VSP: Visual Scanpath Predictor for Egocentric Videos

Sansitha Panchadsaram[0009−0006−6474−3112] and James J. Clark[0000−0002−4512−6171]

McGill University, Montreal QC H3A 0G4, Canada
sansitha.panchadsaram@mail.mcgill.ca
james.clark1@mcgill.ca

Abstract. Scanpath prediction for egocentric videos is a challenging task because human gaze is constantly moving with ongoing actions, while the first-person viewpoint is also continuously changing. Existing approaches incorporate additional inputs, such as hand movements, body poses, or audio signals, which may not always be available and can reduce generalization. In this work, we present VSP, a Visual Scanpath Predictor for egocentric video that predicts the next fixation location using only past RGB video frames and fixation history, without needing additional modalities. Based on the USP architecture and evaluated on the EGTEA Gaze+ dataset, VSP includes a Dynamic Attention Module (DAM), modality-specific encoders, an Attentive Convolutional LSTM (ALSTM), and a convolutional Gated Multimodal Unit (GMU) to predict the next fixation in a scanpath. Unlike prior work that predicts continuous saliency maps, VSP directly predicts the next fixation location, resulting in predictions that align with real-world human scanpaths. Experiments demonstrate that VSP achieves an F1 score of 55.30%, with ablation studies confirming that each architectural component contributes meaningfully to overall performance. Our results suggest that an accurate prediction of egocentric fixation can be achieved without relying on computationally expensive or environment-specific additional inputs.

Keywords: Egocentric video · Scanpath prediction · Attention modeling.

1 Introduction

When a human performs a task, such as preparing a meal or searching for an item in a store, their eyes move and fixate specific locations that they judge to be relevant to the task or that attracts their attention. Their eyes first fixate one location and then explore the environment to form a sequence of fixations, called a scanpath. Predicting this sequence is the goal of scanpath prediction. While saliency prediction accumulates the regions that people tend to focus on into a single spatial map, scanpath prediction instead models how attention evolves over time, which is more aligned with how humans perceive and interact with

their environment. Because modeling scanpath prediction is more complex than identifying salient regions, this may help explain why saliency prediction has been more widely studied in the literature.

Building on this, modeling scanpath prediction becomes more challenging in egocentric video settings where the camera wearer's gaze is tied to his actions at each moment of the task. Recent datasets have examined gaze behavior in egocentric 3D environments, including task-driven scenarios such as visual search in convenience stores [28]. In these settings, each fixation depends on both the previous fixations and the current step of the task. At the same time, the first-person viewpoint is also changing as the head and body move and as objects are manipulated. As a result, a scanpath prediction model must adapt to this dynamic visual scene to accurately model human gaze behavior. Hence, predicting scanpaths in egocentric videos cannot rely solely on visual features, and it must also consider the sequence of actions and the task goal, making it more complex than in static or third-person scenarios.

Recent work in scanpath prediction for egocentric videos showed that accurately modeling human attention in these environments requires incorporating multiple input streams, such as hand movements, body poses, gaze direction [25, 1, 13, 7]. Although these inputs can be useful in tasks where multiple people interact in an environment, they are not always available, they might be noisy, or they can be computationally expensive to extract. Also, relying on such inputs can unnecessarily complicate the model and reduce its ability to generalize to environments where there are no significant hand or body movements.

In addition, current scanpath prediction models for egocentric videos predict the next fixation in the form of a saliency map. These maps represent a continuous probability distribution over the image plane and are visualized as heatmaps that indicate regions that may attract attention. Although this representation is useful for highlighting broad areas of interest, it differs from human gaze behavior. Real-world scanpaths consist of individual fixation points that occur in a specific order, where each fixation depends on the previous ones and the ongoing task. Saliency maps only represent a spatial probability for a single frame and do not provide an exact fixation location, nor the order of fixations. This gap motivates our approach of explicitly predicting fixation locations to align with real-world human scanpaths and model human gaze behavior.

To address these limitations, we propose VSP, a Visual Scanpath Predictor that builds on [1] by only using past video frames and fixations to predict the fixation in the next frame for an egocentric video. While [1] was developed for conversational videos and incorporates social cues, such as facial expressions and gaze direction, our work focuses on egocentric videos, where gaze behavior is primarily influenced by the wearer's actions and visual information, making the incorporation of such social cues in our model less relevant. Furthermore, we perform ablation studies on VSP to isolate the contributions of the various components in the model. Finally, to align with real-world human scanpaths, we predict the next fixation locations rather than saliency maps.

Our contributions are threefold:

1. We present VSP, a next-fixation scanpath predictor for egocentric videos that extends [1] by only receiving as inputs the past video frames and fixation history, which aligns with how human observers view egocentric scenes.
2. Unlike current approaches that output continuous saliency maps in their scanpath prediction model, VSP directly predicts the location of the next fixation, resulting in the model's predictions closely matching real-world human scanpaths.
3. We perform ablation studies that unravel the impact of the DAM, the modality encoders and the GMU to identify the components that are most critical for robust scanpath prediction.

The remainder of this paper is organized as follows. Section 2 reviews related work on saliency and scanpath prediction, with an emphasis on egocentric video settings. Section 3 presents the proposed VSP framework, including the overall architecture, its model components, training objective, and implementation details. Section 4 describes the experimental protocol, including datasets, evaluation metrics, quantitative results, ablation studies, and qualitative analysis. Finally, Section 5 concludes this paper and discusses directions for future work.

2 Related Work

2.1 Visual Saliency Prediction

Visual saliency prediction aims to identify regions in an image or video that attract human attention and then represent it as a continuous spatial heatmap, known as a saliency map. Classical approaches detected and combined low-level visual features such as color, contrast, and orientation to build saliency maps [27, 10, 9, 21, 26]. With the rise of deep learning, large-scale eye-tracking datasets were used to train convolutional neural networks, which produced saliency maps that outperformed prior methods [11, 22, 5]. Recent transformer-based architectures further improved saliency prediction by capturing the spatial relationships between distant image regions [17, 18]. Regarding video saliency, temporal variations were incorporated through recurrent units and optical flow to model how salient regions shift across frames [3, 19]. Even though saliency prediction provides a useful summary of attended regions, it accumulates fixations across observers and time into a single spatial distribution. More specifically, it neglects the sequential nature of human gaze behavior. This motivates focusing on the task of scanpath prediction, which aims to model attention as a function of time.

2.2 Scanpath Prediction

Scanpath prediction aims to model the sequence of fixations of an observer when the latter is viewing a scene. It captures the temporal characteristics of visual attention rather than producing a spatial accumulation of likely fixated regions.

Early scanpath models are based on cognitive models of visual search and probabilistic models to produce fixation sequences from low-level visual features [8, 4]. Later deep learning approaches used recurrent neural networks to model the relationships between successive fixations where each predicted fixation is conditioned on the prior fixations [2, 23]. More recently, transformer-based scanpath models have been adopted to model spatially distant dependencies in fixation sequences and they have achieved strong performance on static images [20, 29].

Adapting scanpath prediction to egocentric video has attracted increasing attention. Lai et al. [12] proposed GLC, an egocentric scanpath model that uses global context, which captures the relationships between distant parts in an image, and the local context around the current fixation to predict the next fixation location. Li et al. [14] introduced EgoM2P, a model that trains on multimodal inputs, namely RGB video, depth, and camera poses to perform the following egocentric tasks: gaze prediction, camera tracking, and depth estimation. More recently, ARGaze [15] proposed a transformer-based model that uses the visual features of the current frame and the gaze heatmaps of the previous frames to predict the gaze heatmap at each timestep. Even though these models have a strong performance, they either rely on additional modalities, including the video frames and the fixation history or they output saliency maps rather than fixation locations. Both of these implementation choices limit real-world applicability because these models cannot be used in scenarios where such additional modalities are unavailable and they output saliency maps, which doesn't align with actual human scanpaths. VSP directly addresses these gaps by conditioning only on past video frames and fixation history and by predicting the next fixation location, rather than a saliency map.

3 Methodology

In this section, we present VSP, a scanpath prediction model for next-fixation prediction from visual input and fixation history in egocentric video. We begin by formalizing the scanpath prediction problem, then describe the architecture and its main components, followed by the training objective.

3.1 Problem Formulation

Let $\mathcal{V} = \{v_i\}_{i=1}^{N}$ represent a dataset of N egocentric videos, where each video v_i is divided into multiple temporal segments k, and each segment is represented as a sequence of T RGB frames $F_i^k = f_t^k t = 1^T$, with $f_t^k \in \mathbb{R}^{H \times W \times 3}$. Each frame f_t^k is associated with a corresponding fixation density map $m_t^k \in \mathbb{R}^{H \times W}$, where the map represents the spatial distribution of human gaze at time step t, obtained by convolving each fixation point with a Gaussian kernel. For each segment, we define $\mathbf{F}_i^k = \{f_t^k\}_{t=1}^{T}$ and $\mathbf{M}_i^k = \{m_t^k\}_{t=1}^{T}$ as the RGB frame sequence and its corresponding fixation density sequence, respectively.

The objective is to learn a predictive mapping $G_\theta : (\mathbf{F}_i^{k,1:T-1}, \mathbf{M}_i^{k,1:T-1}) \rightarrow \hat{m}_T^k$, parameterized by θ, such that given the past $T-1$ frames and their fixation history within a segment, the model predicts the next fixation density map $\hat{m}_T^k$.

Formally, we optimize:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(v_i,k)\sim\mathcal{V}} \left[\mathcal{L}\left(G_\theta(\mathbf{F}_i^{k,1:T-1}, \mathbf{M}_i^{k,1:T-1}), m_T^k \right) \right] \quad (1)$$

where $\mathcal{L}$ is a loss function measuring the error between predicted and ground-truth fixation density maps across all segments and videos.

3.2 Model Architecture

Figure 1 illustrates the full architecture of the proposed VSP model, which is based on the USP framework [1] and adapted for bimodal egocentric fixation prediction. The model takes as input a sequence of $T-1$ bimodal frames, each consisting of an RGB frame $f_t \in \mathbb{R}^{H\times W\times 3}$ and its corresponding fixation density map $m_t \in \mathbb{R}^{H\times W}$, and produces a predicted fixation density map $\hat{m}_T$ for the next time step. The architecture is composed of four main components: a Dynamic Attention Module (DAM), two modality encoders, an Attentive Convolutional LSTM (ALSTM), and a convolutional Gated Multimodal Unit (GMU), each described in detail below.

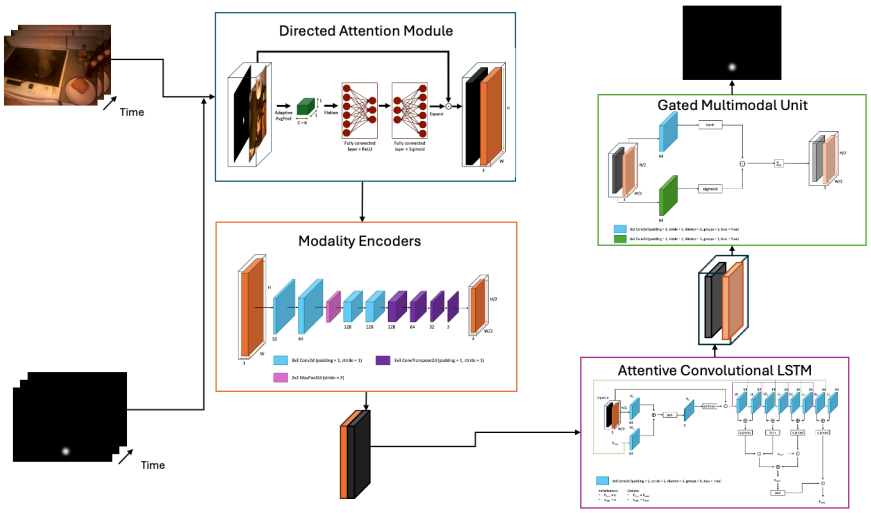

Fig. 1. Architecture of the VSP model. The framework consists of four main components: a Dynamic Attention Module (DAM), two modality encoders, an Attentive Convolutional LSTM (ALSTM), and a convolutional Gated Multimodal Unit (GMU). The DAM emphasizes informative features, the encoders extract modality-specific representations, the ALSTM models temporal dependencies with attention, and the GMU fuses multimodal features for downstream prediction.

Dynamic Attention Module. The DAM is the first component of the model architecture and performs adaptive fusion of the RGB frames and fixation density maps before further processing. As shown in Figure 1, it follows a Squeeze-and-Excitation design [6], where the concatenated multimodal features are first globally pooled to obtain channel descriptors. These are passed through a two-layer gating network to produce channel-wise attention weights. The weights are then applied to the input features via element-wise multiplication, forming reweighted multimodal representations that adaptively balance the contribution of each modality.

Modality Encoders. Each input modality is processed independently by a modality encoder to extract modality-specific latent features. The same encoder weights are shared across all timestamps to ensure consistent representations over time. As shown in Figure 1, the encoder applies convolutional layers to increase the channel depth from 3 to 32, 64, and 128, followed by max-pooling to reduce spatial resolution and capture contextual information. Transposed convolution layers then upsample the features while reducing the channels from 128 to 64, 32, and 3. The resulting feature maps from both modalities are concatenated along the channel dimension to form a joint representation.

Attentive Convolutional LSTM. The concatenated modality-specific feature maps are fed into the ALSTM, whose architecture is shown in Figure 1. The ALSTM models temporal dependencies across the input sequence while preserving spatial structure through convolutional recurrent modules. At each time step, the concatenated feature maps are first refined by a spatial attention mechanism that generates attention weights from the current hidden state and input features to emphasize relevant regions. The attention-modulated features are then passed to a convolutional LSTM cell, where the input, forget, output, and cell-update gates are implemented using 3×3 convolutions. These operations are repeated over the sequence, and the hidden state at the final time step $T - 1$ serves as a compact spatiotemporal representation that is passed to the next module.

Convolutional Gated Multimodal Unit. The output of the ALSTM is passed to a convolutional GMU, whose architecture is shown in Figure 1. This module fuses the modality-specific feature maps into the final fixation density map $\hat{m}_T$ while preserving their 2D spatial structure. The input feature maps are stacked along the channel dimension and processed through two parallel branches: a modality activation branch and a gate activation branch. The modality branch applies grouped 3×3 convolutions independently to each modality, followed by a tanh activation to produce modality-specific features. In parallel, the gate branch applies a standard 3×3 convolution over the concatenated features, followed by a sigmoid activation to generate spatial gating weights. The gated modality features are then combined through element-wise multiplication and summed across modalities to produce the final prediction.

3.3 Loss Function

The model is optimized using a weighted combination of three loss functions [1, 5]:

$$\mathcal{L} = \lambda_{\text{DAM}} \cdot \mathcal{L}_{\text{DAM}} + \lambda_{\text{NLL}} \cdot \mathcal{L}_{\text{NLL}} + \lambda_{\text{KLD}} \cdot \mathcal{L}_{\text{KLD}} \tag{2}$$

where $\mathcal{L}_{\text{NLL}}$ is the negative log-likelihood loss [24] that penalizes low predicted probability at true fixation locations, $\mathcal{L}_{\text{KLD}}$ is the Kullback–Leibler divergence between the predicted and ground-truth fixation density distributions, and $\mathcal{L}_{\text{DAM}}$ is a binary cross-entropy loss that supervises the attention maps of the DAM against the ground-truth fixation maps. Inspired by [1], the optimal loss weights were selected through a hyperparameter search in the range $[0.01, 1]$, with values sampled from a log-normal distribution, producing $\lambda_{\text{DAM}} = 0.9$, $\lambda_{\text{NLL}} = 0.01$, and $\lambda_{\text{KLD}} = 0.9$.

3.4 Implementation Details

The model is implemented in PyTorch and trained on a single NVIDIA H100 GPU. Input video frames are resized to 120×160 pixels and target fixation density maps to 60×80 pixels, with temporal sequences of length $T = 10$ and a batch size of 32. Video sequences are split at the video level into 70% training, 15% validation, and 15% test sets to avoid content leakage across splits. The Adam optimizer is used with a learning rate of 10^{-4} and weight decay of 10^{-3}, with learning rate scheduling applied via ReduceLROnPlateau. Training is performed for up to 20 epochs with early stopping based on validation loss with a patience of 5 epochs.

4 Experiments

4.1 Dataset

The Extended GTEA Gaze+ dataset (EGTEA Gaze+) [16] is a large-scale egocentric video dataset consisting of 29 hours of first-person video recordings from 86 sessions, where 32 subjects perform 7 meal preparation tasks in a kitchen setting. Videos are recorded at 1280×960 pixels and 24 fps, with gaze data captured for every frame. The dataset contains 10,321 labeled action segments across 106 categories, with an average action duration of 3.2 seconds. We use the first train-test split provided by the dataset creators, producing 8,299 action instances for training and 2,022 for testing, using only the action video segments.

4.2 Evaluation Metrics

We evaluate the performance of the model's fixation prediction using precision, recall, and the F1 score, as the task focuses on predicting fixation locations rather than dense saliency maps. Our task results in a highly imbalanced binary

classification problem at the pixel level, where the majority of locations correspond to non-fixated regions, which motivates the use of precision, recall, and F1 score instead of accuracy-based measures.

For each predicted fixation density map, the prediction is first converted to a probability distribution via spatial softmax. Following [12, 15], the ground-truth fixation map is smoothed using a Gaussian filter with $\sigma = 3$ and normalized to the range $[0, 1]$, and binarized using a threshold of 0.001 to obtain the set of true fixation locations. Predicted maps are binarized over 11 adaptive thresholds uniformly sampled from $[0, 0.02]$, and the threshold that maximizes the F1 score is selected for each sample, following [12, 15].

Precision and recall are computed as:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

and the F1 score is defined as their harmonic mean:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives at the pixel level, respectively. The final reported metrics are averaged across all test samples.

4.3 Results

Table 1 reports the fixation prediction performance of our model on the EGTEA Gaze+ test set in terms of precision, recall, and F1 score, compared against recent state-of-the-art methods including EgoM2P, GLC, and ARGaze.

Table 1. Fixation prediction performance on the EGTEA Gaze+ test set. Best results are shown in **bold**.

Method	**Precision (%)**	**Recall (%)**	**F1 (%)**
EgoM2P	35.56	46.70	40.14
GLC	32.41	57.94	41.57
ARGaze	35.22	**60.37**	44.48
VSP (Ours)	**56.43**	54.20	**55.30**

Our method achieves the best overall performance, obtaining an F1 score of 55.30%, outperforming all competing methods. In particular, VSP improves the F1 score by more than 10 percentage points over the strongest baseline, ARGaze (44.48%). While some of the other methods have a higher recall, our model provides a balance of both precision and recall, achieving the highest precision (56.43%) and a competitive recall (54.20%). These results show that using past video frames and fixation history is highly effective in anticipating future fixation locations in egocentric videos.

4.4 Ablation Study

To assess the contribution of each architectural component, we conduct three ablation experiments by removing the Dynamic Attention Module (DAM), the modality encoders, and the Gated Multimodal Unit (GMU) from the full model, respectively, while keeping all other components and training parameters identical.

Table 2. Ablation study results on the EGTEA Gaze+ test set. Best results are shown in **bold**.

Model Variant	Precision (%)	Recall (%)	F1 (%)
Full model (VSP)	**56.43**	**54.20**	**55.30**
Without DAM	53.10	51.04	52.05
Without Modality Encoders	49.80	47.60	48.68
Without GMU	51.90	50.10	50.98

The F1 score decreases from 55.30% to 52.05% when the DAM is removed, which indicates that the role of adaptive modality weighting played by the DAM is important to balance the contribution of the RGB and fixation density map inputs. Without the DAM, both modalities are treated equally, which prevents the model from focusing on the most informative modality at each time step. When the modality encoders are removed, it results in a drop from 55.30% to 48.68% in the F1 score, which signifies that the latent representations specific to the modality learned by the encoders are essential for learning complementary features from each input before feeding them into the ALSTM. Lastly, the removal of the GMU causes the F1 score to decrease from 55.30% to 50.98%, demonstrating that the gated fusion mechanism is important because it helps combine modality-specific representations into a final fixation prediction. Without this mechanism, the model would not be able to accurately weight the contribution of each modality at the spatial level. Together, these results validate the design choices of the full VSP architecture.

4.5 Qualitative Analysis

Figure 2, Figure 3, Figure 4, and Figure 5 present qualitative examples of fixation density map predictions produced by our model on the EGTEA Gaze+ test set during the Greek salad, Cheeseburger, Pasta salad, and Turkey sandwich cooking tasks. Each example presents the input video frame, the predicted fixation density map, the Gaussian-blurred ground-truth fixation density map, and an overlay of both maps, where red regions correspond to our model predictions and green regions correspond to the ground-truth fixations, with yellow indicating overlap. When the subject is manipulating ingredients or utensils in the tasks, the model is able to predict fixations in areas that meaningful to the ongoing action, such that the hands and objects with which the subject interacts, which

aligns with prior findings on egocentric gaze behavior [16]. A blue frame denotes a time step for which no ground-truth fixation was recorded. Even in the absence of a fixation for those frames, our model still produces likely predictions based on the visual scene and temporal information.

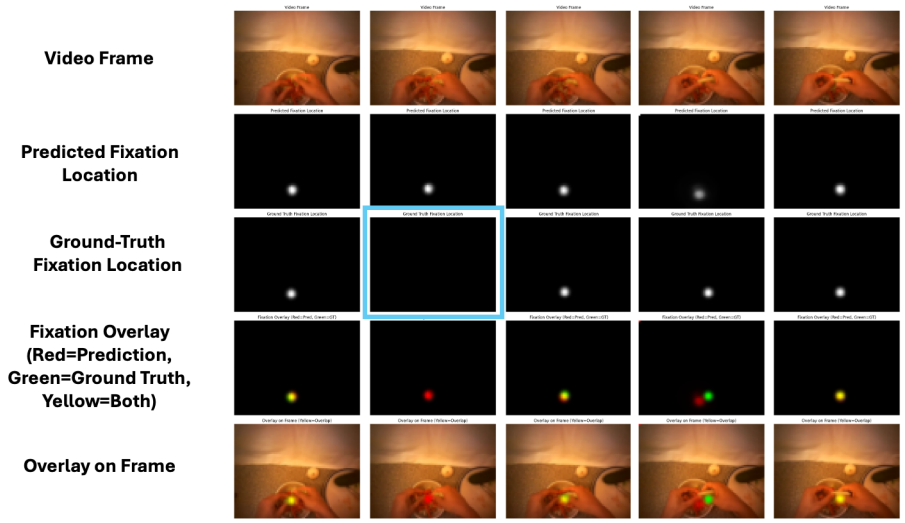

Fig. 2. Qualitative examples of fixation density map predictions on the EGTEA Gaze+ test set during the **Greek salad** cooking task. Red regions indicate model predictions, green regions indicate ground-truth fixations, yellow indicates overlap, and the blue outline around the frame shows a frame with no recorded ground-truth fixation.

5 Conclusion

In this work, we presented VSP, a Visual Scanpath Predictor for egocentric video that predicts the next fixation location using only past RGB video frames and fixation history, without needing additional input modalities such as hand movements, body poses, or gaze direction. By building upon the USP architecture [1] and adapting it to the egocentric fixation prediction domain, VSP integrates a Dynamic Attention Module (DAM), modality-specific encoders, an Attentive Convolutional LSTM (ALSTM), and a convolutional Gated Multimodal Unit (GMU) to jointly model visual appearance and gaze history across time. Unlike prior work that predicts continuous saliency maps, VSP directly predicts the next fixation location, resulting in predictions that align more with real-world human scanpaths. Experiments on the EGTEA Gaze+ dataset [16] show that VSP achieves an F1 score of 55.30%, and our ablation studies confirm that the DAM, modality encoders, and GMU each contribute meaningfully to the overall performance, with the GMU playing a particularly important role in spatially

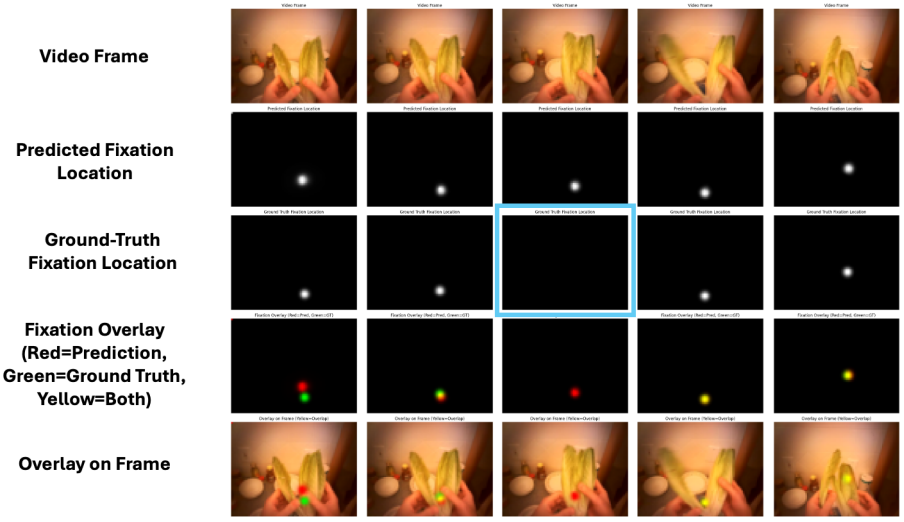

Fig. 3. Qualitative examples of fixation density map predictions on the EGTEA Gaze+ test set during the **Cheeseburger** cooking task. Red regions indicate model predictions, green regions indicate ground-truth fixations, yellow indicates overlap, and the blue outline around the frame shows a frame with no recorded ground-truth fixation.

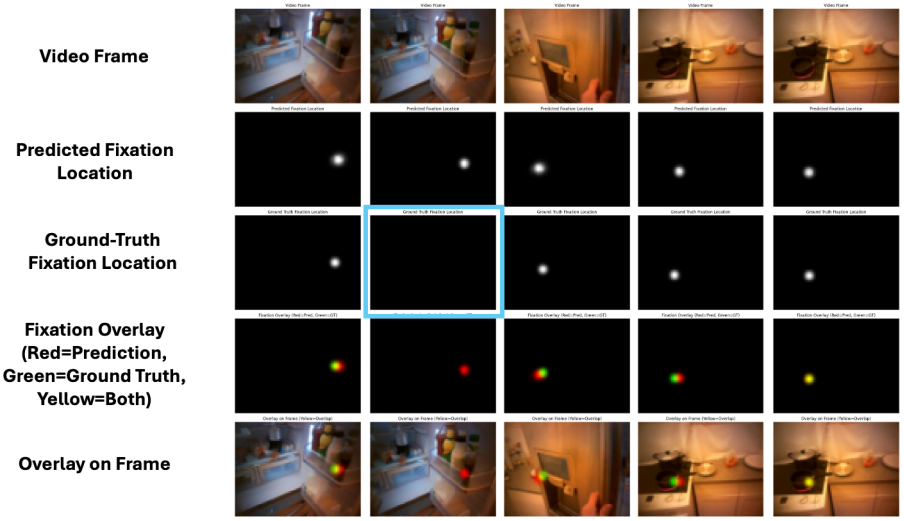

Fig. 4. Qualitative examples of fixation density map predictions on the EGTEA Gaze+ test set during the **Pasta salad** cooking task. Red regions indicate model predictions, green regions indicate ground-truth fixations, yellow indicates overlap, and the blue outline around the frame shows a frame with no recorded ground-truth fixation.

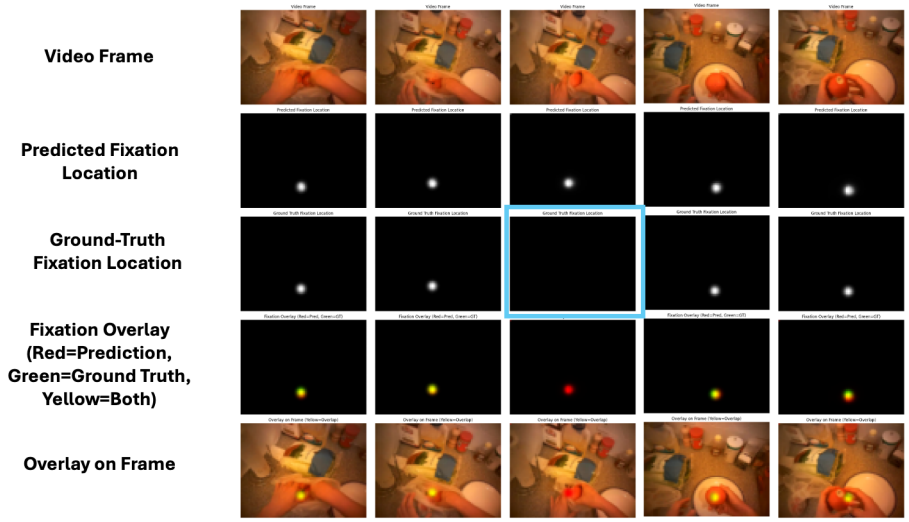

Fig. 5. Qualitative examples of fixation density map predictions on the EGTEA Gaze+ test set during the **Turkey sandwich** cooking task. Red regions indicate model predictions, green regions indicate ground-truth fixations, yellow indicates overlap, and the blue outline around the frame shows a frame with no recorded ground-truth fixation.

merging the two input modalities to form an accurate fixation location prediction.

Despite these promising results, VSP has few limitations. The model is evaluated on a single dataset, and its generalization to other egocentric datasets or activity domains remains to be explored. Furthermore, the model operates on fixed-length temporal sequences, which may limit its ability to capture dependencies across longer action segments. Finally, the current evaluation relies on pixel-level fixation metrics, which do not fully capture the sequential nature of human scanpaths.

Future work will focus on validating VSP across other egocentric datasets and activity domains, extending the model to handle variable-length temporal sequences, and incorporating scanpath-level evaluation metrics that take into account the temporal ordering of human fixations.

References

1. Abawi, F., Fu, D., Wermter, S.: Unified dynamic scanpath predictors outperform individually trained neural models. arXiv preprint arXiv:2405.02929 (2024)
2. Assens, M., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E.: Pathgan: Visual scanpath prediction with generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
3. Bak, C., Kocak, A., Erdem, E., Erdem, A.: Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Transactions on Multimedia **20**(7), 1688–1698 (2017)

4. Boccignone, G., Ferraro, M.: Modelling gaze shift as a constrained random walk. Physica A: Statistical Mechanics and its Applications **331**(1-2), 207–218 (2004)
5. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an lstm-based saliency attentive model. IEEE Transactions on Image Processing **27**(10), 5142–5154 (2018)
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
7. Hu, Z., Xu, J., Schmitt, S., Bulling, A.: Pose2gaze: Eye-body coordination during daily activities for gaze prediction from full-body poses. IEEE Transactions on Visualization and Computer Graphics (2024)
8. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research **40**(10-12), 1489–1506 (2000)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence **20**(11), 1254–1259 (1998)
10. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human neurobiology **4**(4), 219–227 (1985)
11. Kümmerer, M., Theis, L., Bethge, M.: Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045 (2014)
12. Lai, B., Liu, M., Ryan, F., Rehg, J.M.: In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. International Journal of Computer Vision **132**(3), 854–871 (2024)
13. Lai, B., Ryan, F., Jia, W., Liu, M., Rehg, J.M.: Listen to look into the future: Audio-visual egocentric gaze anticipation. In: European Conference on Computer Vision. pp. 192–210. Springer (2024)
14. Li, G., Chen, Y., Wu, Y., Zhao, K., Pollefeys, M., Tang, S.: Egom2p: Egocentric multimodal multitask pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10830–10843 (2025)
15. Li, J., Zhao, W., Deng, S., Lai, B., Wu, Y., Chen, R., Froehlich, J.E., Zhao, Y., Tian, Y.: Argaze: Autoregressive transformers for online egocentric gaze estimation. arXiv preprint arXiv:2602.05132 (2026)
16. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European conference on computer vision (ECCV). pp. 619–635 (2018)
17. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4722–4732 (2021)
18. Lou, J., Lin, H., Marshall, D., Saupe, D., Liu, H.: Transalnet: Towards perceptually relevant visual saliency prediction. Neurocomputing **494**, 455–467 (2022)
19. Min, K., Corso, J.J.: Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2394–2403 (2019)
20. Mondal, S., Yang, Z., Ahn, S., Samaras, D., Zelinsky, G., Hoai, M.: Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1441–1450 (2023)
21. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vision research **45**(2), 205–231 (2005)
22. Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 598–606 (2016)

23. Sun, W., Chen, Z., Wu, F.: Visual scanpath prediction using ior-roi recurrent mixture density network. IEEE transactions on pattern analysis and machine intelligence **43**(6), 2101–2118 (2019)
24. Sun, W., Chen, Z., Wu, F.: Visual scanpath prediction using ior-roi recurrent mixture density network. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(6), 2101–2118 (2021). https://doi.org/10.1109/TPAMI.2019.2956930
25. Thakur, S.K., Beyan, C., Morerio, P., Del Bue, A.: Predicting gaze from egocentric social interaction videos and imu data. In: Proceedings of the 2021 International Conference on Multimodal Interaction. pp. 717–722 (2021)
26. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review **113**(4), 766 (2006)
27. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **12**(1), 97–136 (1980)
28. Wang, Y., Panchadsaram, S., Sherkati, R., Clark, J.J.: An egocentric video and eye-tracking dataset for visual search in convenience stores. Computer Vision and Image Understanding **248**, 104129 (2024)
29. Yang, Z., Mondal, S., Ahn, S., Xue, R., Zelinsky, G., Hoai, M., Samaras, D.: Unifying top-down and bottom-up scanpath prediction using transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1683–1693 (2024)