Predicting Human Scanpaths in Visual Question Answering

Xianyu Chen Ming Jiang Qi Zhao
University of Minnesota
{chen6582, mjiang}@umn.edu, qzhao@cs.umn.edu

Abstract

Attention has been an important mechanism for both humans and computer vision systems. While state-of-the-art models to predict attention focus on estimating a static probabilistic saliency map with free-viewing behavior, real-life scenarios are filled with tasks of varying types and complexities, and visual exploration is a temporal process that contributes to task performance. To bridge the gap, we conduct a first study to understand and predict the temporal sequences of eye fixations (a.k.a. scanpaths) during performing general tasks, and examine how scanpaths affect task performance. We present a new deep reinforcement learning method to predict scanpaths leading to different performances in visual question answering. Conditioned on a task guidance map, the proposed model learns question-specific attention patterns to generate scanpaths. It addresses the exposure bias in scanpath prediction with self-critical sequence training and designs a Consistency-Divergence loss to generate distinguishable scanpaths between correct and incorrect answers. The proposed model not only accurately predicts the spatio-temporal patterns of human behavior in visual question answering, such as fixation position, duration, and order, but also generalizes to free-viewing and visual search tasks, achieving human-level performance in all tasks and significantly outperforming the state of the art.

1. Introduction

Visual attention plays an essential role in everyday tasks. While existing works focus on stimulus-driven attention with free-viewing behavior, underlying daily tasks is another form of attention, *i.e.*, task-driven attention, that selects task-relevant information to make a decision or to accomplish a task. Besides, beyond the static saliency map that highlights the relative importance of a visual input, temporal sequences of eye fixations encode a more comprehensive and natural representation of attention. Understanding and predicting visual scanpaths in general tasks will not only shed light on the decision-making process but also be a useful tool for a variety of computer vision applications.

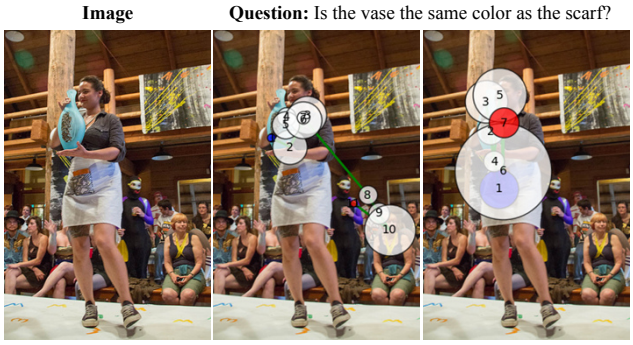

Figure 1. Visual scanpaths of humans can reveal their decision-making strategies and explain their performance. Those who pay attention to relevant visual cues can achieve high levels of task performance. This example compares the scanpaths of people who succeed or fail to answer a question, where the dots represent fixations. The number and radius indicate the fixation order and duration, respectively. The blue and red dots indicate the beginning and the end of the scanpath, respectively.

Task-driven visual scanpaths reflect the visual exploration to accomplish the task, which also strongly correlates with task performance. As an example (Fig. 1), to answer the question "Is the vase the same color as the scarf?" while exploring the scene, humans need to actively explore the scene and search for the vase and the scarf. While looking at the right places at the right time would usually lead to correct answers (Fig. 1, middle), failing to do so may result in incorrect answers (Fig. 1, right).

As a step toward understanding and modeling general task-driven attention, we propose a novel deep reinforcement learning method leveraging task guidance as an important modality to predict the visual exploration behavior of humans performing general tasks. We first introduce a task guidance map to specify task-relevant image regions. The map is designed and demonstrated to generalize across tasks. To address the exposure bias that arises between training- and test-time contexts, we introduce a reinforcement learning method that directly optimizes non-differentiable test-time evaluation metrics [14]. To differentiate eye-movement patterns that lead to different perfor-

mances, we further introduce a novel loss function to account for the consistency and divergence between correct and incorrect scanpaths.

Our work has three distinctions from previous scanpath prediction studies: (1) While state-of-the-art scanpath prediction studies focus on free-viewing [4, 5, 13, 40] or well-structured tasks such as visual search [52], this paper for the first time studies the complex scanpath patterns in general decision-making tasks, and investigates the correlation of scanpaths and performances in this context. (2) Scanpath prediction has not been as popular (compared with saliency prediction) or achieved excellent performance (compared with humans), partly due to the exposure bias – the discrepancy between training-time and test-time contexts. Here we close the gap using self-critical sequence training in the reinforcement learning method, leading to significantly boosted performance that is better than humans. (3) We go beyond a single task and design a new mechanism to encode general task-relevant information that is easily adaptable to other tasks with varying nature and levels of complexity. The proposed method has been demonstrated by three tasks with human-level performance.

In sum, this work makes the following contributions:

1. We develop a deep reinforcement learning model to understand and predict scanpaths in the general task-driven context with visual question answering (VQA). Task performance is for the first time taken into account to predict scanpaths.
2. We propose to explicitly integrate attention maps from task-specific deep neural network models, allowing the encoding of task-relevant information as well as providing an alternative to measure the interpretability of task-specific models through analyzing model *vs.* human attention.
3. To address the discrepancy between training and testing that may have limited the development of scanpath prediction methods, we apply self-critical sequence training to directly optimize non-differentiable evaluation metrics. We further introduce a novel loss function to learn discriminative features and differentiate correct and incorrect scanpaths.
4. The proposed method significantly outperforms the state-of-the-art and shows human-level performance on three tasks: VQA, free-viewing, and visual search, demonstrating the generalizability of the method.

2. Related Work

Scanpath prediction. To precisely predict where humans look is not trivial, as eye movements are governed by several confounding factors [9]. Existing attention models either generate a saliency map where fixations can be sampled based on probability distribution and a winner-take-all strategy [11, 24, 25, 26, 45], or predict a sequence of fixations by modeling their spatio-temporal complexity [4, 5, 8, 13, 22, 31, 34, 40, 41, 43, 46, 47, 49]. Our work is mostly related to the recent studies of task-driven attention [52]. Instead of studying structured vision tasks such as visual search [52], we aim to address a broader scope of general tasks. We use VQA as an example due to its generality and complexity, while further demonstrating the generalizability and flexibility of our method by adapting it for other tasks with various levels of complexity. To the best of our knowledge, our method is the first scanpath prediction method that successfully predicts human eye-movement behavior in the VQA task, and we further take the correctness of answers into account. Our model not only approaches human-level accuracy in the VQA task but is also highly generalizable across different tasks and datasets.

Human and machine attention in VQA. A unique characteristic of our work is the explicit integration of machine attention in the prediction of human scanpaths. With the rapid development of deep neural networks, the attention mechanism has become an essential component for improving the performance and explainability of VQA models [12, 28, 44]. However, due to their intrinsic differences, machine attention disagrees with human attention in many cases [44]. To study the relationship between human attention and machine attention, Chen *et al.* [12] and Jiang *et al.* [28] have developed datasets and computational methods to measure, model, and comparatively analyze the attention maps of humans and VQA models. While these analyses focus on the spatial difference of attention between correct and incorrect answers, our method generates individual fixations to study how people *maintain* and *shift* their attention which also encodes temporal information such as durations and orders. With the explicit incorporation of machine attention, our method also provides an alternative to measure the interpretability of VQA models based on their effectiveness in guiding scanpath prediction.

Reinforcement learning in attention prediction. A plausible approach to human attention prediction is reinforcement learning [27, 35, 36]. Early studies consider selective attention as a Markov decision process [6, 42] that can be optimized using policy iteration and a predefined reward function [27, 35, 36]. Recent scanpath prediction methods [33, 51, 52] adopt inverse reinforcement learning [1, 3] to automatically learn the unknown reward function from humans' eye-movement behavior. Although these methods are promising, there is still a significant performance gap between scanpath prediction models and humans. We hypothesize that the performance gap is mainly caused by the exposure bias that commonly exists in sequence prediction tasks [38]. Exposure bias indicates the contextual discrepancy between the training and test settings. In scanpath prediction studies, many evaluation metrics are based on non-differentiable sequence comparison algorithms. Thus

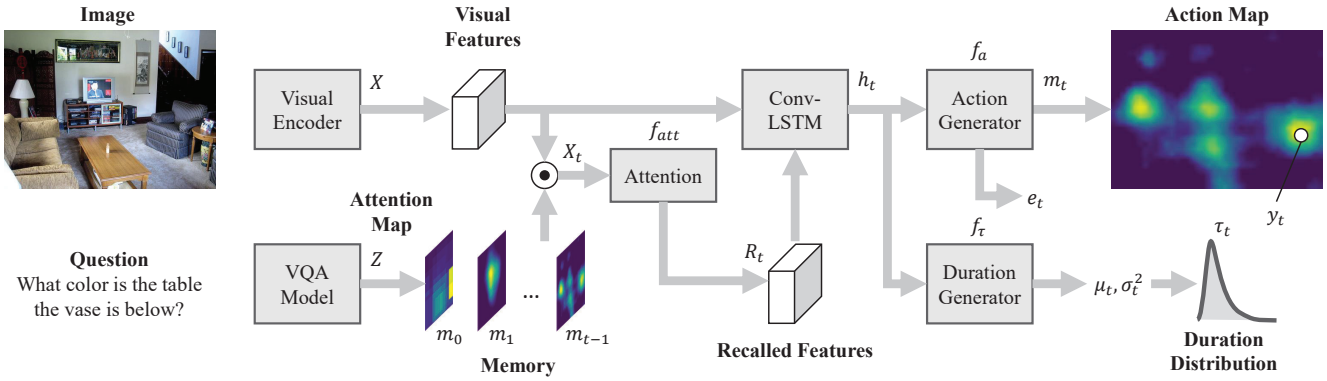

Figure 2. Overview of the proposed scanpath prediction network.

most computational methods are only able to use conventional cross-entropy or saliency evaluation metrics for training, leading to the discrepancy between training-time and test-time contexts. In this work, we adopt self-critical sequence training (SCST) [38] to address this bias by directly optimizing the non-differentiable test-time metrics. Leveraging the effectiveness of SCST, we further introduce a Consistency-Divergence loss to learn the differences between correct and incorrect scanpaths.

3. Method

We develop a deep reinforcement learning model to study and predict complex scanpath patterns in general decision-making tasks, while taking the task performance into account. This section presents the architecture of the proposed network and the machine learning methods to train the network with correct and incorrect scanpaths. Key technical novelties include the creation of a task guidance map to dynamically guide the prediction of fixation positions and durations, a reinforcement learning method with self-critical sequence training to address the exposure bias, and a novel Consistency-Divergence loss to learn the differences between correct and incorrect scanpaths.

3.1. Network Architecture

Where humans look during the VQA task is largely dependent on the input question. Existing task-driven attention models use a one-hot vector [52] or language embeddings [28] to encode the task input. These encoding methods provide semantic guidance to the model, to generate task-dependent outputs, but do not spatially align the task semantics with the visual contents. Differently, we compute a general task guidance map to highlight task-relevant image regions. This task guidance map is designed to be easily adaptable for other tasks. For example, it can be an all-zero matrix for predicting scanpaths in the free-viewing task, or object detection masks can be used to provide task guidance in visual search. In this section, we summarize our method with the general VQA task.

As shown in Fig. 2, we design a neural network model to dynamically generate a sequence of fixation positions and durations. A memory module and an attention mechanism are developed to selectively memorize and recall task-relevant visual information. Specifically, given an image and a question, our goal is to generate a sequence of fixations positions $y = \{y_1, y_2, \cdots, y_T\}$ and durations $\mathcal{T} = \{\tau_1, \tau_2, \cdots, \tau_T\}$. At each step t, the fixation position y_t is sampled from a predicted action map m_t, and the fixation duration τ_t is sampled from a log-normal distribution with two predicted parameters (μ_t, σ_t^2). Besides, a scalar output e_t indicates the end of the scanpath. The specific network design is as follows:

Inputs and task guidance. On the input side, we adopt a CNN-based visual encoder [21] to extract visual features X from the image. The influence of the question is represented as a task guidance map highlighting task-relevant image regions. Trained on large VQA datasets, machine attention can better bridge the task semantics and visual contents by highlighting task-relevant spatial regions that are important for answering the question. Therefore, we guide the prediction of eye fixations using the machine attention of an externally trained VQA model [2, 12, 29, 37]. We preprocess the VQA model's attention into a 2D task guidance map Z with its values normalized within the range of $[0, 1]$.

Memory and attention. Answering complex questions requires dynamically updated memory and attention mechanisms to trace the reasoning process over time [12, 28]. The memory is denoted as $M_t = \{m_0, m_1, \cdots, m_{t-1}\}$, which explicitly maintains all previously computed action maps, as well as the task guidance map $m_0 = Z$. This memory as a whole can be seen as a spatio-temporal attention volume. By applying it to the visual features X, we can obtain the memorized features $X_t = M_t \circ X$, where $\circ$ indicates the Hadamard product. The attention module recalls the most relevant information from the memory, denoted as

$$R_t = f_{att}(X_t; \theta_{att}), \tag{1}$$

where the θ_{att} indicates learnable parameters. It computes

a temporal attention vector indicating the dynamic importance of each historical time step [28], to determine what to recall from the memory for the prediction of the current fixation.

ConvLSTM and outputs. We design a ConvLSTM network to simultaneously predict the distributions of fixation positions and durations. The image features X and the recalled features R_t are fed into a ConvLSTM layer to encode the spatio-temporal patterns of scanpaths. With its current hidden state h_t, the outputs are computed as

$$p_t^a(a_t|a_{1:t-1}) = \text{softmax}(f_a(h_t;\theta_a)), \tag{2}$$

$$[\mu_t, \sigma_t^2] = f_\tau(h_t;\theta_\tau), \tag{3}$$

where f_a and f_τ indicate the output layers and θ_a and θ_τ are learnable parameters. We use $[m_t, e_t] = p_t^a(a_t|a_{1:t-1})$ to represent the distribution of the actions including the action maps m_t and the end-of-scanpath indicator e_t. Finally, we sample the fixation point y_t following the discrete probability values in the action map m_t, and sample the fixation duration following the parametric function $\tau_t \sim p_t^\tau(\tau|\mu_t, \sigma_t^2)$. We model the duration distribution p_t^τ as a log-normal function following previous experimental studies [19, 32].

3.2. Objective

Scanpath prediction is a typical sequential learning task. To address the discrepancy between training and testing contexts in sequential learning, we propose to apply self-critical sequence training (SCST) [38] to directly optimize the non-differentiable evaluation metrics. We further introduce a novel loss function to help differentiate correct and incorrect scanpaths.

Supervised learning. It is widely used in sequential learning to minimize a maximum-likelihood loss at each step. In our context, the objective is to jointly optimize the fixation action a_t and the duration τ_t:

$$L(\theta) = -\sum_{t=1}^{T+1} \log p_t^a(a_t^*|a_{1:t-1}^*;\theta) - \lambda \sum_{t=1}^{T} \log p_t^\tau(\tau_t^*|\mu_t, \sigma_t^2), \tag{4}$$

where T is the length of the ground-truth fixations, a_t^* and τ_t^* are the ground-truth action (one-hot vector indicating the fixation position or end of the scanpath) and fixation duration, respectively. The hyperparameter λ balances the contributions of the two loss terms. With this loss function, we simultaneously train two networks with the correct and incorrect scanpaths. They share most of their parameters, except for the memory and output layers.

However, this objective function does not always produce the best results on the non-differentiable metrics for scanpath evaluation. This discrepancy between training and testing contexts has been observed in similar sequence generation tasks [2, 38]. To address this issue, we propose to use SCST in scanpath prediction and optimize the network based on test-time evaluation metrics.

Reinforcement learning with SCST. Specifically, in the context of scanpath prediction, the objective is to minimize the negative expected reward:

$$L_r(\theta) = -\mathbb{E}_{y,\mathcal{T}}[r(y, \mathcal{T})], \tag{5}$$

where $r(\cdot,\cdot)$ is a reward function (*i.e.,* ScanMatch [14]), while y and $\mathcal{T}$ indicate the sampled fixation positions and durations, respectively. The main idea of SCST is to baseline the REINFORCE algorithm with the reward achieved by the current model under the corresponding evaluation metric used at the test time [38]. To reduce the variance of the gradient estimate and accelerate the training, for each network, we compute the average rewards of k scanpaths and use their mean reward as the corresponding baseline. We denote their corresponding loss functions as $L_r^+(\theta)$ and $L_r^-(\theta)$, respectively. Without loss of generality, in this paper, we use the superscripts $+$ and $-$ to distinguish the notations for correct and incorrect scanpaths, respectively.

Consistency-Divergence loss. The level of difference between correct and incorrect scanpaths is image-specific, so it is difficult to distinguish them by directly learning from the data. We combine the SCST objective with a novel Consistency-Divergence loss (CDL) to explicitly quantify the consistency and divergence of human scanpaths and force the model predictions to resemble such statistics. Specifically, given the correct and incorrect ground-truth scanpaths, we first compute their within-group similarity r^{*+}_{within}, r^{*-}_{within}, and the between-group similarity r^{*}_{between}, by averaging the pair-wise evaluation scores within and between the correct and incorrect groups. The differences $\Delta r^{*+} = r^{*+}_{\text{within}} - r^{*}_{\text{between}}$ and $\Delta r^{*-} = r^{*-}_{\text{within}} - r^{*}_{\text{between}}$ measure the consistency of scanpaths within each group compared with the diversity between the two groups. Intuitively, high within-group similarity and low between-group similarity suggest that the differences between correct and incorrect scanpaths are more distinguishable. Similarly, we can evaluate the predicted scanpaths in the same way to obtain $\Delta r^+(y, \mathcal{T})$ and $\Delta r^-(y, \mathcal{T})$. The objective of the proposed CDL is to let $\Delta r^+(y, \mathcal{T})$ approximate Δr^{*+} and $\Delta r^-(y, \mathcal{T})$ approximate Δr^{*-}, so that the differences between the predicted scanpaths are similar to those of the ground-truth. Therefore, the CDL is computed as

$$\begin{aligned} L_{\text{CD}}(\theta) = \mathbb{E}_{y,\mathcal{T}}\big[|\triangle r^+(y, \mathcal{T}) - \triangle r^{*+}|\big] \\ + \mathbb{E}_{y,\mathcal{T}}\big[|\Delta r^-(y, \mathcal{T}) - \Delta r^{*-}|\big], \end{aligned} \tag{6}$$

Finally, we define the total loss as a linear combination of the negative expected reward and the CDL (6):

$$L'(\theta) = L_r^+(\theta) + L_r^-(\theta) + \gamma L_{\text{CD}}(\theta). \tag{7}$$

The hyperparameter γ balances the contribution of the loss terms in the policy gradient update stage [48].

4. Experiments

We evaluate the proposed method with extensive experiments. Our quantitative and qualitative results demonstrate the performance and generalizability of the proposed method, sheding light on some interesting research questions about scanpath prediction.

4.1. Experiment Settings

Dataset. We conduct our experiments mainly on the AiR dataset [12]. It consists of images and questions selected from the balanced validation set of GQA [23] and provides the eye-tracking data collected from 20 participants who answer the questions. Each question is answered by 10 different participants, and their eye-tracking data are associated with their answers. The numbers of fixations in the recorded scanpaths are similar between the correct answers (10.12 ± 0.99) and the incorrect answers (10.27 ± 1.54). Their spatial priors are also highly similar. These similarities ensure that models do not differentiate between correct scanpaths and incorrect scanpaths based on their prior distributions. We randomly split this dataset into a training set of 1137 questions, a validation set of 142 questions, and a test set of 143 questions. The proportion of correct answers are balanced among these subsets.

Evaluation metrics. To evaluate the models, we generate 10 correct/incorrect scanpaths with each model and compare them with the corresponding ground-truth scanpaths using a combination of four evaluation metrics: The *ScanMatch* [14, 39] measures scanpath similarity based on the Needleman-Wunsch algorithm [7]. It has been commonly used to evaluate scanpath prediction models due to its robustness to the substantial noise inherent in the scanpaths. The *MultiMatch* [17] is a multidimensional evaluation metric, composed of five similarity measures regarding shape, direction, length, position, and duration. The *String-Edit Distance (SED)* [10, 20] is a dissimilarity measure that converts scanpaths into strings by associating each image region with a character. The *Scaled Time-Delay Embedding (STDE)* [46] measures the average of the minimum Euclidean distances of each sub-sequence of the compared scanpaths. For SED and STDE, we report the mean and best evaluation scores. While the mean scores are the averages of all subjects, the best scores are computed based on the most similar human scanpath [18]. These complementary evaluation metrics provide a comprehensive view of the prediction results.

Implementation details. We use ResNet-50 [21] to encode the visual features and use AiR [12] VQA model to compute the task guidance maps. The object-based attention weights are converted to spatial maps by computing a weighted average of their bounding box masks [12]. The resolution of the input image is 240×320. We discrete the fixation position into a 30×40 action map. In supervised learning, we train our model using the Adam [30] optimizer with learning rate 10^{-4} and weight decay 5×10^{-5}. To avoid the divergence of loss, we also adopt the warmup strategy [53] followed by a linear decay of the learning rates. In reinforcement learning, we also use the Adam [30] optimizer with linearly decayed learning rates starting at 5×10^{-5} and weight decay 5×10^{-5}. In SCST, we sample $k = 5$ different scanpaths for the correct and incorrect answers, respectively. The reward function is defined as the harmonic average of the two ScanMatch scores, one with duration and the other without. Our implementation of the ScanMatch metric in training and evaluation follows [14, 39].The hyperparameters λ and γ are empirically set to 1.0 and 2.0, respectively, based on the validation set performance.

4.2. Are the predicted scanpaths plausible?

We first evaluate how well the predicted scanpaths simulate human behavior. Since we are the first to predict scanpaths in the VQA task, for a fair comparison, we customize the most relevant deep-learning-based scanpath prediction models (*i.e.,* SaltiNet [5], PathGAN [4], and IOR-ROI [40]), by combining the BERT embedding [16] of the question with the visual features and jointly predicting the correct and incorrect scanpaths. Following [40, 52], we measure human performance by computing the inter-observer agreements within the correct and incorrect groups, respectively. For each image, we measure the similarity of every pair of human scanpaths from the same group and compute their mean values.

Tab. 1 reports the quantitative results of the compared methods. Our method significantly improves the prediction of both fixation positions and durations. It outperforms the other methods on 9.5/11 metrics by a substantial margin. For example, its ScanMatch scores are over 84% (correct) and 69% (incorrect) higher than the state-of-the-art methods. It even outperforms humans on 6.5/11 metrics.

Fig. 3 presents qualitative examples of the predicted scanpaths. While the state-of-the-art models look at salient objects in general, our predicted scanpaths align better with task-related objects and the human eye-movement behavior regarding fixation positions, durations, and orders. Note that subtle differences of scanpaths can determine the correctness of answers: the incorrect scanpaths consistently miss important objects (*i.e.,* phone and knives).

Note that besides our significant performance boost in predicting correct scanpaths, our method is also effective in predicting scanpaths that lead to incorrect answers thus to be avoided. We find that incorrect scanpaths are less consistent compared with correct ones (also corroborated with Human scores), possibly due to the variety of factors that may lead to an incorrect decision. Yet with the task guidance and the novel CDL loss, our method can capture the subtle differences between the correct and incorrect scan-

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
Human	0.421	0.391	0.945	0.747	0.938	0.879	0.522	7.836	4.804	0.867	0.918
	0.375	0.358	0.938	0.734	0.929	0.851	0.526	8.611	6.406	0.841	0.895
SaltiNet [5]	0.112	0.130	0.933	0.676	0.930	0.696	0.504	10.620	9.264	0.729	0.765
	0.120	0.138	0.930	0.676	0.926	0.696	0.506	10.650	9.750	0.734	0.754
PathGAN [4]	0.210	0.212	0.940	0.637	**0.937**	0.806	0.589	8.658	6.535	0.832	0.862
	0.221	0.218	0.937	0.637	0.927	0.821	0.612	9.071	7.750	0.844	0.861
IOR-ROI [40]	0.171	0.202	0.918	**0.724**	0.908	0.782	0.570	9.210	7.332	0.791	0.818
	0.198	0.216	0.917	**0.737**	0.905	0.793	0.590	9.177	7.945	0.801	0.817
Ours	**0.394**	**0.391**	**0.950**	0.717	0.933	**0.879**	**0.615**	**7.523**	**5.701**	**0.869**	**0.893**
	0.365	**0.368**	**0.946**	0.705	**0.930**	**0.864**	**0.632**	**7.955**	**6.772**	**0.856**	**0.877**

Table 1. Scanpath prediction results on the AiR dataset (VQA). In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. The best results are highlighted in bold. Underlines indicate scores above human performance.

paths, and learn discriminative features relevant to answer correctness to successfully predict both correct and incorrect scanpaths.

4.3. What contributes to the model's performance?

Our proposed method has three major technical contributions: VQA model attention as the task guidance (TG), SCST to address the exposure gap, and the novel Consistency-Divergence loss (CDL). To demonstrate the contribution of each component, we incrementally apply them to a baseline (*i.e.,* a task-ignorant supervised-learning variant of our method). As shown in Tab. 2, each component helps predict both correct and incorrect scanpaths. In particular, though TG results in relatively minor improvements by itself (under supervised learning), it plays a more important role in reinforcement learning with SCST. This observation suggests that SCST can help the model to make better use of the task input to fixate task-relevant regions. Finally, using the new CDL loss together with SCST optimizes the within-group and between-group consistencies of the correct and incorrect scanpaths, thus further increasing the model performance.

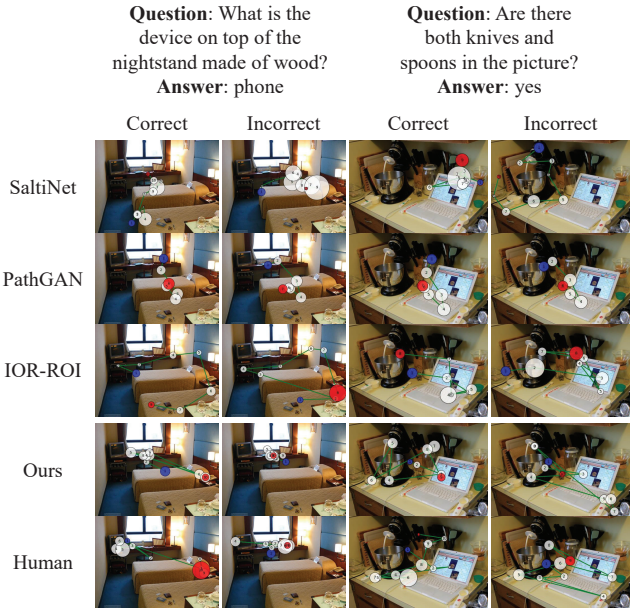

Figure 3. Examples of the predicted scanpaths. Each column compares the prediction results and human scanpaths given specific answer correctness. The number and radius indicate the fixation order and duration, respectively. The blue and red dots indicate the beginning and the end of the scanpath, respectively.

4.4. What do the predicted scanpaths fixate?

To investigate how the predicted scanpaths fixate different objects, we align the fixation positions with the ground-truth object annotations provided by the GQA dataset [23]. We segment each image into three regions: 1) Region of Interest (ROI) is composed of all the objects in the questions and answers; 2) Non-ROI is composed of the other annotated objects that are not included in the ROI; 3) Background is the empty regions without object annotations. For each compared model, we compute the percentage of fixations in each region. As shown in Tab. 3, in general, higher-performance models generate more fixations in the ROI. Our proposed techniques (*i.e.,* TG, SCST, CDL) improve the accuracy of fixating task-relevant objects, allowing our method to perform significantly better than the state-of-the-art methods [4, 5, 40]. The percentage of fixations to ROI of our full model is similar to that of humans. Besides, humans' correct scanpaths fixate the ROI more frequently than the incorrect ones, showing the correlation between their attention allocation and task performance. Our method replicates this correlation, while the compared methods fail to do so. The proposed techniques allow our model to learn

Method			ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
TG	SCST	CDL	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
			0.290	0.323	0.927	**0.719**	0.914	0.845	0.537	8.539	6.829	0.838	0.858
			0.280	0.310	0.920	0.713	0.909	0.831	0.544	8.797	7.667	0.827	0.845
✓			0.296	0.329	0.927	**0.719**	0.914	0.849	0.533	8.438	6.733	0.841	0.862
			0.288	0.317	0.922	**0.717**	0.910	0.837	0.546	8.749	7.682	0.831	0.850
	✓		0.360	0.363	0.948	0.705	0.930	0.865	0.612	7.752	5.961	0.860	0.885
			0.350	0.350	0.943	0.704	0.925	0.852	0.627	8.013	6.818	0.850	0.871
	✓	✓	0.369	0.370	0.949	0.713	**0.933**	0.869	0.605	7.741	5.982	0.860	0.883
			0.350	0.352	0.944	0.716	0.927	0.856	0.616	8.066	6.946	0.849	0.870
✓	✓		0.385	0.383	0.949	0.714	0.932	0.876	0.614	7.569	5.736	0.867	0.891
			0.348	0.354	0.945	0.703	0.928	0.855	0.620	8.011	6.796	0.849	0.873
✓	✓	✓	**0.394**	**0.391**	**0.950**	0.717	**0.933**	**0.879**	**0.615**	**7.523**	**5.701**	**0.869**	**0.893**
			0.365	**0.368**	**0.946**	0.705	**0.930**	**0.864**	**0.632**	**7.955**	**6.772**	**0.856**	**0.877**

Table 2. Ablation study of TG, SCST and CDL on the AiR dataset. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. The best results are highlighted in bold.

Method	Fixations Position %		
	ROI ↑	Non-ROI ↓	Background ↓
Human	26.43	67.48	6.09
	21.60	71.92	6.48
SaltiNet [5]	4.17	77.88	17.95
	3.96	78.49	17.55
PathGAN [4]	7.82	84.34	7.83
	7.17	86.10	6.73
IOR-ROI [40]	9.14	82.99	7.87
	9.79	82.53	7.67
Ours	25.04	69.70	5.26
	22.33	72.27	5.40

Table 3. Percentage of fixations in ROI, non-ROI, and background. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths.

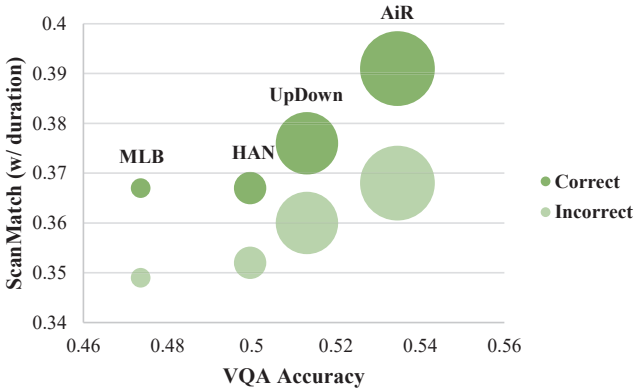

Figure 4. Comparison of VQA models' answer accuracy, scanpath accuracy, and machine attention accuracy (bubble size).

more discriminative features and better distinguish correct and incorrect scanpaths.

4.5. Which VQA model is the most effective?

The explicit use of VQA models in our method allows us to evaluate and visualize VQA models from a human attention's perspective, which has not been explored before. We evaluate the effectiveness of four VQA models: AiR [12], UpDown [2], HAN [37] and MLB [29]. Fig. 4 compares their VQA accuracy on the GQA (test-dev) dataset, machine attention accuracy (AiR-E [12]), and the scanpath prediction performance (ScanMatch w/ duration). As can be seen, both the machine attention accuracy and VQA accuracy are positively correlated with the scanpath prediction performance. Object-based attention maps tend to be more accurate and provide better task guidance: AiR [12] achieves the best performance, thanks to its explicit attention supervision with the ground-truth object annotations. UpDown [2] computes implicitly supervised object-based attention, achieving lower performances in scanpath prediction. HAN [37] relies on attention ground-truth from a specific group of questions [15], which leads to lower performances and difficulties to generalize. MLB [29] is based on image features, so its spatial attention maps may not highlight objects, leading to the lowest performances. In sum, our method suggests that a well-designed machine attention mechanism not only improves the performance of VQA models but also benefits human attention prediction. It also enables further correlational studies between human and machine attention mechanisms.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
Human	0.390	0.386	0.941	0.695	0.931	0.851	0.621	7.486	5.001	0.844	0.906
Itti *et al.* [26]	0.211	0.088	0.824	0.653	0.763	0.685	0.415	8.701	6.529	0.714	0.757
SGC [41]	0.211	–	0.906	0.658	0.870	0.717	–	8.422	6.194	0.771	0.837
Wang *et al.* [46]	0.151	–	0.857	0.641	0.801	0.625	–	9.051	7.129	0.682	0.739
Le Meur *et al.* [34]	0.228	–	0.864	0.657	0.831	0.701	–	8.573	6.536	0.739	0.788
STAR-FC [49]	0.204	–	0.920	0.662	0.900	0.668	–	8.393	6.314	0.751	0.828
SaltiNet [5]	0.169	0.142	0.868	0.647	0.840	0.655	0.566	8.948	7.001	0.706	0.763
PathGAN [4]	0.077	0.079	0.919	0.572	0.905	0.511	0.678	9.414	7.677	0.611	0.691
IOR-ROI [40]	0.267	0.265	0.891	**0.709**	0.860	0.759	0.634	8.180	6.003	0.789	0.844
Ours	**0.383**	**0.377**	**0.943**	0.651	**0.924**	**0.847**	**0.684**	**7.155**	**4.579**	**0.852**	**0.905**

Table 4. Performances on the OSIE dataset (free-viewing). The best results are highlighted in bold. Underlines indicate scores above human performance.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
Human	0.526	0.490	0.944	0.755	0.934	0.913	0.685	2.181	0.359	0.920	0.974
SaltiNet [5]	0.199	0.127	0.909	0.546	0.907	0.740	0.551	4.037	2.742	0.759	0.829
PathGAN [4]	0.277	0.198	0.930	0.561	0.926	0.839	0.604	2.820	1.694	0.847	0.901
IOR-ROI [40]	0.316	0.274	0.919	0.665	0.907	0.834	0.586	4.384	2.595	0.846	0.896
IRL [52]	0.403	–	0.904	0.630	0.887	0.825	–	2.734	1.002	0.898	0.952
Ours	**0.554**	**0.510**	**0.941**	**0.706**	**0.927**	**0.914**	**0.721**	**1.852**	**0.484**	**0.923**	**0.965**

Table 5. Performances on the COCO-Search18 dataset (visual search). The best results are highlighted in bold. Underlines indicate scores above human performance.

4.6. Does the proposed method generalize?

Our method can generalize across tasks with different complexities. Similar to what we observe in the VQA task, results in the free-viewing and visual search tasks also show a significant performance boost, achieving a human-level performance. First, for the free-viewing task (*i.e.,* task guidance and CDL are not applicable), we conduct experiments on the OSIE dataset [50] following the settings of Sun *et al.* [40]. Tab. 4 shows that our method significantly outperforms the state-of-the-art methods [4, 5, 26, 34, 40, 41, 46, 49] on 10/11 metrics with over 42% higher ScanMatch scores. Next, we conduct experiments on COCO-Search18, a visual search dataset [52], using a CenterNet [54] detector to detect the search targets and generate the task guidance maps. As shown in Tab. 5, our method outperforms the state-of-the-art approaches [4, 5, 40, 52] by a large margin and reaches human-level performance on 6/11 metrics. Particularly, our ScanMatch scores are over 37% better than the state-of-the-art [52] and over 5.3% better than humans. These overwhelming performances demonstrate the robustness and generalizability of our method in different task settings.

5. Conclusion

We propose the first model for predicting human scanpaths during visual question answering. By explicitly integrating a task guidance map, the model learns to predict a sequence of task-driven scanpaths that lead to correct or incorrect answers. To address the exposure bias, we propose an SCST approach that optimizes the model based on scanpath evaluation metrics and a Consistency-Divergence loss to distinguish between correct and incorrect scanpaths. Our method significantly outperforms the state-of-the-art methods on multiple datasets and tasks. Our experiments suggest that our model can predict human-like scanpaths and reveal the critical fixation patterns that determine the task performance. The improved performance of human scanpath prediction will push forward the research on task-driven attention and advance a wide range of applications in the development of intelligent robots, automatic design and advertising systems, human-computer interaction systems, and diagnostic tools for mental healthcare.

Acknowledgements

This work is supported by NSF Grants 1908711 and 1849107.

References

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *arXiv preprint arXiv:1806.06877*, 2018.

[4] Marc Assens, Xavier Giro-i-Nieto, Kevin McGuinness, and Noel E. O'Connor. PathGAN: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*, 2018.

[5] Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto, and Noel E. O'Connor. SaltiNet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[6] Dimitri P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, Belmont, Massachusetts, USA, 1st edition, 2019.

[7] Saul B.Needleman and Christian D.Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology (JMB)*, 1970.

[8] Giuseppe Boccignonea and Mario Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications (PHYSICA A)*, 2004.

[9] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2013.

[10] Stephan A. Brandt and Lawrence W. Stark. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience (JCN)*, 1997.

[11] Dirk Brockmann and Theo Geisel. The ecology of gaze shifts. *Neurocomputing*, 2000.

[12] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with reasoning capability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[13] Zhenzhong Chen and Wanjie Sun. Scanpath prediction for visual attention using IOR-ROI LSTM. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[14] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods (BRM)*, 2010.

[15] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods (BRM)*, 2012.

[18] Lapo Faggi, Alessandro Betti, Dario Zanca, Stefano Melacci, and Marco Gori. Wave propagation of visual stimuli in focus of attention. *arXiv preprint arXiv:2006.11035*, 2020.

[19] Gary Feng. Eye movements as time-series random variables: A stochastic model of eye movement control in reading. *Cognitive Systems Research (CSR)*, 2006.

[20] Tom Foulsham and Geoffrey Underwood. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision (JoV)*, 2008.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: searching for coding length increments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2008.

[23] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[24] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research (VR)*, 2000.

[25] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience (NRN)*, 2001.

[26] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 1998.

[27] Ming Jiang, Xavier Boix, Gemma Roig, Juan Xu, Luc Van Gool, and Qi Zhao. Learning to predict sequences of human visual fixations. *IEEE Transactions on Neural Networks and Learning Systems (IEEE TNNLS)*, 2016.

[28] Ming Jiang, Shi Chen, Jinhui Yang, and Qi Zhao. Fantastic answers and where to find them: Immersive question-directed visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[29] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[31] Huiying Liu, Dong Xu, Qingming Huang, Wen Li, Min Xu, and Stephen Lin. Semantically-based human scanpath estimation with HMMs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[32] Arthur J. Lugtigheid. Distributions of fixation durations and visual acquisition rates. *Ph.D. dissertation*, 2007.

[33] Stefan Mathe and Cristian Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[34] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research (VR)*, 2015.

[35] Silviu Minut and Sridhar Mahadevan. A reinforcement learning model of selective visual attention. In *Proceedings of the International Conference on Autonomous Agents (AGENTS)*, 2001.

[36] Dimitri Ognibene, Christian Balkenius, and Gianluca Baldassarre. A reinforcement-learning model of top-down attention based on a potential-action map. *The Challenge of Anticipation Anticipatory Approach (CAAA)*, 2008.

[37] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[38] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[39] Hiroyuki Sogo. Gazeparser: an open-source and multiplatform library for low-cost eye tracking and analysis. *Behavior Reserch Methods (BRM)*, 2013.

[40] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using IOR-ROI recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[41] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Xian-Ming Liu. Toward statistical modeling of saccadic eye-movement and visual saliency. *IEEE Transactions on Image Processing (IEEE TIP)*, 2014.

[42] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts, USA, 2nd edition, 2018.

[43] Hamed Rezazadegan Tavakoli, Esa Rahtu, and Janne Heikkilä. Stochastic bottom-up fixation prediction and saccade generation. *Image Vision Computing (IVC)*, 2013.

[44] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[45] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks (NN)*, 2006.

[46] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. Simulating human saccadic scanpaths on natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[47] Yixiu Wang, Bin Wang, Xiaofeng Wu, and Liming Zhang. Scanpath estimation based on foveated image saliency. *Cognitive Processing (CP)*, 2017.

[48] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning (ML)*, 1992.

[49] Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[50] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision (JoV)*, 2014.

[51] Mai Xu, Yuhang Song, Jianyi Wang, Minglang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[52] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[53] Zhi Zhang, Tong He, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of freebies for training object detection neural networks. *arXiv preprint arXiv:1902.04103*, 2019.

[54] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.