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

Bolin Lai, Miao Liu, Fiona Ryan, James M. Rehg

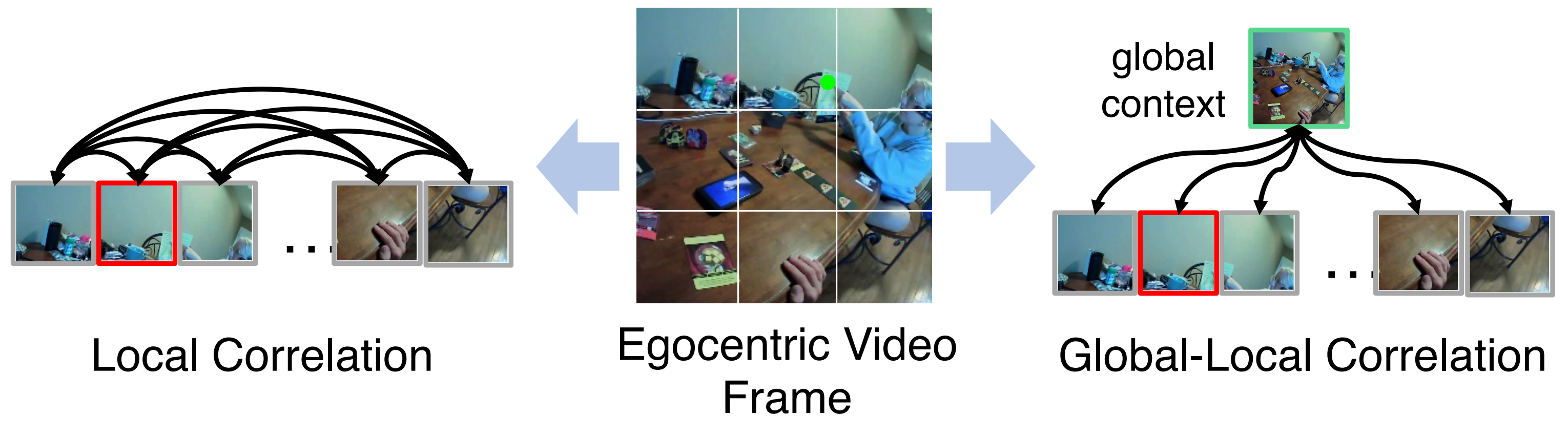
Georgia Institute of Technology

BMVC 2022

Motivation

- Egocentric gaze implies human's attention in daily activities, which is critical for applications in AR/VR.
- There are too many items disturbing our prediction in a complicated scene.
- Global-local correlation is not well captured in self-attention mechanism.

How do you know where I am looking at?



Gaze estimation in a holistic view:

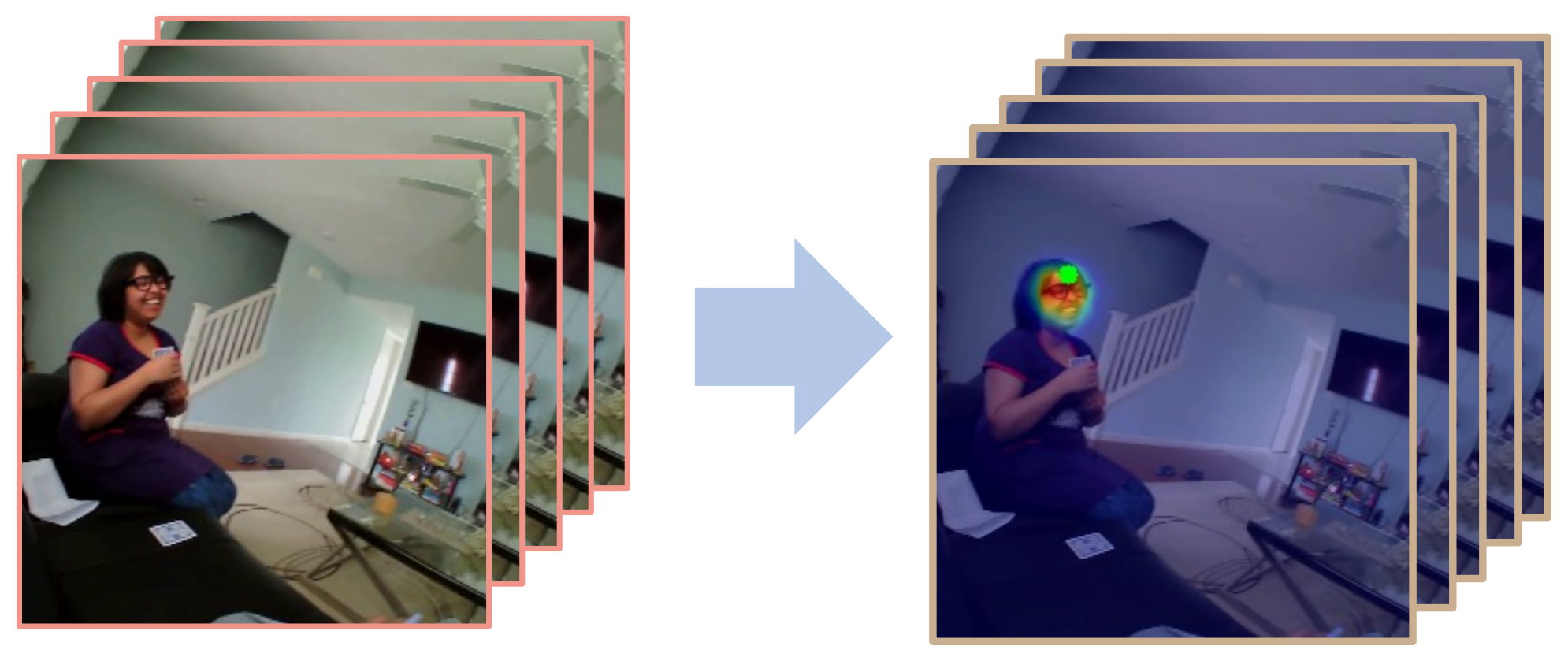
Another person is pointing at and looking at the sheet she holds.

Correlations across only local visual tokens are insufficient to distinguish gaze fixation in complex background!

Objective

Input: egocentric video sequence

Output: gaze prediction in each frame (heatmap)



Gaze labels are captured by eye trackers.

Challenge:

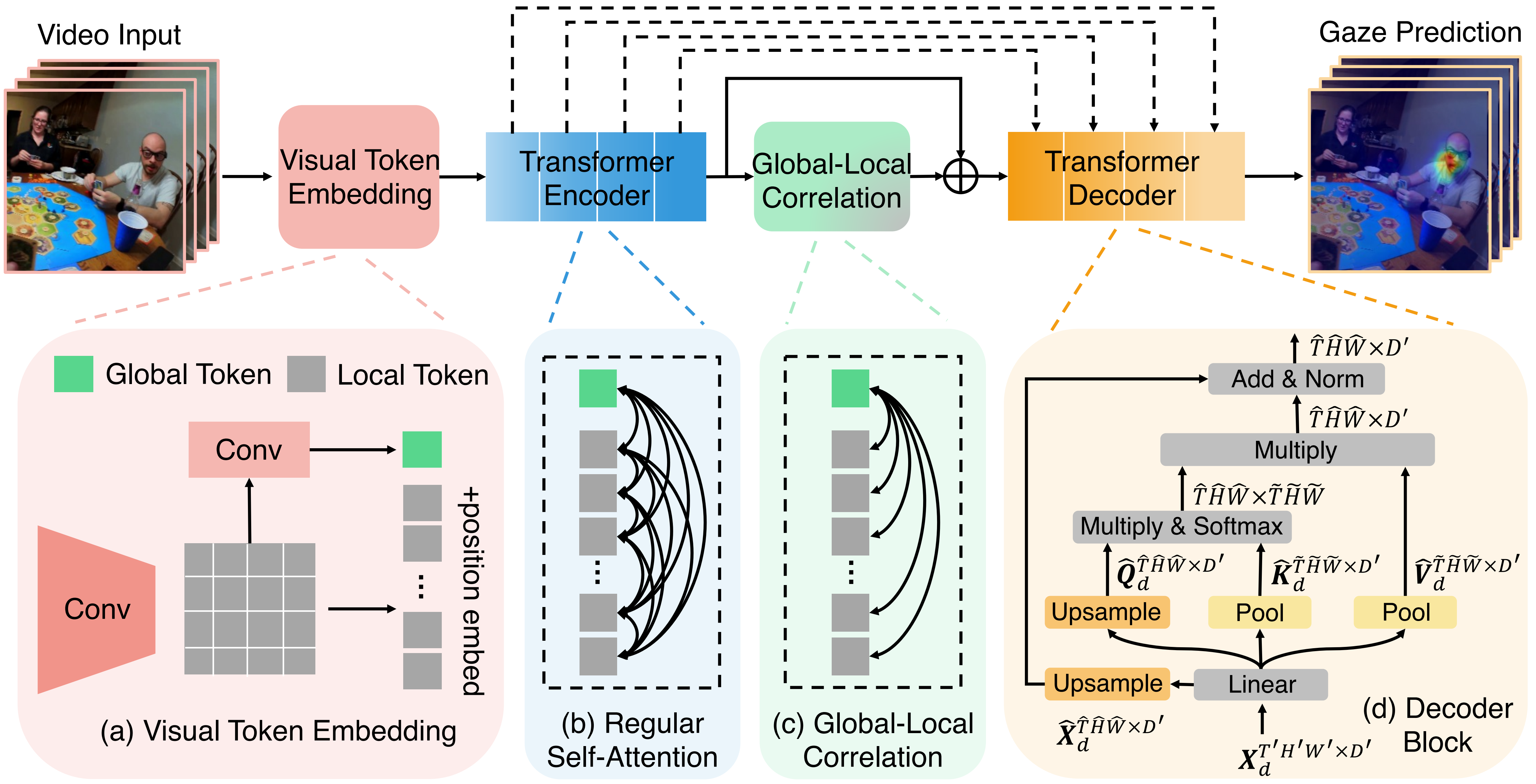
- Gaze can move fast and background changes drastically.
- We need to integrate cues from a global view into a holistic analysis of visual attention.

Key Idea:

- Encoding the global context into an additional token.
- Highlighting the correlation of global token and each local token in a specifically designed module.

Contact:

Overview of Architecture


Video Input
Visual Token Embedding
Transformer Encoder
Global-Local Correlation
Transformer Decoder
Gaze Prediction
Global Token
Local Token
Conv
Conv
+position embed
(a) Visual Token Embedding
(b) Regular Self-Attention
(c) Global-Local Correlation
$\hat{T}\hat{H}\hat{W}\times D'$
Add & Norm
$\hat{T}\hat{H}\hat{W}\times D'$
Multiply
$\hat{T}\hat{H}\hat{W}\times\tilde{T}\tilde{H}\tilde{W}$
Multiply & Softmax
$\hat{Q}_d^{\hat{T}\hat{H}\hat{W}\times D'}$
$\hat{K}_d^{\tilde{T}\tilde{H}\tilde{W}\times D'}$
$\hat{V}_d^{\tilde{T}\tilde{H}\tilde{W}\times D'}$
Upsample
Pool
Pool
Upsample
Linear
$\hat{X}_d^{\hat{T}\hat{H}\hat{W}\times D'}$
$X_d^{T'H'W'\times D'}$
(d) Decoder Block

Regular Self-Attention vs. Global-Local Correlation

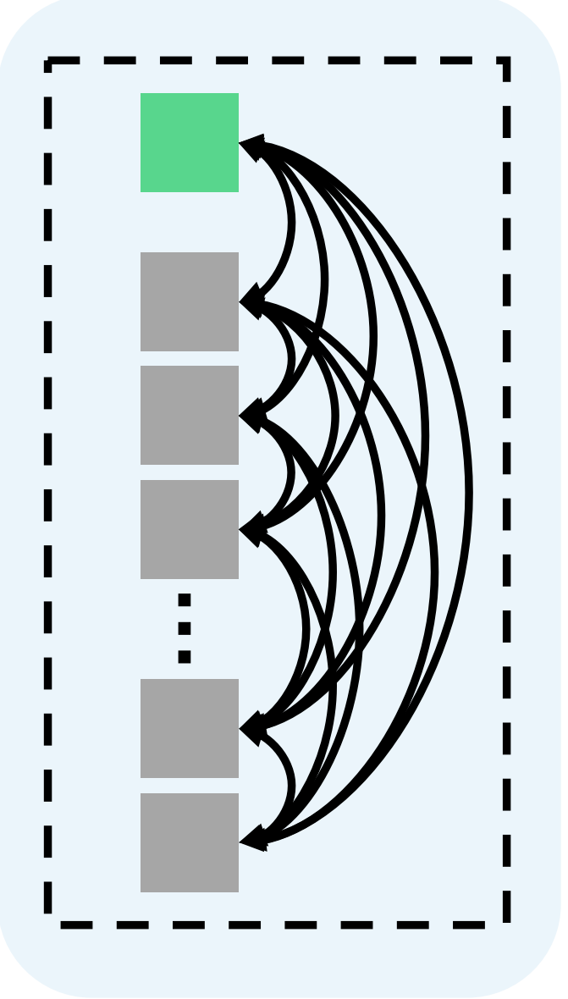
Input: $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{N+1} \end{bmatrix} \in \mathbb{R}^{(N+1)\times D}$

global token

local tokens

$$[\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}] = Linear(\boldsymbol{X}) \quad (\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} \in \mathbb{R}^{(N+1)\times D})$$

Linearly map each token x_i to query, key and value vectors.



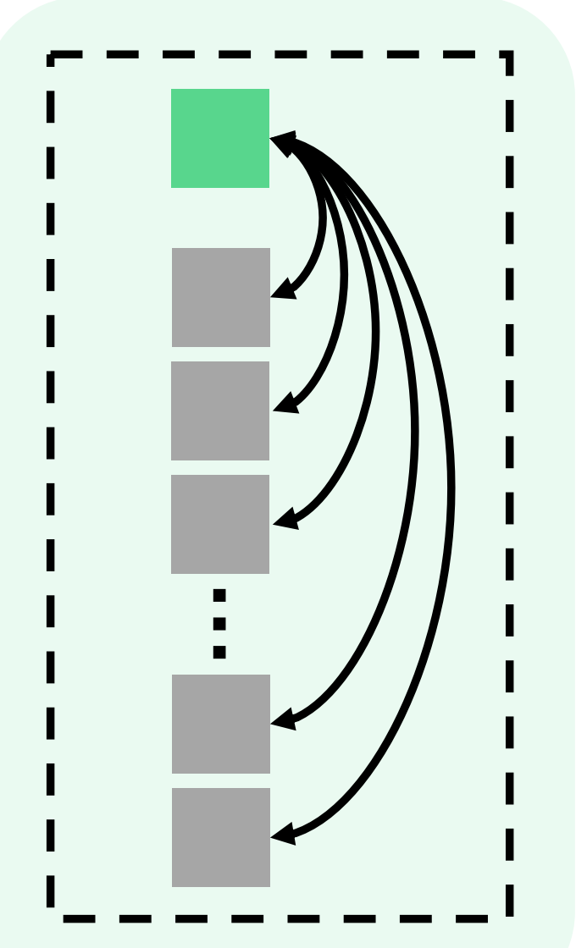
Regular Self-attention:

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

Global-Local Correlation:

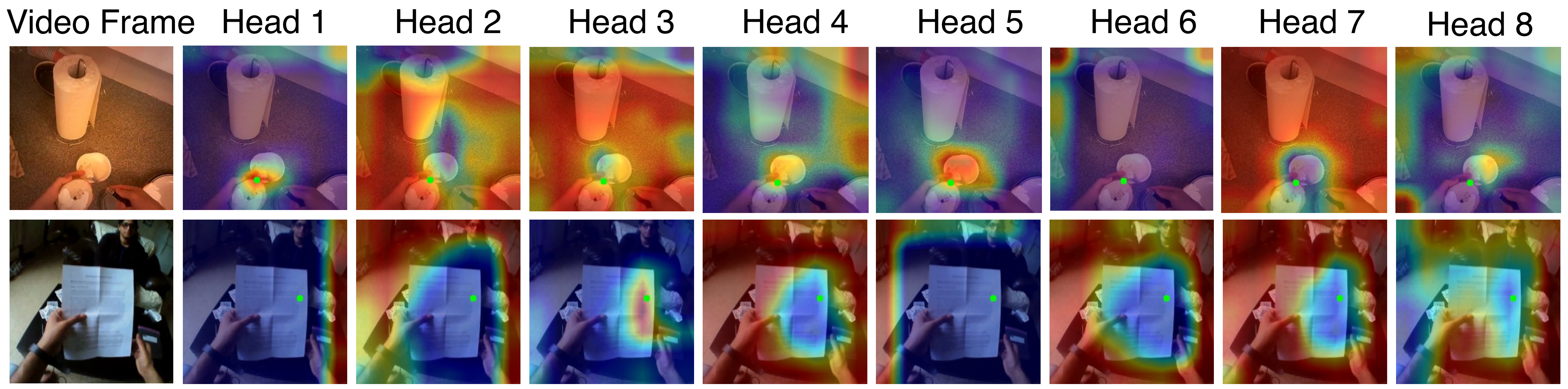
$$\boldsymbol{S}^{(N+1)\times(N+1)} = [s_{i,j}] \quad s_{i,j} = \begin{cases} 0, & if\ i=j\ or\ j=1 \\ 10^8, & otherwise \end{cases}$$

$$GLC(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left((\boldsymbol{Q}\boldsymbol{K}^T-\boldsymbol{S})/\sqrt{D}\right)\boldsymbol{V}$$

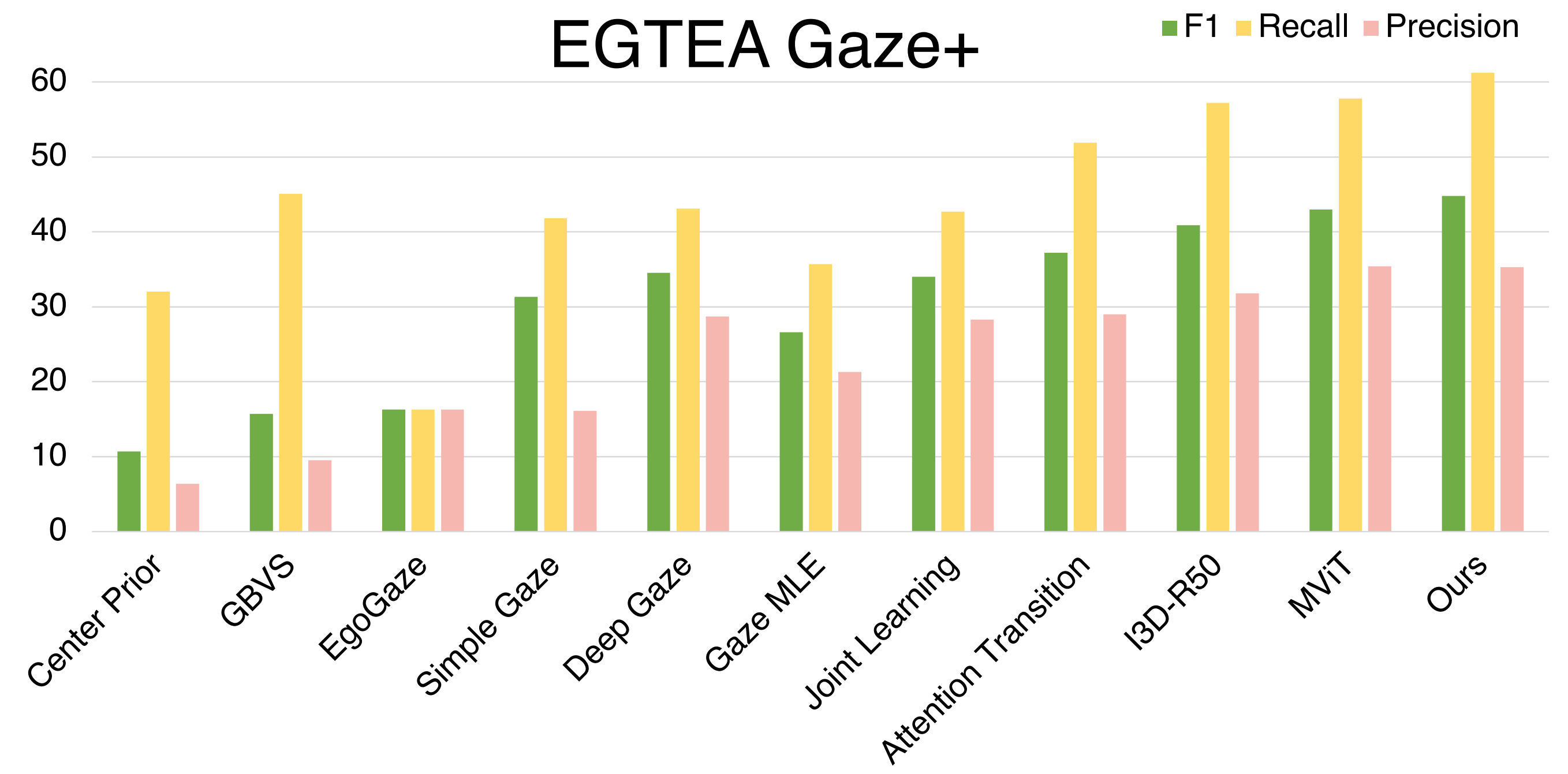


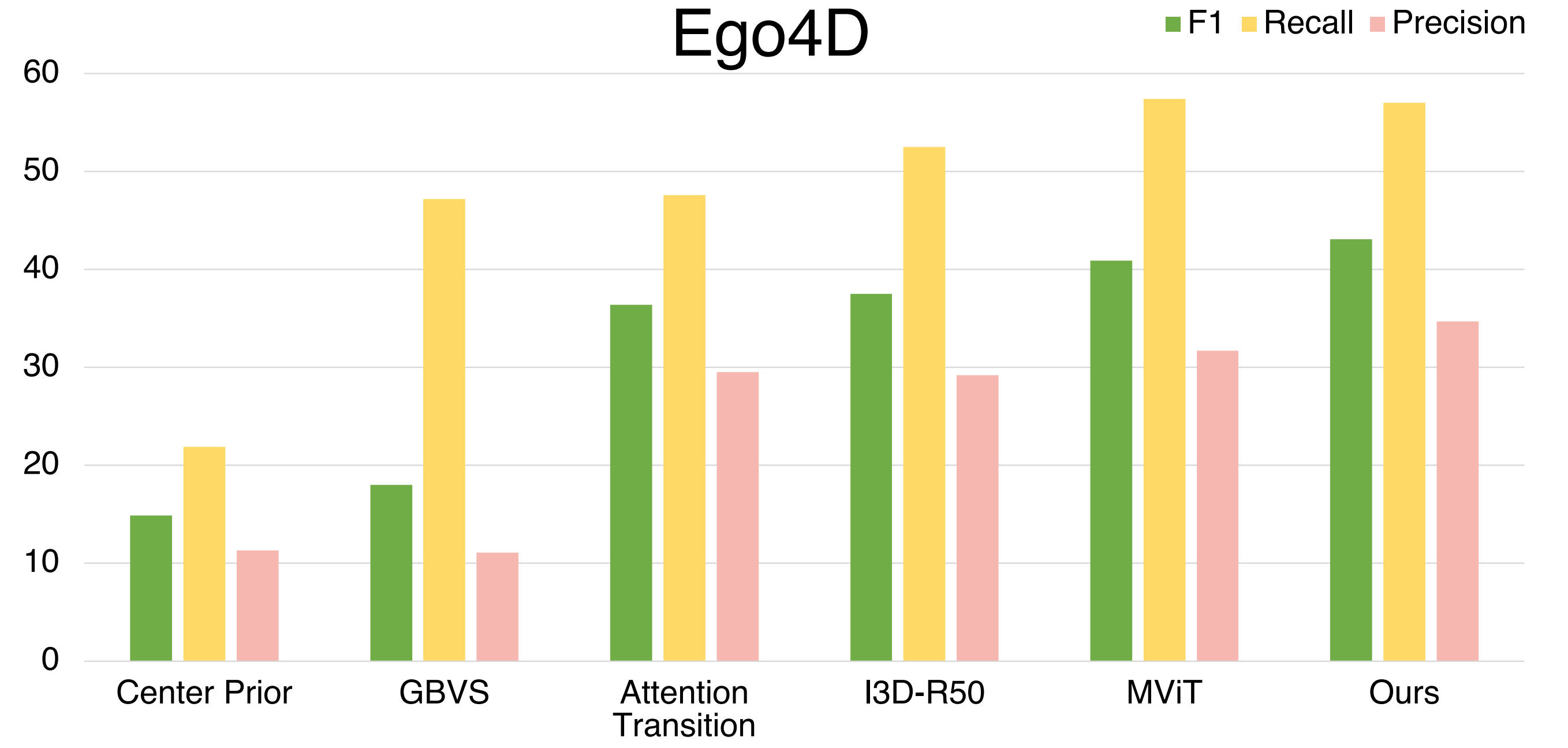
Key point: Correlations across local tokens are ignored by subtracting a suppression matrix.

Visualization of Attention in GLC

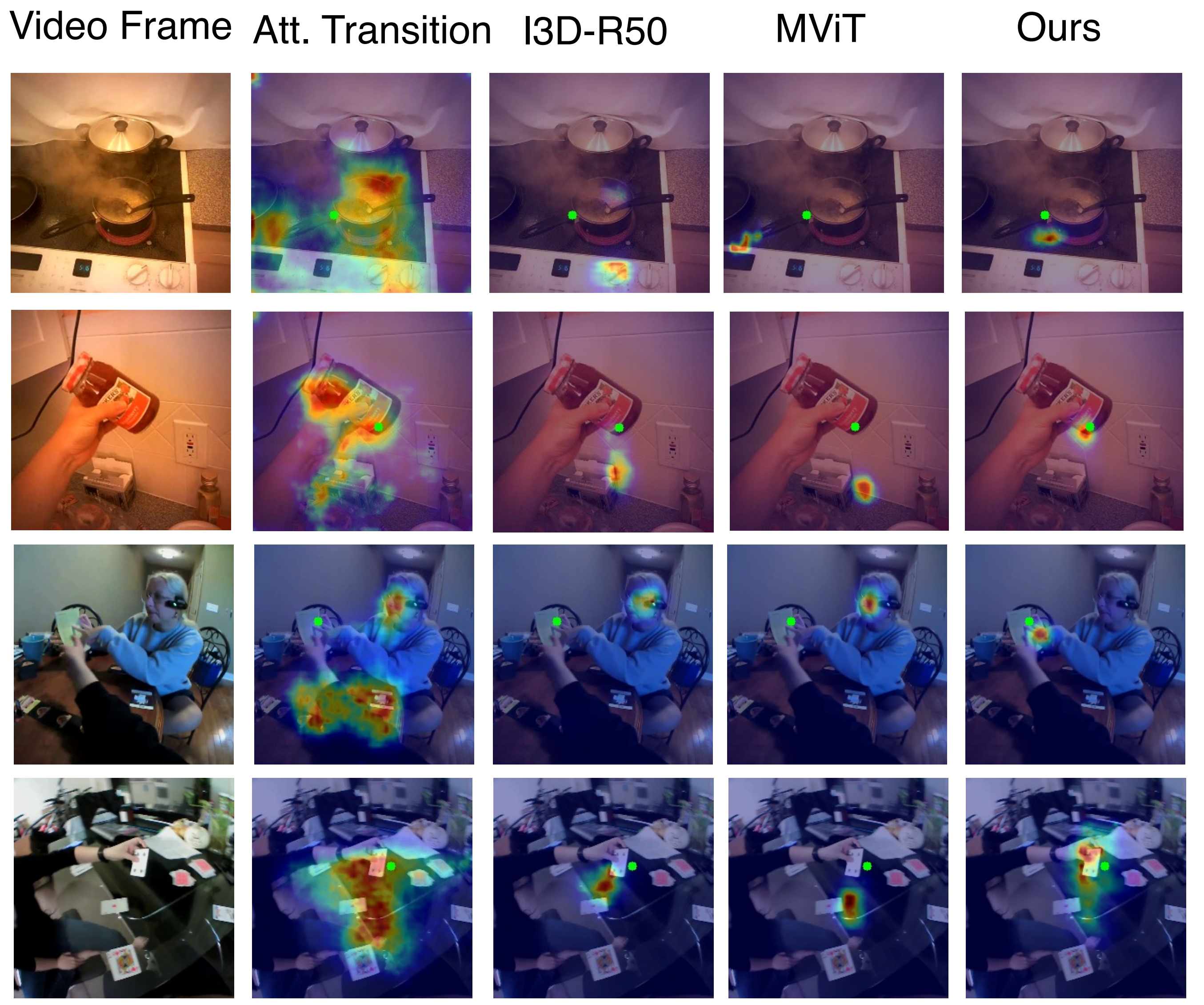


Experiments and Results





Visualization of Prediction



Conclusion

We develop the first transformer-based model for gaze estimation on egocentric videos. Our proposed method facilitates strong gaze representation learning and achieves new state of the art.

Acknowledgments. Portions of this project were supported by a gift from Meta.