In the Eye of Transformer: Global-Local Correlation for Egocentric Gaze Estimation

Bolin Lai
bolin.lai@gatech.edu

Miao Liu
mliu328@gatech.edu

Fiona Ryan
fkryan@gatech.edu

James M. Rehg
rehg@gatech.edu

College of Computing
Georgia Institute of Technology
Atlanta, GA

Abstract

In this paper, we present the first transformer-based model to address the challenging problem of egocentric gaze estimation. We observe that the connection between the global scene context and local visual information is vital for localizing the gaze fixation from egocentric video frames. To this end, we design the transformer encoder to embed the global context as one additional visual token and further propose a novel Global-Local Correlation (GLC) module to explicitly model the correlation of the global token and each local token. We validate our model on two egocentric video datasets – EGTEA Gaze+ and Ego4D. Our detailed ablation studies demonstrate the benefits of our method. In addition, our approach exceeds previous state-of-the-arts by a large margin. We also provide additional visualizations to support our claim that global-local correlation serves a key representation for predicting gaze fixation from egocentric videos. More details can be found in our website (https://bolinlai.github.io/GLC-EgoGazeEst).

1 Introduction

Findings in cognitive neuroscience suggest that eye movements reflect cognitive processes [53], which are essential for understanding human intention during daily activities [21]. Such an understanding of visual attention and intention can be valuable for many applications, including Augmented Reality (AR), Virtual Reality (VR), and Human-Robot Interaction (HRI). While wearable eye trackers are a standard way to obtain measurements of gaze behavior, they require calibration, consume significant power, and add substantial cost and complexity to wearable platforms. Alternatively, prior works [1, 22, 23, 24, 34, 36, 44, 49, 52, 53] seek to estimate the visual attention of the camera-wearer from videos captured from a first-person perspective. This task is known as egocentric gaze estimation.

The key challenge in egocentric gaze estimation is to effectively integrate multiple gaze cues into a *holistic* analysis of visual attention. Cues include the likelihood that different scene objects are gaze targets (i.e. salience), the relative location of gaze targets within the
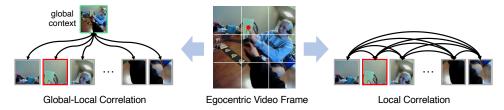
Figure 1: Example of local correlation and global-local correlation for the task of egocentric gaze estimation (predicting where the camera-wearer is looking using egocentric video alone). The red dot represents the gaze ground truth (from a wearable eye tracker) and the image patch that contains the gaze target has red edges. Global-local correlation models the connections between the global context and each local patch, making it possible to capture, e.g., the camera wearer and social partner are pointing at the salient object. In contrast, local-local correlations may not yield an effective representation of the scene context.

video frame (i.e. center prior), and the patterns of camera movement that are reflective of visual attention (i.e. head motions accompanying a gaze shift). Recently, the transformer has achieved great success in various vision tasks by modeling the spatio-temporal correlation among local visual tokens [12, 16, 32, 39, 41, 48, 50, 52]. However, the pairwise comparisons performed by standard transformer Self-Attention (SA) are not optimized for interpreting local video features in the context of the global scene. The example Fig. 1 illustrates the key role of comparisons between local patches and global context - the gaze target is a salient object pointed at by both the camera wearer and another person. Such a salient object can not be easily localized by only modeling the correlation of local patches.

This paper introduces a novel transformer-based deep model that explicitly embeds global context and calculates spatio-temporal global-local correlation for egocentric gaze estimation. Specifically, we design a transformer encoder that adopts a global visual token embedding strategy to incorporate the global scene context. We then introduce a novel Global-Local Correlation (GLC) module that highlights the connection between global and local visual tokens. Finally, we adopt a transformer-based decoder to produce gaze prediction output. We evaluate our approach on two egocentric video datasets – EGTEA Gaze+ [35] and Ego4D [18]. Our proposed model is easy to incorporate into existing transformer-based video analysis architectures, and we show that it yields an *improvement of more than* 3.9% *in F1 score* over SOTA methods for egocentric gaze estimation. The code and pretrained models will be made publicly available to the research community. In summary, this work makes the following contributions:

- We introduce the first transformer-based approach to address the challenging task of egocentric gaze estimation.

- We utilize a global visual token embedding strategy to incorporate global visual context into self-attention, and further introduce a novel Global-Local Correlation module to explicitly model the correlation between global context and each local visual token.

- Our novel design obtains consistent improvement on the EGTEA Gaze+ [35] and Ego4D [18] datasets and outperforms previous state-of-the-art methods by at least 3.9% on EGTEA and 5.6% on Ego4D in F1 score. Importantly, this is the first work that uses the Ego4D dataset for egocentric gaze estimation, which serves as important benchmark for future research in this direction.

2 Related Work

The task of egocentric gaze estimation, is distinct from prior work on eye tracking [30, 43, 64] and gaze target prediction from the third person video [9, 10, 28, 46]. In the interest of space, we limit our discussion to prior work on egocentric gaze estimation and related works on transformer-based video representation learning and video saliency prediction.

Egocentric Gaze Estimation. Previous works have shown that gaze behavior plays an important role in recognizing human daily actions from egocentric videos [1, 22, 23, 24, 34, 36, 37, 38, 44, 49, 52, 53, 65]. Here, we discuss the most relevant works that develop deep models for egocentric gaze estimation. Zhang et al. [65] used deep models and an adversarial network to predict egocentric gaze location in future video frames. Huang et al. [22] incorporated temporal attention transition into saliency-based models for gaze estimation. Tavakoli et al. [52] studied both top-down and bottom-up cues that contribute to gaze guidance. Park et al. [49] introduced the novel problem of predicting joint attention during social interaction using egocentric videos. Huang et al. [24] collected a new egocentric video dataset and developed a graphical model to detect joint attention. Thakur et al. [53] proposed a multi-modal network that uses both video and inertial measurement unit data for more accurate egocentric gaze estimation. Naas et al. [44] developed a tiling scheme for gaze prediction which enables a more efficient VR content delivery. Importantly, we are the first to develop a transformer-based architecture to address the problem of egocentric gaze estimation.

Vision Transformer. Recently, vision transformers [14] have demonstrated superior performance on image classification [13, 32, 40, 48, 58, 62], detection [5, 11, 12, 16], segmentation [8, 50, 55, 66, 67], saliency prediction [39, 41, 42] and video analysis [2, 4, 15, 33, 45, 56]. In this section, we focus on reviewing previous works that use vision transformers for pixel-wise visual prediction and video understanding. Strudel et al. [50] developed the first transformer-based architecture for semantic segmentation. Cheng et al. [8] further unified semantic, instance, and panoptic segmentation in one transformer architecture. Bertasius et al. [4] proposed TimeSformer for video action recognition. A similar idea is also explored in [2]. Fan et al. [15] designed a multiscale video transformer balancing computational cost and action recognition performance. Ma et al. [42] expanded transformers to visual saliency forecasting by using self-attention to capture the correlation between past and future frames. Liu et al. [39] built a transformer-based model to detect salient objects on RGB-D images. Inspired by these successful applications of transformer architectures, we present the first work that uses a vision transformer to address the challenging task of egocentric gaze estimation. In addition, we introduce the novel Global-Local Correlation (GLC) module that provides additional insight into video representation learning with self-attention.

Visual Saliency. Visual saliency prediction has been well studied in computer vision [3, 6, 7, 25, 26, 29, 31, 41, 47, 51, 54, 57, 60, 61]. We mainly review previous works on saliency prediction from videos. Wang et al. [57] expanded image saliency models to videos by incorporating a new branch to handle temporal information. Wu et al. [61] proposed SalSAC, which shuffles features of different CNN layers and feeds them to a correlation-based ConvLSTM. Wang et al. [60] used multiple spatio-temporal self-attention modules to address the limitation of fixed kernel size in 3D models and to model long-range temporal dependencies. Chen et al. [7] decomposed video saliency prediction into spatial pattern capture and spatio-temporal reasoning. Lou et al. [41] combined a convolutional network and transformer architecture to model the long-range spatial context. While visual saliency prediction localizes interesting spatial regions as potential attention targets, egocentric gaze estimation seeks to determine the gaze target of the camera wearers as they interact with a
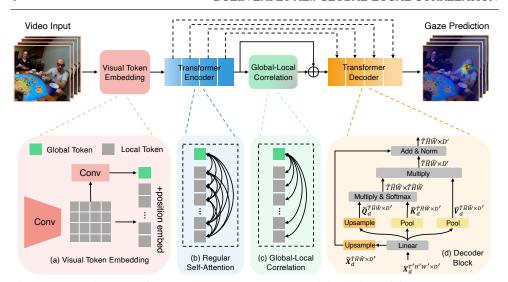
Figure 2: Architecture of the proposed model. The model consists of four modules – (a) **Visual Token Embedding Module** encodes the input into local tokens and one global token, (b) **Transformer Encoder** is composed of multiple regular self-attention and linear layers, (c) **Global-Local Correlation Module** models the correlation of global and local tokens, and (d) **Transformer Decoder** maps encoded video features from Transformer Encoder and GLC to gaze prediction. $\oplus$ denotes concatenation along the channel dimension.

scene. Moreover, the scene context captured from egocentric video is complex and rapidly changing, which requires a gaze estimation model with the ability of explicitly reasoning about the correlation between local visual features and global scene context. In our experiment section, we demonstrate that our proposed GLC module can significantly benefit gaze estimation performance under this challenging setting.

3 Method

Given an input egocentric video clip with fixed length T and spatial dimension $H \times W$, our goal is to predict the gaze location in each video frame. Following [35], we consider the gaze prediction as a probabilistic distribution defined on the 2D image plane.

Fig. 2 presents an overview of our proposed method. We use the recent multi-scale video transformer (MViT) architecture [15] as the backbone network for video representation learning. We extend MViT by designing the *Visual Token Embedding Module* to generate the spatio-temporal tokens of both local visual patches and global visual context and feed them into the standard *Multi-Head Self-Attention Module*. We then utilize a novel *Global-Local Correlation (GLC) Module* to explicitly model the correlation between global and local visual tokens for gaze estimation. Finally, we make use of the *Decoder Network* to predict the gaze distribution based on the learned video representation from the GLC module.

3.1 Transformer Encoder with Global Visual Token Embedding

Visual Token Embedding. We split the input video sequence into non-overlapping patches with size $s_T \times s_H \times s_W$ and adopt a linear mapping function to project each flattened patch into

D-dimension vector space. Following MViT [15], this is equivalent to a convolutional layer with a stride of $s_T \times s_H \times s_W$ and a number of output channels of D. This operation results into N tokens where $N = \frac{T}{s_T} \times \frac{H}{s_H} \times \frac{W}{s_W}$. In addition, the learnable positional embedding $\boldsymbol{E} \in \mathbb{R}^{N \times D}$ is added to the local tokens. Our key insight is to further embed global information into a global visual token using convolutional operations, as illustrated in Fig. 2(a). Since there is a single global token, it does not require positional embedding. In our experiments, we examine multiple strategies to embed the global visual token.

Multi-Head Self-Attention Module. The N local tokens and one global token are fed into a transformer encoder consisting of multiple self-attention blocks. The number of local tokens is downsampled after each self-attention block, while the number of global tokens remains 1. Suppose the input of the j-th layer of encoder is $\boldsymbol{X}_e^{(j)} = [\boldsymbol{x}_i^{(j)}]_{i=1}^{N_j+1} \in \mathbb{R}^{(N_j+1) \times D_j}$, where N_j is the number of local tokens, D_j is the vector length of each token and $\boldsymbol{x}_i^{(j)}$ is the i-th row of $\boldsymbol{X}_e^{(j)}$ denoting the i-th token of size $1 \times D_j$. For simplicity, we omit subscript and superscript of j and multi-head operation in the following equations. In each self-attention layer, correlations are calculated in each token pair as shown in Fig. 2(b). They are used to reweight values of each token after softmax. Formally, we denote the query, key and value matrices of each self-attention layer in an encoder block as $\boldsymbol{Q}_e^{(N+1) \times D} = [\boldsymbol{q}_i]_{i=1}^{N+1}$, $\boldsymbol{K}_e^{(N+1) \times D} = [\boldsymbol{k}_i]_{i=1}^{N+1}$ and $\boldsymbol{V}_e^{(N+1) \times D} = [\boldsymbol{v}_i]_{i=1}^{N+1}$. The self-attention in transformer encoder is formulated as

$$Attention(\boldsymbol{Q}_e, \boldsymbol{K}_e, \boldsymbol{V}_e) = Softmax(\boldsymbol{Q}_e \boldsymbol{K}_e^T / \sqrt{D}) \boldsymbol{V}_e \in \mathbb{R}^{(N+1) \times D}. \tag{1}$$

Finally, we attach a standard linear layer after the self-attention operation.

3.2 Global-Local Correlation

Even though global information has been explicitly embedded into the global visual token in our model, the transformer encoder treats the global and local tokens equivalently as shown in Eq. 1 and Fig. 2(b). In this case, global-local correlation is diluted by correlations among the local tokens, limiting its impact on gaze estimation. In order to address this problem, we propose to increase the available capacity to model global-local token interactions. Our solution is a novel Global-Local Correlation module described in Fig. 2(c).

Formally, we denote the global token as the first row vector of $\boldsymbol{X}_e$, i.e., $\boldsymbol{x}_1$. Thus $\boldsymbol{q}_1$, $\boldsymbol{k}_1$ and $\boldsymbol{v}_1$ are the query, key and value projected from the global token, respectively. To explicitly model the connection between global and local visual features, we only calculate the correlation between each local token and the global token, i.e., $Correlation(\boldsymbol{x}_i, \boldsymbol{x}_1)$, as well as its self-correlation, i.e., $Correlation(\boldsymbol{x}_i, \boldsymbol{x}_i)$. Then correlation scores are normalized by softmax to further re-weight the values. We exploit a suppression matrix [40] $\boldsymbol{S}^{(N+1) \times (N+1)}$ to suppress the correlation of other tokens, where

$$\boldsymbol{S}^{(N+1) \times (N+1)} = [s_{ij}], \quad s_{ij} = \begin{cases} 0, & \text{if } i = j \text{ or } j = 1 \\ \lambda, & \text{otherwise.} \end{cases} \tag{2}$$

We assign zeros to the diagonal and the first column in $\boldsymbol{S}$ and set a large value λ for the other elements. We follow the empirical choice from the implementation of [40] and set $\lambda = 10^8$ in our experiments. Formally, the proposed GLC can be formulated as the following:

$$GLC(\boldsymbol{Q}_e, \boldsymbol{K}_e, \boldsymbol{V}_e) = Softmax((\boldsymbol{Q}_e \boldsymbol{K}_e^T - \boldsymbol{S}) / \sqrt{D}) \boldsymbol{V}_e \in \mathbb{R}^{(N+1) \times D} \tag{3}$$

In this way, we keep the values on the first column and the diagonal, and map them into probability distributions, while values in other positions are nearly "masked out" after the softmax. Residual connections and linear layers are also used in the GLC module as in the regular self-attention block. Finally, the output tokens from the GLC are concatenated with those from the transformer encoder in the channel dimension. We denote outputs of the GLC and the last encoder block as $\boldsymbol{X}_e^{GLC} \in \mathbb{R}^{(N+1)\times D}$ and $\boldsymbol{X}_e^{SA} \in \mathbb{R}^{(N+1)\times D}$. The concatenation can then be formulated as $\boldsymbol{X}_e = \boldsymbol{X}_e^{SA} \oplus \boldsymbol{X}_e^{GLC} \in \mathbb{R}^{(N+1)\times 2D}$. The fused tokens $\boldsymbol{X}_e$ are subsequently fed into the transformer decoder for gaze estimation.

3.3   Transformer Decoder

To produce the gaze distribution with the desired spatio-temporal resolution, we adopt a decoder to upsample the encoded features. We utilize a transformer decoder based on the multiscale self-attention block of MViT [15]. Suppose each decoder layer takes visual features $\boldsymbol{X}_d \in \mathbb{R}^{T'H'W'\times D'}$ as inputs and the corresponding query, key and value matrices are $\boldsymbol{Q}_d^{T'H'W'\times D'}$, $\boldsymbol{K}_d^{T'H'W'\times D'}$ and $\boldsymbol{V}_d^{T'H'W'\times D'}$. As shown in Fig. 2(d), we replace the original pooling operation for the query matrix with an upsampling operation implemented with trilinear interpolation and keep the pooling for the key and value matrices. Following [15], $\widehat{\boldsymbol{Q}}_d$ is obtained by applying a deconvolutional operation on $\boldsymbol{Q}_d$, while $\widehat{\boldsymbol{K}}_d$ and $\widehat{\boldsymbol{V}}_d$ are obtained by applying convolutional operations on $\boldsymbol{K}_d$ and $\boldsymbol{V}_d$. Then, the output of self-attention is calculated in the same way as Eq. 1. In addition, we keep the skip connection in the self-attention layers and replace the pooling operation in skip connections with trilinear interpolation, which produces the upsampled output with dimension $\widehat{T}\widehat{H}\widehat{W} \times D'$. Our decoder is composed of 4 decoding blocks. Skip connections are used to combine intermediate features of the encoder with corresponding decoder features. Finally, another linear mapping function is used to output the final gaze prediction.

3.4   Network Architecture and Model Training

We adopt MViT [15] as the backbone, with weights initialized from Kinetics-400 pretraining [27]. The token embedding stride is set as $s_T = 2$, $s_H = 4$ and $s_W = 4$ and the embedding dimension is $D = 96$. The encoder is composed of 16 self-attention layers that are divided into 4 blocks. The number of tokens is downsampled at the transition between two blocks. We build the decoder with 4 decoder blocks corresponding to the 4 blocks in the encoder. After getting raw output from decoder, softmax is applied on each frame with a temperature τ. This can be formally written as $\hat{p}_{ij} = \frac{\exp(\hat{y}_{ij}/\tau)}{\sum_{i,j}\exp(\hat{y}_{ij}/\tau)}$ where $\hat{y}_{ij}$ is the logit at location (i,j) from the model and $\hat{p}_{ij}$ is probability after softmax. In experiments, τ is empirically set as 2. We use KL-divergence loss to capture the difference between labels and predictions. More details of the training parameters can be found in the supplementary.

4   Experiment

4.1   Datasets and Metrics

We conducted our experiments on two egocentric video datasets with gaze tracking data to serve as ground truth – EGTEA Gaze+ [35] and Ego4D [18]. The EGTEA dataset is captured under the meal preparation setting, which involves a great deal of hand-object interactions.

Methods	EGTEA Gaze+			Ego4D		
	F1	Recall	Precision	F1	Recall	Precision
MViT	43.0	57.8	35.4	40.9	57.4	31.7
MViT + (a)	43.4	58.4	34.5	41.5	56.8	32.6
MViT + (b)	43.5	59.2	34.4	41.4	57.3	32.4
MViT + (c)	43.7	58.3	34.9	41.3	57.5	32.2
MViT + (d)	43.9	59.0	34.9	41.7	57.6	32.7
MViT + (d) + SA	44.1	58.8	**35.3**	42.1	**58.5**	32.9
MViT + (d) + GLC	**44.8**	**61.2**	**35.3**	**43.1**	57.0	**34.7**

Table 1: Evaluation of different global embedding approaches and global-local correlation module. (a)(b)(c)(d) are different global embedding strategies elaborated in Section 4.2. *SA* and *GLC* denote regular self-attention and global-local correlation module, respectively.

We used the first train/test split from EGTEA in our experiments (8299 clips for training and 2022 clips for testing). The Ego4D dataset includes 27 videos of 80 participants totaling 31 hours with gaze tracking data captured under the social setting. We split the long videos into 5-second video clips and pick clips containing gaze fixation. More details of the selection of clips with gaze fixations are discussed in supplementary. We used 20 videos (15310 clips) for training and the other 7 videos (5202 clips) for testing. This is the first work that uses the Ego4D dataset for egocentric gaze estimation, and we will make our split publicly available to drive future research on this topic. Following [35, 36], we adopt F1 score, recall, and precision as the evaluation metrics.

4.2   Experimental Results

The Design Choice of Global Visual Embedding. Our key insight is embedding the global visual information into the transformer architecture for egocentric gaze estimation. Here, we first explore 4 global visual embedding strategies – (a) direct max pooling on the input, (b) max pooling on the unflattened local tokens, (c) convolutional layers on the input and (d) convolutional layers on the unflattened local tokens.

As shown in Table 1, all four global embedding strategies improve the performance of vanilla MViT model on both the EGTEA dataset and the Ego4D dataset. This result supports our claim that global context is essential for gaze estimation. Among the four embedding strategies, (d) achieves the largest performance improvement on both datasets (+0.9% on EGTEA and +0.8% on Ego4D). This indicates that convolutional layers and the embedded local tokens can facilitate the learning of global context. Thus, we use this strategy in the following experiments. Note that all baseline methods use the same transformer decoder.

Evaluation of Global-Local Correlation. We also evaluate the Global-Local Correlation (GLC) module of our model. As presented in Table 1, our full model – MViT+(d)+GLC outperforms the baseline MViT by +1.8% on EGTEA dataset and +2.2% on Ego4D dataset. Specifically, the GLC module contributes to a performance gain of +0.9% on EGTEA Gaze+ and +1.4% on Ego4D (comparing to MViT+(d)). This result suggests that the GLC can break down the mathematical equivalence of global and local tokens in regular self-attention, thereby "highlighting" the global-local connection in the learned representation.

Does the Performance Improvement Come from Additional Parameters? It is possible that the performance of our model benefits from additional parameters in the GLC module. In Table 1, we report the results of another baseline model, where we replace the GLC

Methods	F1	Recall	Precision
Center Prior	10.7	32.0	6.4
GBVS [20]	15.7	45.1	9.5
EgoGaze [34]	16.3	16.3	16.3
SimpleGaze	31.3	41.8	16.1
Deep Gaze [65]	34.5	43.1	28.7
Gaze MLE [36]	26.6	35.7	21.3
Joint Learning [36]	34.0	42.7	28.3
Attention Transition [22]	37.2	51.9	29.0
I3D-R50 [17]	40.9	57.2	31.8
MViT	43.0	57.8	35.4
Ours	**44.8**	**61.2**	**35.3**

Table 2: Comparison with previous methods on EGTEA Gaze+. Our complete model is highlighted. The proposed model outperforms previous approaches by a significant margin.

Methods	F1	Recall	Precision
Center Prior	14.9	21.9	11.3
GBVS [20]	18.0	47.2	11.1
Attention Transition [22]	36.4	47.6	29.5
I3D-R50 [17]	37.5	52.5	29.2
MViT	40.9	57.4	31.7
Ours	**43.1**	**57.0**	**34.7**

Table 3: Comparison with previous methods on Ego4D. Our complete model is highlighted. The model shows consistent superiority over other state of the arts on all metrics.

module with a regular self-attention (SA) layer. Interestingly, the additional SA layer has minor influence on the overall performance (+0.2% on EGTEA and 0.4% on Ego4D). In contrast, our model outperforms this baseline by +0.7% on EGTEA and +1.0% on Ego4D. This result indicates that the performance boost of our method does not simply come from the additional parameters of GLC. Instead, the explicit modeling of the connection between global and local visual features is the key factor in the performance gain.

Comparison with Previous State-of-the-Art. In addition to these studies to evaluate the components of our model, we compare our approach with prior work. Results are presented in Table 2 and Table 3. Interestingly, the baseline MViT model easily outperforms all previous works that use CNN-based architectures on both the EGTEA dataset and the Ego4D dataset. In addition, our proposed method outperforms the best CNN model by +3.9% on F1, +4.0% on recall and +3.5% on precision. On Ego4D, our method surpasses the best CNN model by +5.6% on F1, +4.5% on recall and +5.5% on precision. These results demonstrate the superiority of using a transformer-based architecture for egocentric gaze estimation as well as the effectiveness and robustness of our proposed method. We note that the improvement of our model is more prominent on Ego4D than EGTEA Gaze+. We speculate that this is because the Ego4D videos with gaze tracking data are captured under social interaction scenarios that contain interactions with both people and objects, and thus require the model to more heavily consider the global-local connections (*e.g.* the visual information about a social partner's gesture to an object) to predict the gaze. Another possible reason is that the Ego4D dataset has more samples to train the transformer-based model.
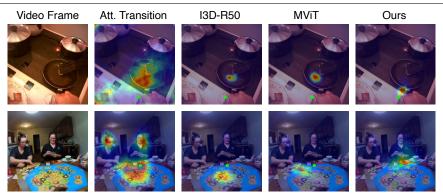
Figure 3: Visualization of gaze estimation. The first sample is from EGTEA Gaze+ and the second is from Ego4D. Estimated gaze is represented as a heatmap overlayed on input frames. Green dots denote the ground truth gaze location.
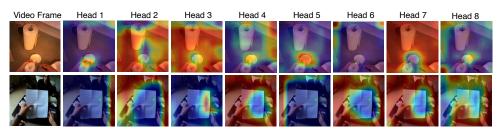

Figure 4: Visualization of the eight heads in global-local correlation module. The first sample is from EGTEA Gaze+ and the second is from Ego4D. Green dots denote gaze location.

4.3 Remarks

Visualization of Predictions. We visualize predictions of our model and other previous methods in Fig. 3. Attention transition [22] tends to overestimate gaze area which includes more uncertainty and ambiguity. I3D-R50 [17] and vanilla MViT [15] architectures run into failure modes when there are multiple objects and people in the scene. In contrast, our model, by explicitly modeling the connection between the global and local visual tokens, more robustly predicts the egocentric gaze distribution from the input video clip.

What has been learned by the Global-Local Correlation module? We additionally empirically analyze our proposed GLC module. We first calculate the correlation of the global token and each local token, and then normalize the calculated weights into a probabilistic distribution. A higher score suggests that the GLC captures a stronger connection between the particular local token and the global context. We reshape and upsample these weight distributions to form a heatmap, which we overlay with the original input. Since the GLC module applies a multi-head operation, we visualize the results from different heads in Fig. 4. Interestingly, the correlations captured by the GLC heads are quite diverse. We observe that the GLC module does assign different weights to local tokens, thereby capturing the different global-local connections for each token. Another important finding is that some heads learn to attend to background pixels to prevent the model from omitting important scene context. We provide further commentary in the supplementary.

5 Conclusion

In this paper, we develop a transformer-based architecture to address the task of estimating the camera wear's gaze fixation based only on egocentric video frames. Our key insight is that our global visual token embedding strategy, which encodes global visual information into the self-attention mechanism, and our global-local correlation (GLC) module, which explicitly reasons about the connection between global and local visual tokens, facilitate strong representation learning for egocentric gaze estimation. Our experiments on the EGTEA Gaze+ and Ego4D datasets demonstrate the effectiveness of our approach. We believe our work serves as an essential step in analyzing gaze behavior from egocentric videos and provides valuable insight into learning video representations with transformer architectures.

References

[1] Mohammad Al-Naser, Shoaib Ahmed Siddiqui, Hiroki Ohashi, Sheraz Ahmed, Nakamura Katsuyki, Sato Takuto, and Andreas Dengel. Ogaze: Gaze prediction in egocentric videos for attentional object selection. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2019.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[3] Giovanni Bellitto, Federica Proietto Salanitri, Simone Palazzo, Francesco Rundo, Daniela Giordano, and Concetto Spampinato. Hierarchical domain-adapted feature learning for video saliency prediction. *International Journal of Computer Vision*, 129 (12):3216–3232, 2021.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095*, 2(3):4, 2021.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[6] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. How is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing*, 29:2287–2300, 2019.

[7] Jiazhong Chen, Zongyi Li, Yi Jin, Dakai Ren, and Hefei Ling. Video saliency prediction via spatio-temporal reasoning. *Neurocomputing*, 462:59–68, 2021.

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[9] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398, 2018.

[10] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020.

[11] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2988–2997, 2021.

[12] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1601–1610, 2021.

[13] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2022.

[15] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[16] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34, 2021.

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[18] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[19] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. Group contextualization for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 928–938, 2022.

[20] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19, 2006.

[21] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.

[22] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018.

[23] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.

[24] Yifei Huang, Minjie Cai, and Yoichi Sato. An ego-vision system for discovering human joint attention. *IEEE Transactions on Human-Machine Systems*, 50(4):306–316, 2020.

[25] Sen Jia and Neil DB Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and vision computing*, 95:103887, 2020.

[26] Lai Jiang, Yifei Li, Shengxi Li, Mai Xu, Se Lei, Yichen Guo, and Bo Huang. Does text attract attention on e-commerce images: A novel saliency prediction dataset and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2088–2097, 2022.

[27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[28] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[29] Apoorv Khattar, Srinidhi Hegde, and Ramya Hebbalaguppe. Cross-domain multi-task learning for object detection and saliency estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3639–3648, 2021.

[30] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[31] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020.

[32] Youngwan Lee, Jonghee Kim, Jeff Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

[33] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[34] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pages 3216–3223, 2013.

[35] Yin Li, Miao Liu, and James M Rehg. 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)*, pages 619–635, 2018.

[36] Yin Li, Miao Liu, and Jame Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[37] Miao Liu, Xin Chen, Yun Zhang, Yin Li, and James M Rehg. Attention distillation for learning video representations. *British Machine Vision Conference*, 2019.

[38] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–721. Springer, 2020.

[39] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4722–4732, 2021.

[40] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[41] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Visual saliency prediction using transformers. *arXiv preprint arXiv:2110.03593*, 2021.

[42] Cheng Ma, Haowen Sun, Yongming Rao, Jie Zhou, and Jiwen Lu. Video saliency forecasting transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[43] Jeff J MacInnes, Shariq Iqbal, John Pearson, and Elizabeth N Johnson. Wearable eye-tracking for research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices. *BioRxiv*, page 299925, 2018.

[44] Si-Ahmed Naas, Xiaolan Jiang, Stephan Sigg, and Yusheng Ji. Functional gaze prediction in egocentric video. In *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, pages 40–47, 2020.

[45] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021.

[46] Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2192–2201, June 2022.

[47] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[48] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[49] Hyun Soo Park and Jianbo Shi. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2015.

[50] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[51] Yubao Sun, Mengyang Zhao, Kai Hu, and Shaojing Fan. Visual saliency prediction using multi-scale attention gated network. *Multimedia Systems*, 28(1):131–139, 2022.

[52] Hamed Rezazadegan Tavakoli, Esa Rahtu, Juho Kannala, and Ali Borji. Digging deeper into egocentric gaze prediction. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 273–282. IEEE, 2019.

[53] Sanket Kumar Thakur, Cigdem Beyan, Pietro Morerio, and Alessio Del Bue. Predicting gaze from egocentric social interaction videos and imu data. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 717–722, 2021.

[54] Antigoni Tsiami, Petros Koutras, and Petros Maragos. Stavis: Spatio-temporal audio-visual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4766–4776, 2020.

[55] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021.

[56] Jue Wang and Lorenzo Torresani. Deformable video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14053–14062, 2022.

[57] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017.

[58] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[59] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12249–12256, 2020.

[60] Ziqiang Wang, Zhi Liu, Gongyang Li, Yang Wang, Tianhong Zhang, Lihua Xu, and Jijun Wang. Spatio-temporal self-attention network for video saliency prediction. *IEEE Transactions on Multimedia*, 2021.

[61] Xinyi Wu, Zhenyao Wu, Jinglin Zhang, Lili Ju, and Song Wang. Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12410–12417, 2020.

[62] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.

[63] Alfred L Yarbus. *Eye movements and vision*. Springer, 2013.

[64] Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D Abowd, and James M Rehg. Detecting eye contact using wearable eye-tracking glasses. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 699–704, 2012.

[65] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4372–4381, 2017.

[66] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12083–12093, 2022.

[67] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

Supplementary

This is supplemental material for the paper titled "In the Eye of Transformer: Global-Local Correlation for Egocentric Gaze Estimation". We organize the content in the following way:

- A – Data Processing

- B – Implementation Details

- C – Experiments on Action Recognition

- D – Details of Different Global Visual Embedding Strategies

- E – More Visualization Examples of Gaze Estimation

- F – Future Work

A   Data Processing

At training time, we randomly sample 8 frames from each video with a sampling interval of 8 as input (*i.e.* selecting 8 frames from a 72-frame window with equal spacing). All videos are spatially downsampled to 256 in height while keeping the original aspect ratio. We further implement multiple data augmentations including random flipping, shifting, and resizing. We then randomly crop each frame to get an input with dimensions $8 \times 256 \times 256$. The output from the decoder is a downsampled heatmap with dimension $8 \times 56 \times 56$. For visualization, the output heatmap is upsampled to match the input size by trilinear interpolation. At inference time, the input clip is center-cropped. For gaze labels, we generate a gaussian kernel centered at the gaze location in each input frame with a kernel size of 19 following [10]. We use a uniform distribution for frames where gaze is not tracked in training and only calculate metrics on frames with fixated gaze in testing as in [35]. For the EGTEA Gaze+ [35] dataset, we determine which frames to calculate metrics on by using the provided label of gaze fixations and saccades. On the Ego4D [18] dataset, no label of gaze type is available. We calculate the euclidean spatial distance of gaze between adjacent frames and consider the tracked gaze to be a saccade if the distance is above a threshold, and treat it as fixation otherwise. We adopt an empirical threshold of 40.

B   Implementation Details

We show the parameter details of each layer in Table 5. Data is input to the local token embedding module to get local tokens. Then, these tokens are fed to the global token embedding module which consists of three convolutional layers and one linear layer. Both local and global tokens are flattened into vectors of length of 96. In the following encoder, Global-Local Correlation Module (GLC), and decoder blocks, the number of local tokens is either downsampled or upsampled, while the number of global tokens remains as one. Hence we write the number of tokens in the output size as *(1 global token + number of local tokens)*.

Methods	Cls Token	Pooling	Top1-Acc	Top5-Acc	Mean Cls Acc
MViT [15]	✓		64.64	89.22	54.02
MViT [15]		✓	63.45	88.72	55.34
MViT + Global Token	✓		64.44	88.72	55.28
MViT + Global Token		✓	63.06	88.53	54.15
MViT + Global Token + GLC	✓		64.79	88.67	56.77
MViT + Global Token + GLC		✓	65.33	**89.12**	**57.26**

Table 4: Results of action recognition on EGTEA Gaze+. We implemented two methods for classification – adding an additional class token or using global average pooling. "-"" means the result is unavailable. The complete models are highlighted.

After generating the output from decoder block4, a convolutional layer is applied only on the local tokens to compress the 8 channels to 1. We then convert this to a probability distribution by applying softmax to each frame.

C   Experiments on Action Recognition

In addition to egocentric gaze estimation in the main paper, we also examine the application of our GLC module to the egocentric video action recognition task, and find that our method performs competitively with methods designed specifically for this task on EGTEA Gaze+. To this end, we remove the decoder in the gaze estimation model and keep only the visual token embedding, transformer encoder, and GLC modules. Generally, there are two ways to obtain activity class category prediction: adding a class embedding token at the first layer of transformer, or using pooling across all global tokens to obtain a final embedding. Then a fully-connected layer followed by softmax is used to predict probabilities for each category. We implement both strategies and compare our approaches with previous works in Table 4. We conduct these experiments only on EGTEA Gaze+ [35] using the same split as gaze estimation. Note that the Ego4D [18] social benchmark does not contain action labels.

For vanilla MViT [15], class token embedding performs better than the pooling operation. For both methods, simply adding global embedding has a minor influence on the overall performance (-0.2% on top1 accuracy, -0.5% on top5 accuracy and $+1.32\%$ on mean class accuracy while using the class token, and -0.39%, on top1 accuracy, -0.19% on top5 accuracy and -1.19% on mean class accuracy while using pooling layer). This result suggests that simply adding global context as an additional token has minor influence on the action recognition performance.

In addition, adding our GLC module can only improve the model performance by a small margin when using class token embedding to predict action classes. We hypothesize that this is because only the class token is input into the linear layer for final prediction and re-weighted tokens from GLC are left unused. In contrast, when applying global average pooling on all local tokens, GLC improves top1, top5 and mean class accuracy over the counterpart that doesn't use GLC (*MViT+Global Token*) by $+2.27\%$, $+0.59\%$ and $+3.11\%$, respectively. Gains over corresponding the MViT baseline are $+1.88\%$, $+0.4\%$ and $+1.92\%$ on the three metrics. These results indicate our proposed GLC module is a robust and general design that also improves the action recognition performance. However, the impact on action recognition is smaller compared with egocentric gaze estimation.

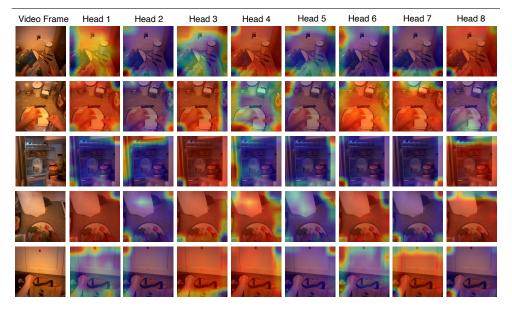We note that our model achieves a competitive performance for action recognition on

Figure 5: Visualization of the eight heads in global-local correlation module for action recognition.

EGTEA Gaze+ without additional design for this specific task. Our top1 accuracy of 65.33% exceeds Wang et al. (2020) [59] by +1.23%, and is only a −1.17% difference from Hao et al. (2022)'s [19] recent state-of-the-art method for this benchmark of 66.5%. We also want to emphasize that we conduct these action recognition experiments to demonstrate the generalization ability of our proposed GLC module rather than aim to produce SOTA results on action recognition.

Additionally, we visualize the global-local correlation weights of the GLC in Fig. 5. Importantly, the learned global-local correlation is vastly different from the gaze distribution when the model is trained for action recognition; in contrast, a stronger connection between the learned global-local correlation and gaze distribution can be observed when the model is trained for gaze estimation (see Fig. 8). How to design a weakly-supervised model for egocentric gaze estimation remains an open question.

D   Details of Different Global Visual Embedding Strategies

We present further details of the four global visual embedding strategies we studied in Section 4.2 of the main paper. As demonstrated in Fig. 6, (a) implements max pooling on input frames directly, and (b) implements max pooling on local visual tokens. For (c) and (d), we replace max pooling operations in (a) and (b) with a sequence of convolutional layers. The specific parameters of (d) are detailed in Table 5. For global embedding in (c), input video frames are fed into a convolutional layer that is identical to the layer used for local token embedding (i.e., kernel is $3 \times 7 \times 7$ and stride is $2 \times 4 \times 4$.) Then, the output is passed to a sequence of convolutional layers identical to (d).
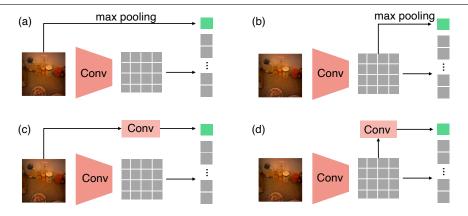
Figure 6: Four different approaches of global visual token embedding.

E More Visualization Examples of Gaze Estimation

More visualizations of gaze prediction of both our model and previous state-of-the-art approaches are presented in Fig. 7. Our proposed model can accurately predict the gaze distribution even when the scene context is very complicated, while the other three approaches may be misled by background objects or produce predictions with too much uncertainty.

We provide more examples of GLC visualizations in Fig. 8. The 8 heads capture features of different areas which is consistent with the examples in the main paper. On the EGTEA Gaze+ dataset, the maps produced by heads 1, 4, 5, and 8 highlight pixels around the gaze point with different uncertainty (which is illustrated by the size of highlighted area). The other four heads focus on surrounding objects and leave gaze areas unattended. As for the Ego4D data, only head 3 captures the wearers' attention, while the other heads fully focus on the backgrounds in different aspects. This supports our key conclusion in the main paper that our GLC module learns to model human attention by setting different weights from local to global tokens, capturing many facets of scene information (both around the gaze target and in the background) in the multi-headed attention mechanism.

F Future Work

In this paper, we studied the explicit integration of global scene context for egocentric gaze estimation and proposed a novel modeling approach for this problem. We also showed the results of our proposed architecture on egocentric action recognition in this supplementary material to demonstrate our model's generalization ability. Our findings also point to several exciting future research directions:

- Our proposed GLC module has the potential to address other video understanding tasks including visual saliency prediction in third-person video, active object detection, and future forecasting. We plan to study the effect of our method on those tasks in our future work.

- Our modeling work can be expanded to understanding human gaze behavior associated with multiple sensing modalities, especially in the social conversation setting. An
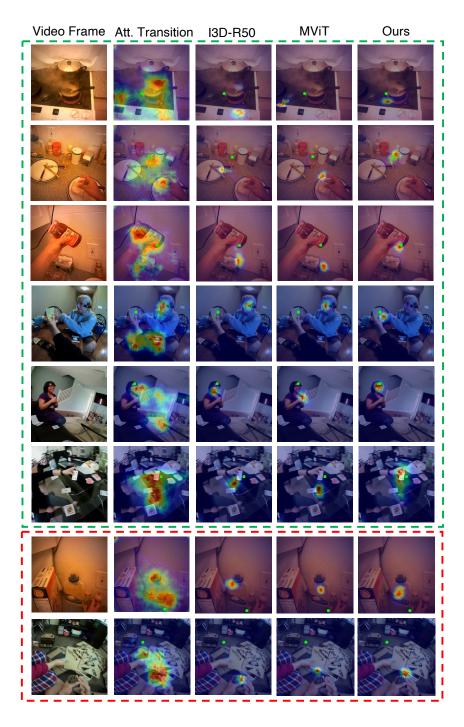
Figure 7: Visualization of gaze estimation. Both successful cases (in green box) and failure cases (in red box) of our model are demonstrated. Green dots present ground truth.
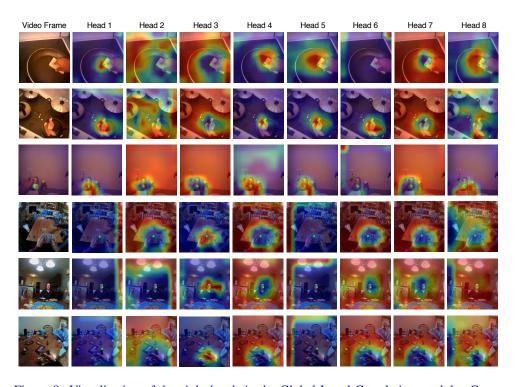
Figure 8: Visualization of the eight heads in the Global-Local Correlation module. Green dots represent the ground truth.

exciting future direction is incorporating audio signals into egocentric gaze estimation.

- Our proposed GLC fails to learn the gaze distribution when the model is trained to predict the action labels. How to design a weakly supervised model for egocentric gaze estimation using action labels is an interesting problem.

Stages	Operators			Output Size
data	-			$8 \times 256 \times 256$
local token embedding	$Conv(3 \times 7 \times 7,\ 96)$ $stride\ 2 \times 4 \times 4$			$96 \times 4 \times 64 \times 64$
global token embedding	$Conv(3 \times 3 \times 3,\ 96)$ $Conv(3 \times 3 \times 3,\ 96)$ $Conv(3 \times 3 \times 3,\ 96)$ $Linear(24576)$ $stride\ of\ each\ conv\ 1 \times 2 \times 2$			96×1
tokenization	flattening and concatenation			$96 \times (1 + 4 \times 64 \times 64)$
encoder block1		$MSA(96)$ $MLP(384)$	$\times 1$	$192 \times (1 + 4 \times 64 \times 64)$
encoder block2		$MSA(192)$ $MLP(768)$	$\times 2$	$384 \times (1 + 4 \times 32 \times 32)$
encoder block3		$MSA(384)$ $MLP(1536)$	$\times 11$	$768 \times (1 + 4 \times 16 \times 16)$
encoder block4		$MSA(768)$ $MLP(3072)$	$\times 2$	$768 \times (1 + 4 \times 8 \times 8)$
global-local correlation	$GLC(768)$ $MLP(3072)$ $concatenation\ in\ channel$		$\times 1$	$1536 \times (1 + 4 \times 8 \times 8)$
decoder block1		$MSA(1536)$ $MLP(3072)$	$\times 1$	$768 \times (1 + 4 \times 16 \times 16)$
decoder block2		$MSA(768)$ $MLP(1536)$	$\times 1$	$384 \times (1 + 4 \times 32 \times 32)$
decoder block3		$MSA(384)$ $MLP(768)$	$\times 1$	$192 \times (1 + 4 \times 64 \times 64)$
decoder block4		$MSA(192)$ $MLP(384)$	$\times 1$	$96 \times (1 + 8 \times 64 \times 64)$
head	$Conv(1 \times 1 \times 1,\ 1)$ $stride\ 1 \times 1 \times 1$			$8 \times 64 \times 64$

Table 5: Architecture of the proposed model. Convolutional layers are denoted as *Conv(kernel size, output channels)*. Numbers of input channels of multi-head self-attention are shown in the parenthesis of *MSA*. Dimensions of the hidden layer in multi-layer perceptrons are listed in parenthesis of *MLP*. In tokenization, local and global tokens are reshaped and concatenated. In global-local correlation, the output is concatenated with its input in the channel dimension. Head only takes local tokens as input.