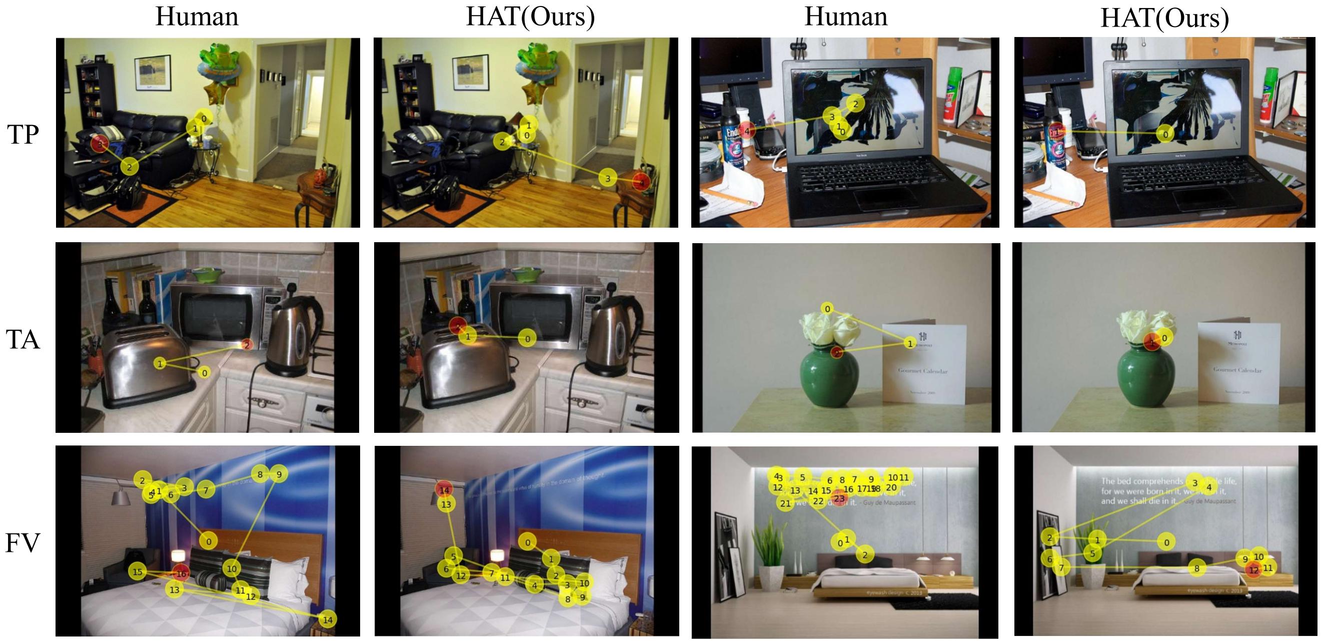


Figure 5. **Failure cases.** The first row is two failure cases for laptop and tv search, respectively. The second row is two failure cases for sink and clock search, respectively. The third row is two failure cases for free viewing.

the metallic object (because knives are usually metallic), whereas other methods either missed the target or did not show any distractions to the metallic object. This shows the capacity of HAT in modeling human attention control in visual search. Fig. 7 shows that HAT learns to leverage the context cues in predicting target-absent fixations, e.g., when the search target is microwave, HAT correctly predicted the fixations on the counter-top and table, where microwaves are often found. Similarly, HAT also generates the most human-like scanpaths in free-viewing task (see Fig. 8), capturing all important aspects of scanpaths, such as the locations (where), the semantics (what), and the order (when) of the fixations.

6. Implementation details

6.1. Network structure.

HAT has four modules as shown in Fig. 2 of the main text. By default, the feature extraction module employs ResNet-50 [6] as the pixel encoder and MSDeformAttn [17] as the pixel decoder. Sec. 4 presents the results with other pixel encoders, ResNet-101 and Swin Transformer [10], and pixel decoder, FPN [9]. The number of channels of the feature maps C is set to 256. For the foveation module, the transformer encoder has three layers. The transformer decoder in the aggregation module has six layers (i.e,. $L = 6$). All transformer encoder and decoder layers in HAT have 4 attention heads. The number of queries $N = 18$ for visual search scanpath prediction as COCO-Search18 contains 18 target categories and $N = 1$ for free-viewing scanpath prediction. Finally, the MLP in the fixation prediction module has two linear layers with 512 hidden dimensions and a ReLU activation function.

6.2. Training settings.

Following [15, 16], we resize all images to 320×512 for computational efficiency during training and inference. We use the AdamW [11] with the learning rate of 0.0001 and train HAT for 30 epochs with a batch size of 32. No data augmentation is used during training. Note that we keep the pixel encoder fixed during training and we use the COCO-pretrained weights for panoptic segmentation from [5] as an initialization for the pixel encoder and pixel decoder. Following [15], we set the maximum length of each predicted scanpath to 6 and 10 (excluding the initial fixation) for target-present and target-absent search scanpath prediction,

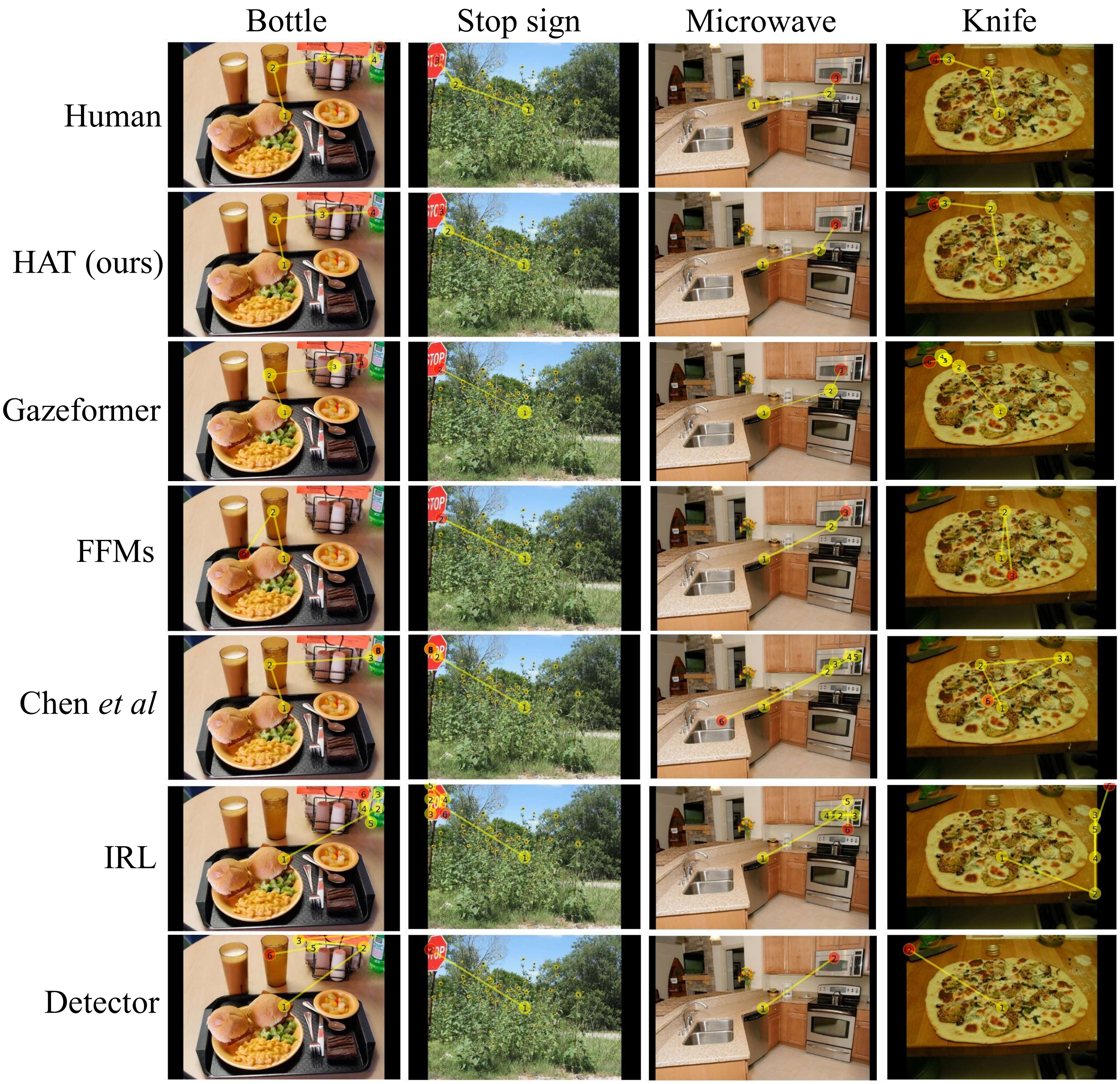


Figure 6. **Target-present scanpath visualization**. We show the scanpaths of seven methods (rows) for four different targets (columns) which are bottle, stop sign, microwave and knife. The final fixation of each scanpath is highlighted in red circle. For methods without termination prediction, i.e., IRL and detector, we visualize the first 6 fixations.

respectively. For free viewing, the maximum scanpath length is set to 20.

6.3. Additional details on heuristic baselines

Detector: The detector network consists of a feature pyramid network (FPN) for feature extraction (1024 channels) with a ResNet50 pretrained on ImageNet as the backbone and two convolution layers with batch normalization and a ReLU activation layer in between for detection of 18 different targets. The kernel size and hidden dimension of the first convolutional layer is 3 and 128, respectively. The detector network predicts a 2D spatial probability map of the target from the image input and is trained using the ground-truth

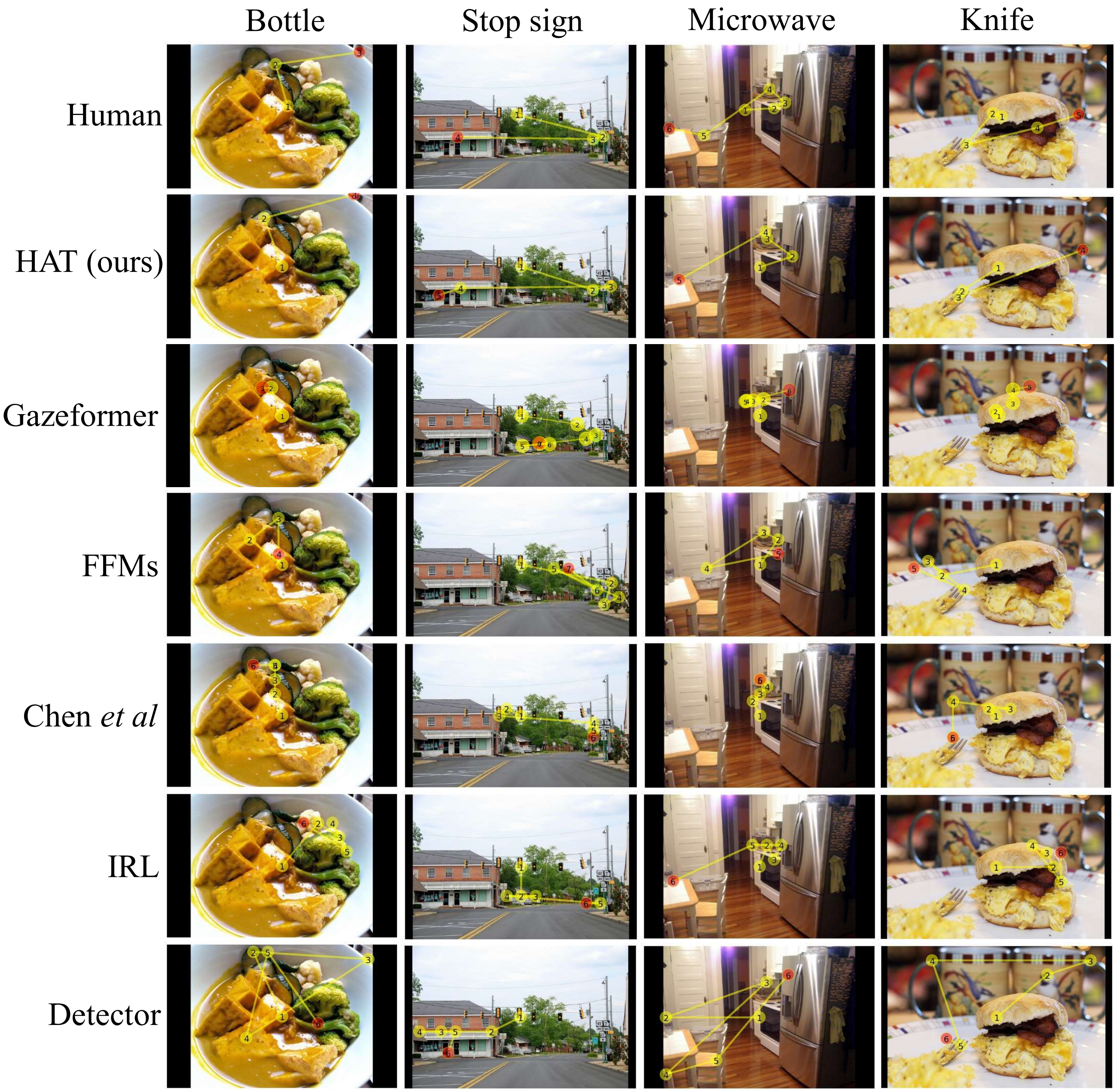


Figure 7. **Target-absent scanpath visualization**. We show the scanpaths of seven methods (rows) for four different targets (columns) which are bottle, stop sign, microwave and knife. The final fixation of each scanpath is highlighted in red circle. For methods without termination prediction, i.e., IRL and detector, we visualize the first 6 fixations.

location of the target. Another similar baseline is **Fixation Heuristic**. This network shares exactly the same network architecture with the detector baseline but it is trained with behavioral fixations in the form of spatial fixation density map, which is generated from 10 subjects on the training images.

6.4. Scanpath generation

Most methods except human consistency generate a new spatial priority map or action map at every step, while the predicted priority map is fixed over all steps for the Detector, Fixation Heuristic and IVSN baselines. Prior works like [15] measure model perfor-

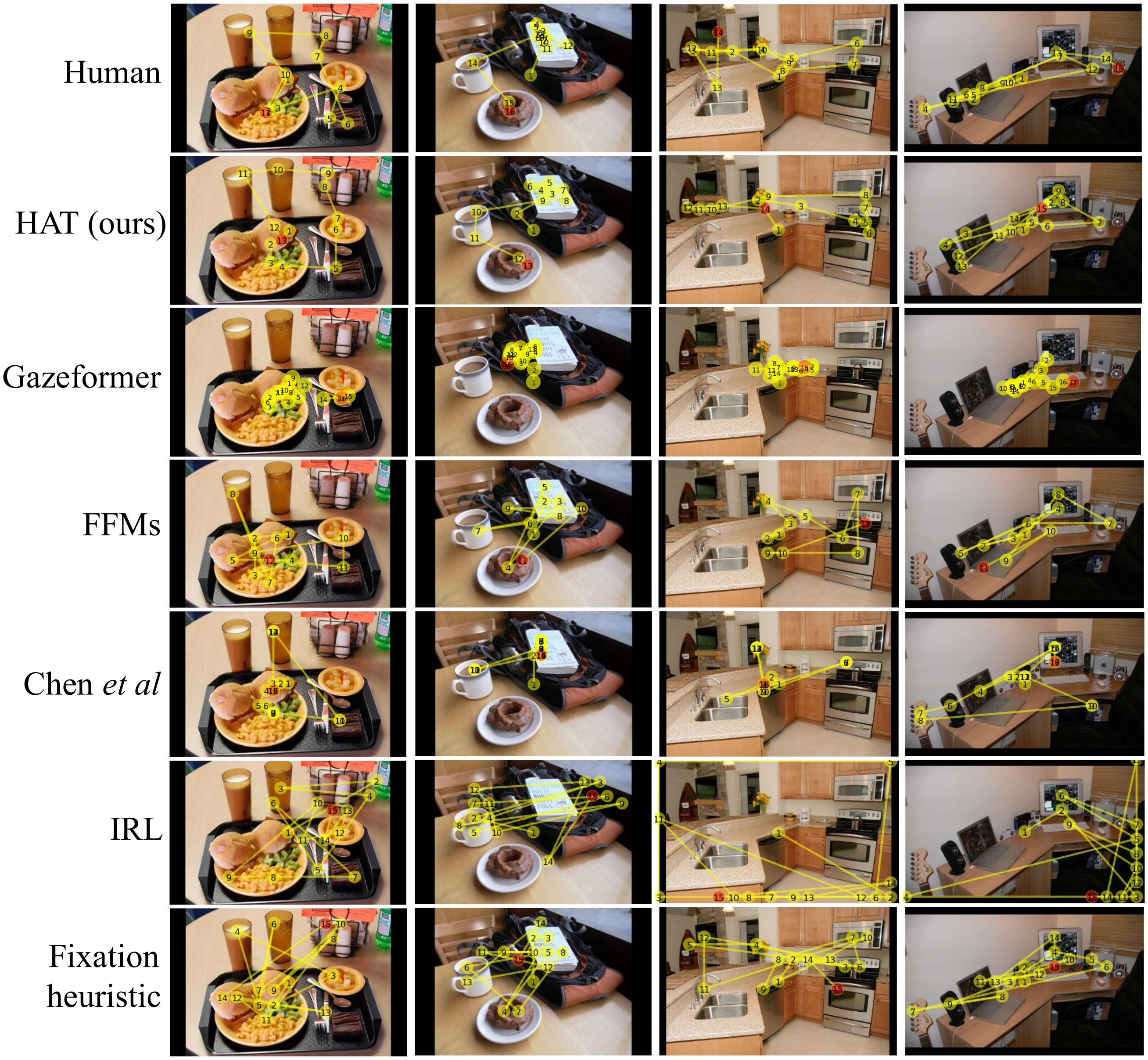


Figure 8. **Free-viewing scanpath visualization**. We show the scanpaths of seven methods (rows) for four example images. The final fixation of each scanpath is highlighted in red circle. For methods without termination prediction, i.e., IRL and detector, we visualize the first 15 fixations.

mance based on multiple randomly sampled scanpaths, which, however, can be unfairly bias toward models that repeatedly sample the best (greedy) scanpath. Therefore, in this work we directly compare different methods using their best predictions. When generating scanpaths, for all methods we follow [16] and predict one scanpath for each testing image in a greedy fashion—a fixation location is determined by selecting the most probable fixation location in the predicted priority map. At evaluation, we compare the predicted "greedy" scanpath against all GT scanpaths which helps measure how well a model (at its best) captures human scanpath consistency.

6.5. Implementation of cIG

cIG denotes the amount of information gain from the predicted fixation map (the model is provided with all previous fixations) over a baseline in predicting the ground-truth fixation. Here, the baseline is a fixation density map constructed by averaging the smoothed density (with a Gaussian kernel of one degree of visual angle) maps of all training fixations. For target-present and target-absent visual search settings, we use a (target) category-wise fixation density map, following [16]. For the heuristic models (i.e., target detector and saliency heuristic) which apply the winner-take-all strategy on a static fixation map to generate the scanpath prediction, we use the same static fixation map for all fixations in a scanpath to compute cIG, cNSS and cAUC. To obtain the predicted fixation maps for Chen *et al*.'s model [2], we use the ground-truth fixation map (Gaussian smoothed with a kernel size of 2) as input to obtain the predicted action map for the next fixation (i.e., the predicted fixation map). Note that all predicted fixation maps in computing cIG, cNSS and cAUC, are resized to 320×512 for fair comparison.

7. A note on human consistency

For the same image, there are multiple ground-truth scanpaths from different human subjects. As in [15, 16], for each image human consistency is computed by averaging the similarities of every pair of human scanpaths. A model's performance, however, is measured by the average similarity of the predicted "mean" scanpath to every human scanpath. Consider scanpaths as 2D points whose similarity can be measured by Euclidean distance. The average pairwise similarity between these points (human consistency) is smaller than the average similarity of these points to their arithmetic mean (model performance). This explains how it is possible for a good model to exceed the human consistency. Taking a triangle as analogy: the average distance of a point (predicted scanpath) within the triangle to all vertices can be smaller than the average edge length (human consistency).

8. Further discussion on applications

Models that predict top-down attention (TP/TA search fixations), modulated by an external goal, have wide applicability to attention-centric HCI. For example, faster attention-based rendering that leverages the prediction of a user's attention as they play a VR/AR game and home robots incorporating search-fixation-prediction models will be better at inferring a user's need (i.e., their search target). Home robots incorporating search-fixation-prediction models will be better able to infer a users' need (i.e., their search target) and autonomous driving systems can attend to image input like an undistracted driving expert. Applications of FV attention prediction exist in foveated rendering [8] and online video streaming [13].

References

[1] Sage EP Boettcher, Dejan Draschkow, Eric Dienhart, and Melissa L-H Võ. Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of vision*, 18(13):11–11, 2018. 5

[2] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *CVPR*, 2021. 1, 2, 6, 11

[3] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):1–11, 2021. 2, 6

[4] Yupei Chen, Zhibo Yang, Souradeep Chakraborty, Sounak Mondal, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Characterizing target-absent human attention. In *CVPR Workshops*, 2022. 5

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 7

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 7

[7] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009. 1

[8] Anton S Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. 11

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4, 7

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4, 7

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 7

[12] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *CVPR*, 2023. 4, 6

[13] Sohee Park, Arani Bhattacharya, Zhibo Yang, Samir R Das, and Dimitris Samaras. Mosaic: Advancing user quality of experience in 360-degree video streaming with machine learning. *IEEE Transactions on Network and Service Management*, 18(1):1000–1015, 2021. 11

[14] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 1

[15] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *CVPR*, 2020. 6, 7, 9, 11

[16] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *ECCV*, 2022. 6, 7, 10, 11

[17] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 4, 7