Action-Segmentation based Gaze Anticipation on Egocentric Video

Sebin Lee Jaewoo Park Jiwon Lee Yerang Mok
Seoul National University
{wallen, 1qkrwodn1, zzwon1212, yelangmog}@snu.ac.kr

Abstract

Forecasting human gaze in egocentric videos provides a window into attention dynamics during daily activities. While recent methods have shown progress by leveraging visual features from RGB frames, they often fail to model the semantic intent underlying gaze behavior, such as the observer's ongoing actions. As a result, gaze predictions tend to be spatially imprecise and biased toward the image center. In this paper, we propose an ***Action-Segmentation based*** *framework that anticipates future gaze by conditioning on both visual and semantic cues. Our model is composed of four key modules: (1) a transformer-based visual encoder, (2) an action feature extractor that predicts verb/noun and produces an activity-aware representation, (3) a segmentation module that extracts gaze-aligned object masks using SAM, and (4) a decoder that fuses visual and semantic features to generate gaze heatmaps. We validate our model on the EGTEA Gaze+ dataset and show that action-segmentation based modeling significantly enhances egocentric gaze forecasting performance.*

1. Introduction

Egocentric vision understanding has emerged as a novel and challenging research field in computer vision with the rapid development of wearable devices. Unlike conventional third-person vision, first-person visual data acquired via cameras or sensors worn on human body offer a unique perspective that reflects human visual experiences [11]. Such research can be useful not only in augmented reality (AR), virtual reality (VR), and human–computer interaction (HCI), but also in the emerging research domain of embodied AI where artificial agents interact with the physical world, coupled with robotics.

Gaze represents distinctive information, inherently present in egocentric vision. A person's eye movements, which reflect the observer's intentions and goals, are essential for understanding egocentric video data, as the video frames themselves change according to the observer's head and body movements driven by these intentions. In particular, the problem of *egocentric gaze anticipation*, predicting where gaze will move in future frames of first-person videos, can elicit a more comprehensive understanding of the relationship between egocentric scene and gaze behavior than a rather simpler gaze estimation task in the current frame. In addition, gaze anticipation enables predictive computation and is beneficial for many applications [23]. For instance, in virtual reality environments, predicting the direction of a user's gaze movement can be used to pre-load content that the user is likely to focus on, thereby reducing latency in VR rendering processes. However, the gaze anticipation task is largely understudied due to the complexity of egocentric scenes and the dynamic nature of gaze behaviors [10].

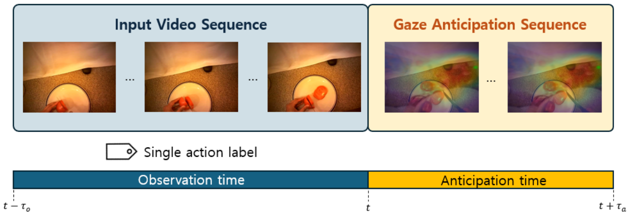

Figure 1. The problem setting of egocentric gaze anticipation. During the observation time (1-3 seconds), the model receives video frames and a single action label. The model predicts two outcomes for the anticipation time (1 second): (1) whether the current action will continue, and (2) the gaze distribution as a heatmap.

We argue that both action-based and segmentation-guided cues are essential for egocentric gaze anticipation. Gaze patterns are highly dependent on the observer's ongoing task. During free-viewing scenarios such as environmental exploration or passive observation, gaze tends to follow bottom-up visual saliency, gravitating toward the most visually prominent stimuli. In contrast, in task-driven settings — e.g., cooking or object manipulation — gaze typically shifts toward task-relevant objects that are semantically associated with the next intended action.

Prior works have primarily relied on bottom-up saliency cues extracted from RGB inputs. However, such approaches have shown limited effectiveness, especially under conditions of rapid gaze transitions or when multiple salient regions are present [18]. These limitations highlight the need for models that incorporate higher-level semantic understanding.

To this end, we propose a model that jointly leverages two complementary sources of information: (1) gaze-guided segmentation masks that provide spatially localized, semantic representations of potential gaze targets, and (2) action-aware features that encode the observer's behavioral intent. This dual semantic conditioning enables more accurate gaze forecasting by capturing both the 'where' and 'why' behind gaze behavior. Our approach is particularly effective in complex egocentric scenes, where saliency alone fails to explain human attention.

Our contributions are summarized as follows:

- We introduce the first gaze anticipation framework that fuses action semantics and gaze-guided object segmentation to model both behavioral intent and observed object as well as visual saliency in egocentric videos.
- Our model successfully enhances gaze forecasting accuracy in scenarios with complex interactions and rapid gaze transitions by jointly leveraging high-level semantic cues.

2. Related Work

Egocentric Gaze Modeling. In understanding human gaze behavior in egocentric videos, the majority of previous research has focused on gaze estimation, which aims to infer the gaze point in the current video frame. Early research started from saliency prediction using handcrafted features [12, 20, 21]. As entering the deep learning era, gaze estimation models leveraging CNNs [6, 7, 19], LSTMs, and Transformers [9, 14] have been developed. Huang *et al.* [6] developed a framework for learning temporal attention shifts from video features that capture significant gaze movements. Lai *et al.* [9] explicitly formulated global-local relationships within visual embeddings for gaze estimation.

In contrast, egocentric gaze anticipation, which aims to forecast future gaze targets based on previous video frames, addresses a relatively unexplored aspect of gaze modeling. Currently there are only few gaze anticipation models in the literature. Zhang *et al.* [23] first introduced a novel challenge defined as gaze anticipation in egocentric videos and proposed a Generative Adversarial Network (GAN [4]) model called Deep Future Gaze (DFG) to generate future frames, adopting a 3D-CNN architecture. In the follow-up study [24], they improved their model by expanding to a dual-branch structure. Yun *et al.* [22] introduced the Multisensory Spherical World-Locked Transformer (MuST) framework, which transforms audiovisual information relative to head pose, thereby compensating for self-motion effects and improving the accuracy of gaze anticipation. Recently, Lai *et al.* [10] proposed a Contrastive Spatial-Temporal Separable (CSTS) approach, utilizing both video and audio modality for the first time in egocentric gaze understanding and achieved state-of-the-art performance on two egocentric video dataset, Ego4D [5] and AEA [16]. All previous egocentric gaze anticipation models learn only from the bottom-up sensory properties. In this work, we introduce our model which leverages another crucial factor influencing gaze behavior, the observer's action.

Gaze Understanding and Action. The relationship between gaze patterns and actions was explored in several previous works. Borji *et al.* [1] established a direct mapping between low-level visual features and motor actions derived from top-down processes in driving simulation. Fathi *et al.* [3] proposed a probabilistic generative model which uses verb-noun pairs describing actions as a prior for gaze. Huang *et al.* [7] and Li *et al.* [13] jointly modeled gaze and action using CNNs, thereby constructed a unified framework for gaze estimation and action recognition. These studies have all been conducted in the gaze estimation task, and no paper has explicitly investigated the effect of action-related information on performance in the gaze anticipation task. Therefore, we will examine whether action-related information also helps to predict where gaze will be directed in the future.

3. Method

3.1. Overview

An overview of the full architecture is illustrated in Fig. 2. Our proposed method aims to forecast future gaze locations from egocentric video by integrating both visual and semantic cues. The architecture is composed of four main components: (1) a transformer-based visual encoder, (2) an action feature extractor that predicts verb/noun and produces an activity-aware representation, (3) a segmentation module that extracts gaze-aligned object masks using SAM, and (4) a transformer-based decoder that fuses visual and semantic features to generate gaze heatmaps. The input to our model is a sequence of egocentric RGB frames $\{I_t\}_{t=0}^{T}$ captured over a variable temporal range between 1 and 3 seconds and ground-truth. Regardless of the input duration, we uniformly sample 8 frames to represent the temporal span, assuming this subsampling retains sufficient information for future gaze prediction. Our goal is to predict the user's gaze distribution $\hat{G}$ over the upcoming one-second interval in the form of a dense spatial heatmap.

3.2. Network Architecture

Spatiotemporal Feature Encoder. We adopt the Multiscale Vision Transformer (MViT) [2] to encode egocentric video clips. MViT captures both spatial and temporal fea-

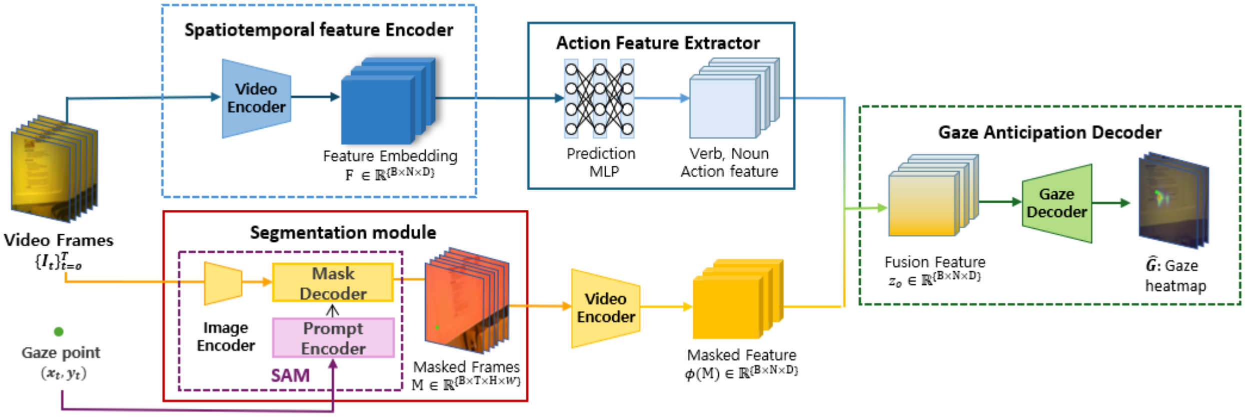

Figure 2. Overview of the proposed model. Given egocentric video frames $\{I_t\}_{t=0}^{T}$ and gaze point (x_t, y_t), our model extracts both action semantics and object-centric features to forecast future gaze. The video encoder first computes spatiotemporal embeddings $\mathbf{F} \in \mathbb{R}^{B\times N\times D}$. The action feature extractor generates action-aware features via a prediction MLP, while the segmentation module applies SAM to generate spatial masks conditioned on gaze cues. The masked video frames are encoded again to produce $\phi(\mathbf{M})$, and the fused features are decoded to produce the gaze heatmap $\hat{G}$.

tures across multiple scales, making it well-suited for dynamic first-person scenarios. Each input clip is uniformly subsampled to 8 frames to balance computational efficiency and temporal context, under the inductive bias that short (under 3 seconds) egocentric sequences are sufficient for downstream tasks.

Given an input video clip $\mathbf{X} \in \mathbb{R}^{B\times T\times H\times W\times 3}$, the encoder produces patch-level feature embeddings:

$$\mathbf{F} = \phi(\mathbf{X}) \in \mathbb{R}^{B\times N\times D}$$

where B is the batch size, T the number of frames, $H \times W$ the spatial resolution, N the number of spatiotemporal tokens, and D the embedding dimension.

Action Feature Extractor. This module infers the semantic activity context from the encoded visual representations. A lightweight MLP-based head predicts the verb and noun associated with the user's ongoing action:

$$\hat{v} \in \mathbb{R}^{B\times C_v}, \quad \hat{n} \in \mathbb{R}^{B\times C_n}, \quad \mathbf{f}_a \in \mathbb{R}^{B\times N\times D} = E_{\text{action}}(\mathbf{F})$$

where C_v and C_n denote the number of verb and noun classes, respectively. The feature $\mathbf{f}_a$ encodes action-aware semantic information for downstream use.

Segmentation Module. We employ the Segment Anything Model (SAM) [8] to extract object masks guided by predicted 2D gaze coordinates. Given the current frame I_t and predicted gaze point $g_t = (x_t, y_t)$, SAM generates a binary mask $M_t \in \{0,1\}^{H\times W}$. Stacking across time produces $\mathbf{M} \in \mathbb{R}^{B\times T\times H\times W}$, which is encoded via the same backbone:

$$\mathbf{f}_m = \phi(\mathbf{M}) \in \mathbb{R}^{B\times N\times D}$$

Gaze Forecasting Decoder. The decoder predicts spatiotemporal gaze heatmaps by fusing the action-aware features $\mathbf{f}_a$ and the mask features $\mathbf{f}_m$:

$$\mathbf{z}_0 = \text{Fuse}(\mathbf{f}_a,\, \mathbf{f}_m) \in \mathbb{R}^{B\times N\times D}$$

The fused embedding $\mathbf{z}_0$ is processed by transformer decoding layers to produce gaze heatmaps:

$$\hat{G} = \text{Decoder}(\mathbf{z}_0) \in \mathbb{R}^{B\times 1\times T\times H'\times W'}$$

where H', W' denote the spatial resolution of the predicted heatmaps. The output $\hat{G}$ represents the predicted spatiotemporal gaze distribution across the future T frames.

3.3. Training

Two-stage Training Strategy. We adopt a two-stage training scheme to ensure that the gaze forecasting module benefits explicitly from the action semantics. In the first stage, we train the egocentric video encoder and the action feature extractor jointly using only action supervision. This allows the model to learn discriminative activity-aware visual representations that capture the user's semantic context.

In the second stage, we introduce the gaze forecasting decoder and train the entire model end-to-end using both action and gaze supervision. Rather than freezing the action encoder, we allow its parameters to be fine-tuned alongside the decoder, enabling tighter semantic alignment between action understanding and gaze prediction.

Loss Function. We use a combination of classification and distribution-based losses for the two tasks. For action prediction, we apply a multi-part cross-entropy loss that accounts for verb and noun classification:

$$\mathcal{L}_{\text{action}} = \lambda_v \cdot \text{CE}(\hat{a}_{\text{verb}}, a_{\text{verb}}) + (1 - \lambda_v) \cdot \text{CE}(\hat{a}_{\text{noun}}, a_{\text{noun}}) \tag{1}$$

where λ_v is a hyperparameter that controls the relative importance of verb versus noun prediction.

For gaze prediction, we use the Kullback–Leibler divergence to measure the discrepancy between the predicted and ground-truth gaze heatmaps over the target interval:

$$\mathcal{L}_{\text{gaze}} = \text{KL}(\hat{G} \parallel G) \tag{2}$$

The total loss is computed as the weighted sum of the two task-specific losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gaze}} + \mathcal{L}_{\text{action}} \tag{3}$$

This formulation allows the model to jointly optimize gaze forecasting and action understanding in an end-to-end manner, reinforcing semantic alignment between behavioral cues and spatial attention.

4. Experiments

4.1. Experimental Setup

Dataset. The EGTEA Gaze+ [13] includes HD videos, frame-level action annotations, and gaze tracking data. It was collected from 32 subjects in 86 sessions and was captured over 28 hours of cooking activities. Videos and gaze points have the same resolution of 1280×960 and 24 fps fixed frame rate. The action annotations of 106 unique classes are based on 19 verbs and 53 nouns. Original action-segmented clips are trimmed in accordance with our task settings, as we only use segments between 2 to 4 seconds. Clips shorter than 2 seconds are discarded, and those longer than 4 seconds are partitioned into non-overlapping 4-second video segments. For any residual segment at the end of a clip, the same criterion is applied: segments shorter than 2 seconds are discarded while those of 2 seconds or longer are retained. After preprocessing, there are 8,589 training clips and 2,019 test clips. Other popular datasets such as Ego4D [5] and AEA [16], which were used in CSTS [10], are not suitable for our model due to the absence of action labels.

Evaluation metrics. Following previous studies on egocentric gaze anticipation [10, 13], we adpot F1 score, recall, and precision as evaluation metrics.

4.2. Implementation Details

For each training sample, we use video frames from $t - i$ to t as input, where $i \in [1, 3]$, and predict gaze heatmaps from t to $t + 1$. During training, we sample 8 frames from the input video, and resize them to 256×256. We adopt the MViT backbone to the encoder and the decoder with the same structure as CSTS [10]. For SAM, we denormalize input frame back to its original RGB before feeding it to SAM. For each frame, SAM outputs three candidate masks along with confidence scores; we select the single mask with the highest score. We set a threshold of 0.5 to binarize the selected mask. We use AdamW [15] optimizer with initial learning rate of 1e-4 and cosine annealing scheduler. The momentum and weight decay are 0.9 and 0.05 respectively. The model is trained for 15 epochs with a batch size of 8 per GPU using 4 NVIDIA RTX 3090 GPUs. Other hyperparameters are listed in Tab. 1

Module	**Parameter**	**Value**
Input	Frame size	256
	Num. frames	8
Patch embedding	Embedding dim	96
	Patch kernel	(3, 7, 7)
	Patch stride	(2, 4, 4)
	Patch padding	(1, 3, 3)
Action Head	Verb / Noun classes	19 / 53
	Verb / Noun weights (stage1)	1.0 / 1.0
	Verb / Noun weights (stage2)	0.5 / 0.5
Training	Loss functions	KLDiv + CE
	Optimizer	AdamW
	Learning rate	1×10^{-4}
	Momentum	0.9
	Weight decay	0.05
	Max epochs	15

Table 1. Key hyperparameter settings used in each module of the proposed model.

4.3. Main Results

Baselines. We compared our model's performance with three other models: GLC [9], MViT [2], and EgoVideo [17]. Although the current state-of-the-art (SOTA) model is CSTS [10], direct comparison on the dataset was not feasible, as EGTEA Gaze+ does not include audio data. Since CSTS relies on an audio–video fusion mechanism, its fusion layers become ineffective without audio, rendering the CSTS functionally identical to MViT. Therefore, we decided to compare with MViT backbone rather than CSTS. We also compare our method with EgoVideo [17], a newly released egocentric foundation model as an encoder. Since EgoVideo was specifically designed to understand the unique characteristics of egocentric video, we expected this model would exhibit comparable performance to MViT, which is trained on general video data. Finally, we adapt the SOTA egocentric gaze estimation model, GLC [9] to the anticipation setting. In [10], GLC also demonstrated performance slightly lower than CSTS but higher than MViT in the anticipation task.

Methods	F1 Score	Recall	Precision
GLC [9]	49.1	52.8	46.0
MViT [2]	49.1	54.2	44.8
EgoVideo [17]	45.4	53.8	39.6
Ours	**54.4**	**65.7**	**46.4**

Table 2. Performance comparison on EGTEA Gaze+. Existing gaze estimation model is also employed as anticipation setting for more thorough comparison. The best results are highlighted in bold.

Methods	F1 Score	Recall	Precision
vanilla MViT	49.1	54.2	44.8
Action only	46.4	61.2	37.4
SAM only	**54.8**	64.5	**47.6**
Ours	54.4	**65.7**	46.4

Table 3. Results of ablation study.

As summarized in Tab. 2, our action-segmentation guided gaze anticipation model outperforms other models across all evaluation metrics. The result shows that our method to combine the action and segmentation approach is effective for gaze anticipation. Meanwhile, contrary to expectations, EgoVideo showed lower performance than MViT. While this could be an issue with the performance of encoder itself, it could be related to the difference in model architecture, as EgoVideo model uses a convolutional decoder instead of an MViT decoder due to the absence of intermediate features.

4.4. Ablation Study

Tab. 3 shows the result of ablation study on EGTEA Gaze+ dataset. To study the effect of each module in our model, we compare our full model with ablated versions: vanilla MViT as our pure backbone model, "Action only" as the model using only the action features without mask features from segmentation module, and "SAM only" as the model using only the segmentation branch without action features. Unexpectedly, the results are contrary to our initial hypothesis. The "Action only" model performed lower than the MViT backbone, and even our full model exhibited slightly inferior performance compared to the "SAM only" model in terms of F1 Score and Precision. These results consistently suggest that the action module had a detrimental impact on performance enhancement. It appears that the majority of the performance gain was driven by the SAM module.

Our initial hypothesis that action features would be beneficial for gaze anticipation appears to be incorrect. We assume that it is not because action information itself is unhelpful, but rather because the current module, utilizing an MLP layer, failed to effectively integrate the action features into the model. The action context features, which are added with the original frame embedding in the action module, were designed for verb/noun classification. Consequently, this might lead the decoder to be influenced by confusing information that is irrelevant to the original gaze anticipation task. Also, the same context embedding is broadcasted and added to all N tokens of each sample within a batch. Given that verb/noun information is global while visual tokens are local, injecting the same global information into all tokens might constitute an excessive overgeneralization, potentially weakening the spatio-temporal characteristics of the original frame embeddings.

4.5. Visualization

Fig. 3 and illustrates the visualization of the anticipation gaze heatmaps of our model and the baselines. Our method shows clear advantages in actions involving significant object movement within the frame, such as pouring liquid or cutting ingredients. In such cases, gaze transitions are strongly influenced by the trajectory of moving objects, and our model is able to better anticipate these shifts by effectively integrating action semantics and temporal dynamics. In contrast, bottom-up approaches like MViT and GLC perform reasonably well on simpler actions with minimal object motion but struggle when the task requires tracking rapidly moving targets. These methods tend to rely heavily on static visual saliency and lack the capacity to model high-level task-driven gaze behavior. We presume that the regions from gaze-guided segmentation masks can provide more localized information about the objects currently being observed by the viewer to the model.

We also illustrate a failure case in Fig. 4. When the gaze shifts abruptly, the model cannot anticipate the correct gaze position. This is a common limitation of all previous gaze anticipation models. This problem is hard to resolve due to an inherent limitation of the gaze anticipation task itself, which necessitates forecasting gaze without access to future frames.

5. Conclusion

In this paper, we addressed the challenging problem of egocentric gaze anticipation, which is crucial for understanding human intent and interaction in first-person videos. We propose a novel action-segmentation based approach to forecast future gaze targets. Our key contribution is that our model estimate future gaze points more accurately by leveraging gaze point information from the observation time. Furthermore, we attempted to introduce action-related information into the gaze anticipation task, albeit with results that differed from our initial hypothesis.

There are some limitations in our study. First, the model is heavier and slower than baselines due to the use of SAM.

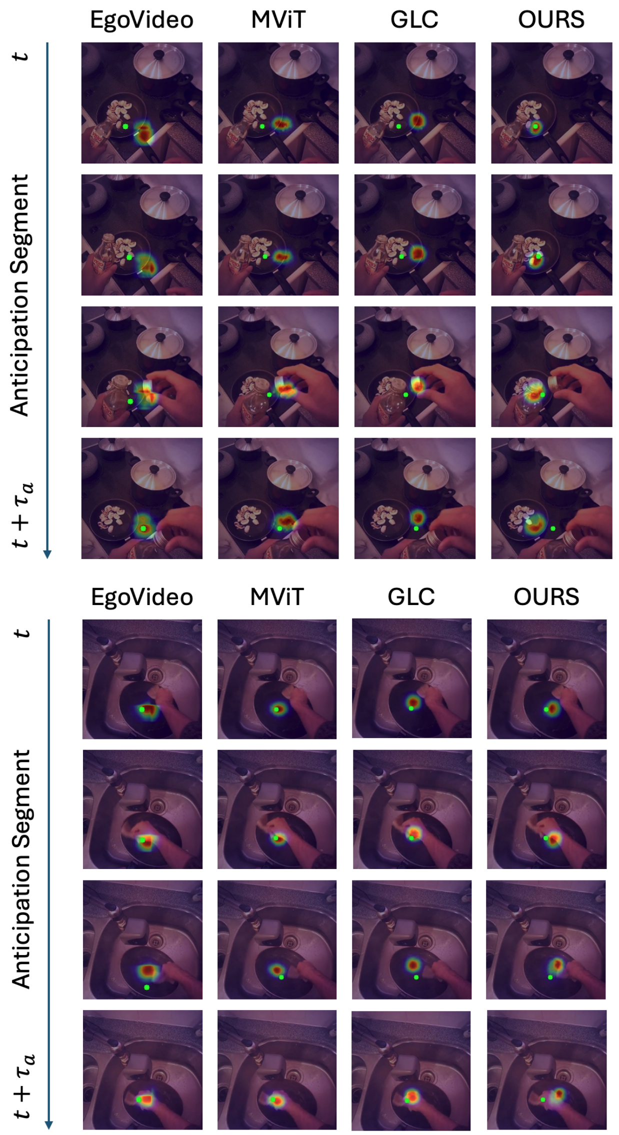

Figure 3. Egocentric gaze anticipation results from our model and other baselines. Four future time steps are uniformly sampled from the anticipation segments. Green dots indicate the ground truth gaze location. Above (”Pour oil”) is the case where our model predicts more accurately than other baseline models, while below (”Wash dish”) is the case where the result of our model and baselines is similar.

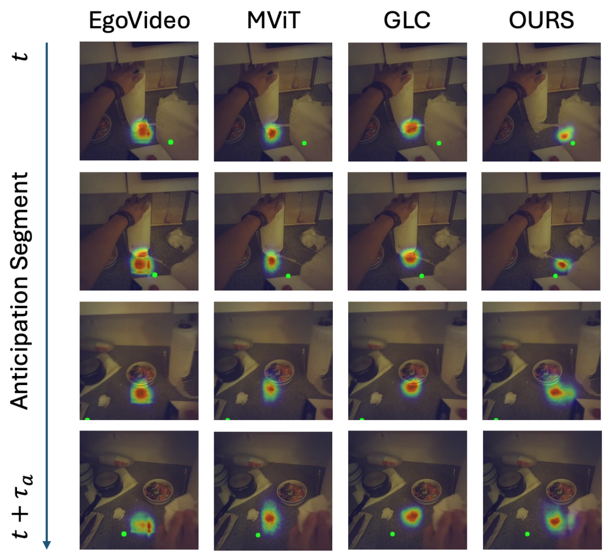

Figure 4. Failure cases of our model and baselines. Green dots indicate the ground truth gaze location. (Action annotation: ”Divide/Pull Apart paper towel”)

The overhead appears to largely stem from internal structural issues within SAM. Specifically, the mask derivation operation within SAM is not processed in a batched manner; instead, it is handled individually for each frame. Furthermore, SAM requires input as NumPy arrays rather than tensors, which causes frequent data transfers between the CPU and GPU. If we use an improved segmentation module addressing these architectural inefficiencies, we could significantly reduce the overhead.

Second, the dataset is limited to cooking situation. This is primarily due to the scarcity of datasets that possess both gaze and action annotations simultaneously. Attaching fine-grained action annotations per frame is a highly labor-intensive and costly process. Therefore, a promising direction for future study could try a self-supervised learning, or use a pretrained video caption model so as to automatically generate action labels of each video clip.

Third, our model use ground-truth gaze points at observation time instead of predicted gaze points. Although using information prior to the anticipation time is valid, relying on ground-truth gaze data limits inference to only cases where such data are available, thereby reducing its practical applicability. Since this study confirmed the benefit of gaze point information from the observation time, future work could explore leveraging gaze points inferenced by other gaze estimation models instead of ground-truth, enhancing the method’s generalizability and practical use.

Lastly, our model fails to completely overcome the limitation of previous approaches in predicting drastic gaze movements. As mentioned earlier, this is due to the inherent nature of the gaze anticipation task itself, which necessitates forecasting gaze without access to future frames. To solve this problem, we need to figure out the relationship between future frame movement and gaze movement. Future studies may account for these two components simultaneously.

In spite of these limitations, our work provides important insights into the relationship of gaze, action, and observed objects. Since several unresolved challenges are still left in egocentric gaze anticipation, we hope our research encourages further studies into this domain.

References

[1] Ali Borji, Dicky N Sihite, and Laurent Itti. Probabilistic learning of task-specific visual attention. In *2012 IEEE Conference on computer vision and pattern recognition*, pages 470–477. IEEE, 2012.

[2] 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.

[3] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 314–327. Springer, 2012.

[4] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[5] 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.

[6] 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.

[7] 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.

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[9] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, 132(3):854–871, 2024.

[10] Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, pages 192–210. Springer, 2024.

[11] Xiang Li, Heqian Qiu, Lanxiao Wang, Hanwen Zhang, Chenghao Qi, Linfeng Han, Huiyu Xiong, and Hongliang Li. Challenges and trends in egocentric vision: A survey. *arXiv preprint arXiv:2503.15275*, 2025.

[12] 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.

[13] 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.

[14] Yujie Li, Xinghe Wang, Zihang Ma, Yifu Wang, and Michael C Meyer. Swingaze: Egocentric gaze estimation with video swin transformer. In *2023 IEEE 16th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, pages 123–127. IEEE, 2023.

[15] Ilya Loshchilov and Frank Hutte. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[16] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024.

[17] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, and Yu Qiao. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024.

[18] 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.

[19] 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.

[20] Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. Can saliency map models predict human egocentric visual attention? In *Asian conference on computer vision*, pages 420–429. Springer, 2010.

[21] Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. Attention prediction in egocentric video using motion and visual saliency. In *Advances in Image and Video Technology: 5th Pacific Rim Symposium, PSIVT 2011, Gwangju, South Korea, November 20-23, 2011, Proceedings, Part I 5*, pages 277–288. Springer, 2012.

[22] Heeseung Yun, Ruohan Gao, Ishwarya Ananthabhotla, Anurag Kumar, Jacob Donley, Chao Li, Gunhee Kim, Vamsi Krishna Ithapu, and Calvin Murdock. Spherical world-locking for audio-visual localization in egocentric videos. In *European Conference on Computer Vision*, pages 256–274. Springer, 2024.

[23] 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.

[24] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Anticipating where people will look using adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1783–1796, 2018.