Supplementary Material for Anticipating Where People Will Look Using Adversarial Networks

Mengmi Zhang, *Student Member, IEEE,* Keng Teck Ma, *Member, IEEE,* Joo Hwee Lim, *Senior Member, IEEE,* Qi Zhao, *Member, IEEE,* and Jiashi Feng, *Member, IEEE,*

1 Gaze anticipation on GTEAplus

We provide evaluation results of gaze anticipation in GTEAplus dataset in Figure S1. Our DFG model outperforms all the competitive baselines in this dataset which is consistent with the other three datasets introduced in the main text.

2 Qualitative results on gaze anticipation

We show more qualitative examples on gaze anticipation on both egocentric and third person videos in Figure S2, S3, S4, S5 and S6. It covers various activities, such as fridge opening, cooking, kissing, and person fighting. Our DFG model demonstrates its gaze anticipation capability in diverse activities.

3 Spatial Bias Analysis - Gaze Distribution Map

Compared with third-person videos, researchers have found that egocentric gaze has smaller variance in space and the gaze preference is often toward the bottom part of the image in object manipulation tasks [1]. We compute a 2D gaze distribution map by collecting all the human fixations from the training set for each dataset and report the two variations of utilizing the gaze distribution map: (1). the 2D gaze distribution map alone as the predicted temporal saliency map on all future frames; (2) we replace **DFG-P** in our DFG model with the gaze distribution map.

4 Head Motion Effect on Gaze anticipation

In the main text,we analyze how head motion influences gaze anticipation performance. Here we show two examples where **Generator** fail to synthesize realistic future frames due to large head motion. We quantify the large head motion as the averaged magnitude of head motion vector to be larger than 6 pixels calculated based on optical flow on boundary pixels over the next 31 future frames. In Figure S7a, the anticipated gaze location still matches the ground truth despite the large head motion but in Figure S7b, it fails. This again validates the point that egocentric videos have characteristics of having small gaze shifts in space as they often get compensated by the head motion. However, this phenomenon does not imply that either center bias or the gaze distribution map from all human fixations in the training set is sufficient for gaze anticipation. See Section 4.6 in the main text for more discussions.

5 How humans perform in gaze anticipation

In the main text, we discussed about how humans perform in the gaze anticipation task. In this section, we provide detailed description of the psychophysics experiment.

Ethical Statement

All the experiments were conducted with the subjects' informed consent and according to the protocols approved by the Institutional Review Board.

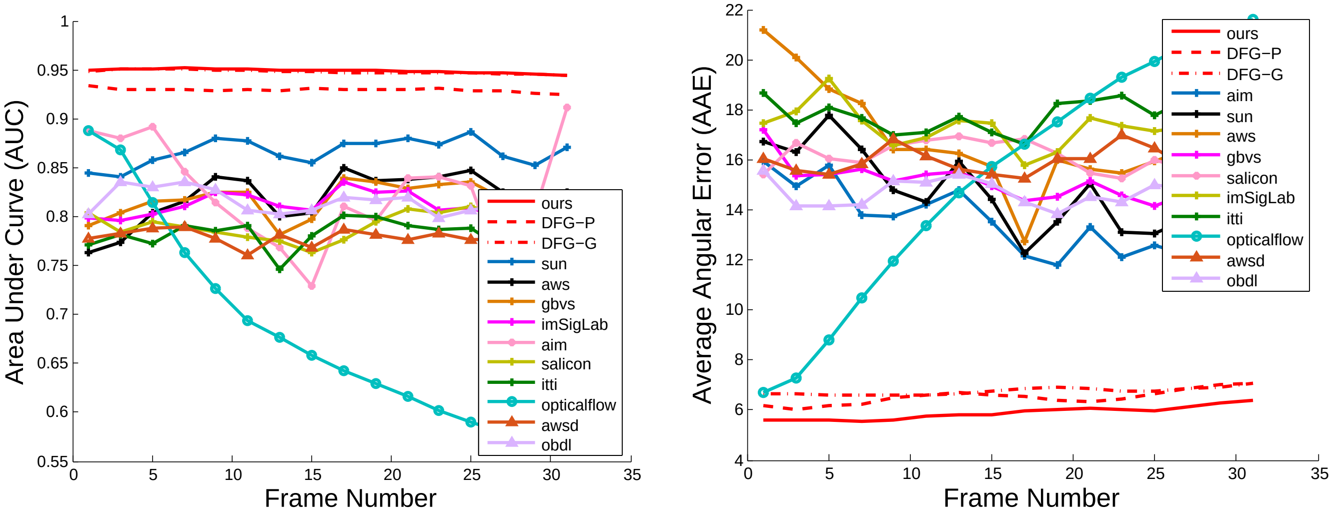
Experimental procedures

The experiment started with a briefing on the study's objectives and procedures. During briefing, all participants are instructed on the objectives of the study: comparison between algorithms and human performance on the gaze anticipation task. The gaze anticipation task is prediction of gaze point on future unseen frames from a single video frame. Participants were given unlimited time to complete the task. There are 2 sessions with each session containing 50 testing video clips either from GTEA or GTEAplus datasets.

In each session, there are 2 training phases and 1 testing phase. Training phase 1 is to familiarize the participants with the system. Training phase 2 is similar to the supervised learning of our model. Testing phase is the same setup as our machine experiments, that is given one frame, anticipate the gaze positions for some future frames.

In training phase 1, the participant was shown a video frame. It is the ego-centric view of the scene with the recorded gaze (red circle) overlay on it. This is repeated for all frames of each video clip. There are 5 video clips during this training phase.

In training phase 2, participant was shown a video frame followed by a blank gray screen. The participant was then instructed to imagine the next frame and click



(a) Evaluation using Area Under the Curve (AUC)

(b) Evaluation using Average Angular Error (AAE)

Fig. S1. Evaluation of Gaze Anticipation using Area Under the Curve (AUC) on the current frame as well as 31 future frames (2.7 sec ahead) in GTEAplus Dataset. Larger is better for AUC. Smaller is better for AAE. The algorithms in the legend are introduced in Section 4.3 in the main text.

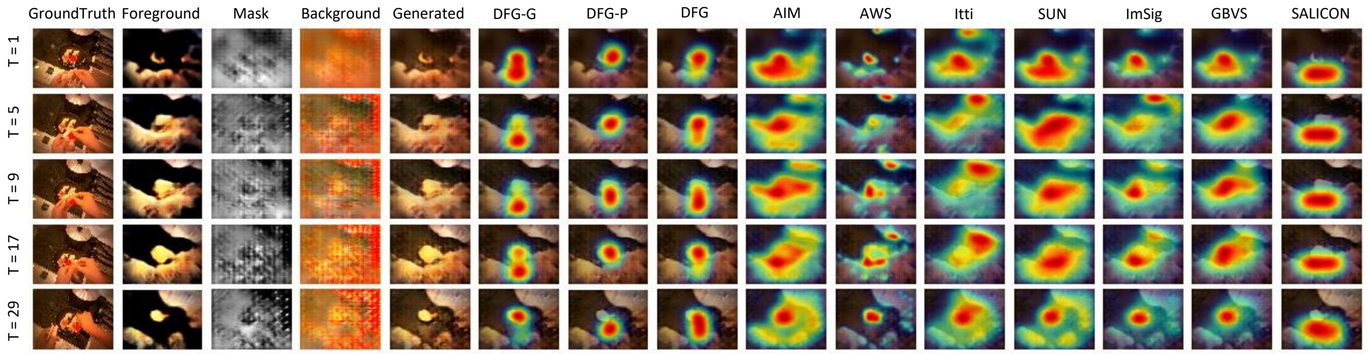


Fig. S2. Example results of gaze anticipation on egocentric video datasets. Our DFG model produces 31 future frames based on the current frame. From first to last rows, results on future frames #1, 5, 9, 17, 29 with respect to the current frame are shown. The leftmost column shows the ground truth (GT) with red circle denoting human gaze locations. Column 2, 3, 4 (FG, mask, BG) show the foreground $F(\cdot)$, the mask $M(\cdot)$, and the background $B(\cdot)$ learnt by **Generator** respectively. Column 5 shows the generated future frames (GEN). Column 6 and 7 show the corresponding predicted temporal saliency maps from two pathways **DFG-G** and **DFG-P** in our model. Column 8 show the final integrated temporal saliency maps predicted by our model. Column 9 and onwards show the predicted temporal saliency maps by all baselines (See Section 4.3 in the main text). Best viewed in color.

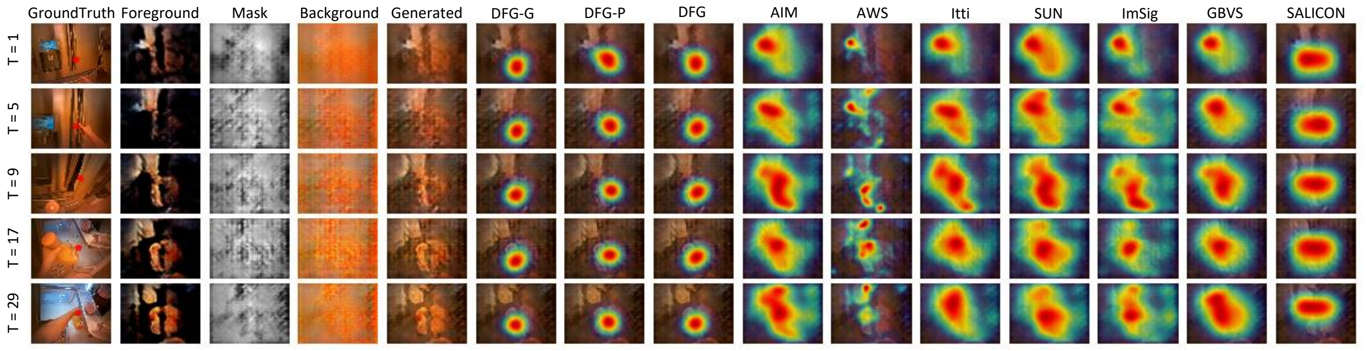


Fig. S3. Example results of gaze anticipation on egocentric video datasets. The format and conventions follow those in Figure S2.

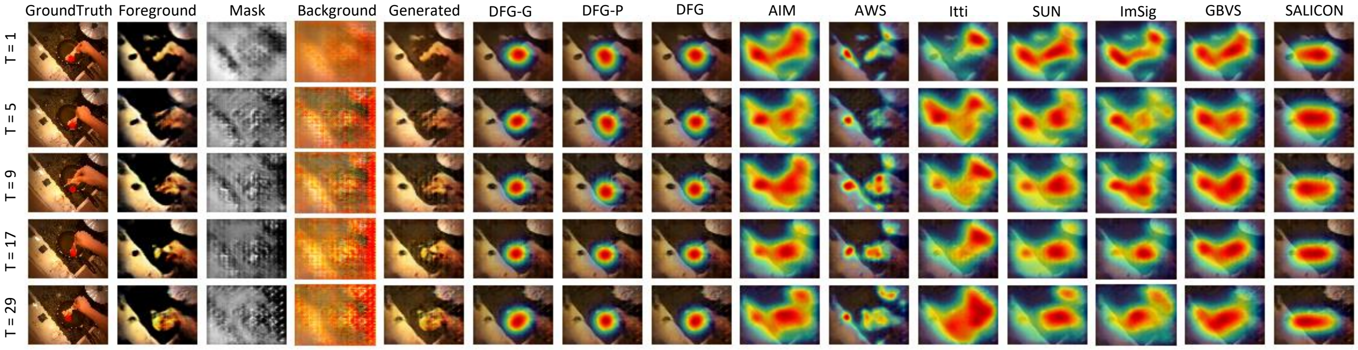


Fig. S4. Example results of gaze anticipation on egocentric video datasets. The format and conventions follow those in Figure S2.

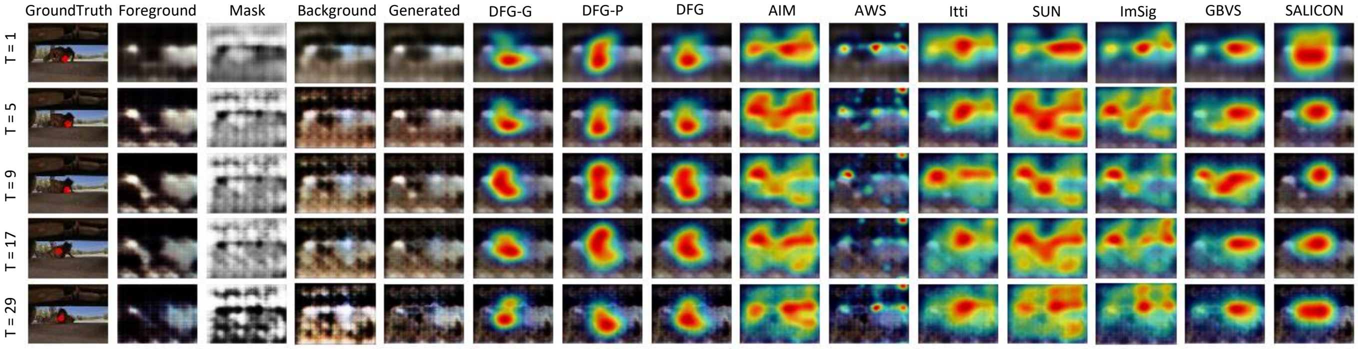


Fig. S5. Example results of gaze anticipation on Hollywood2 third person video dataset. The format and conventions follow those in Figure S2.

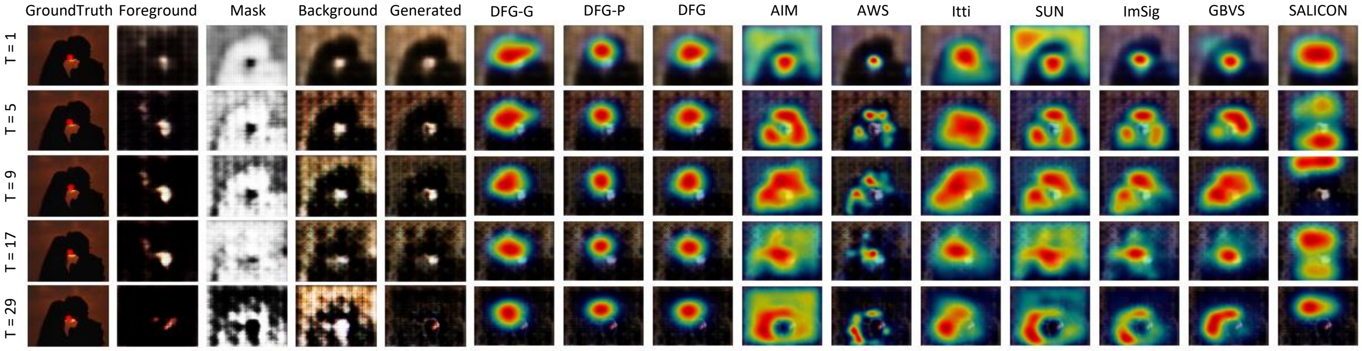


Fig. S6. Example results of gaze anticipation on Hollywood2 third person video dataset. The format and conventions follow those in Figure S2.

on their anticipated gaze location for the next frame. The participant was shown the recorded gaze (red circle) overlay on the next frame (i.e. ground truth). The user's mouse click position (blue cross) was also overlaid as the feedback to the participant. This was repeated for all frames of all video clips. There are also 5 video clips for this training phase.

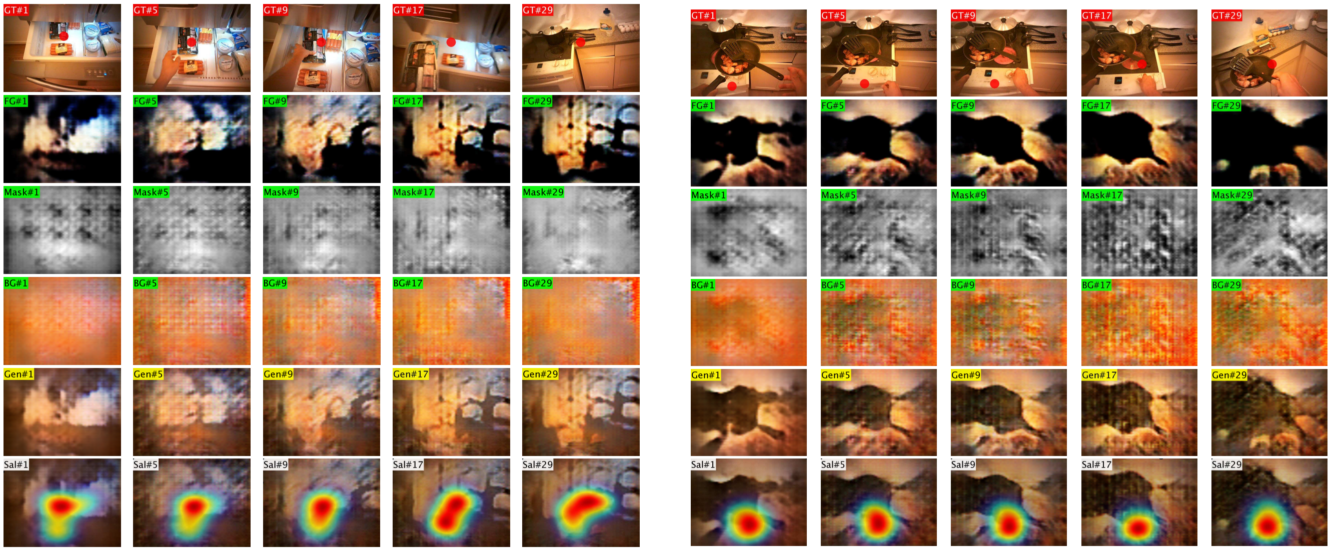
In testing phase, participant was shown a video frame. The participant was then instructed to click on their location of the predicted gaze for this frame. A blank gray screen was shown to the participant. The participant was then instructed to imagine the next frame and click on the anticipated gaze location for the subsequent frame. The blank gray screen was repeated for 32 frames of the video clip. For each blank screen, participant imagined the future frame and clicked on the anticipated gaze location.

6 Schematics of Ablated Models

In order to study the effect of the individual component of DFG on both egocentric and third person videos, we conduct an ablation study and test on GTEA, OST and Hollywood2 datasets by removing *only* one component in DFG at one time while the rest of the architecture remains the same. In the main text, we introduce five ablated models. Figure S8, S9, S10, S11 and S12 show the schematics of these five ablated models.

7 Analysis on Frame numbers

In video analysis, the number of consecutive frames is a key parameter in practice. To study the effect of the number of frames on which we anticipate gaze, we assign the scalar



(a) Example 1 (Head motion = 6.4)

(b) Example 2 (Head motion = 27.8)

Fig. S7. Example results of gaze anticipation when there is large head motion. In each example, frames #1, 5, 9, 17, 29 are shown (left to right columns). The topmost row shows the ground truth with red circle denoting human gaze locations. Row 2, 3, 4 show the foreground $F(\cdot)$, the mask $M(\cdot)$, and the background $B(\cdot)$ learnt by **Generator Network** respectively. Row 5 shows the generated future frames. Row 6 shows the corresponding predicted temporal saliency maps. Best viewed in color.

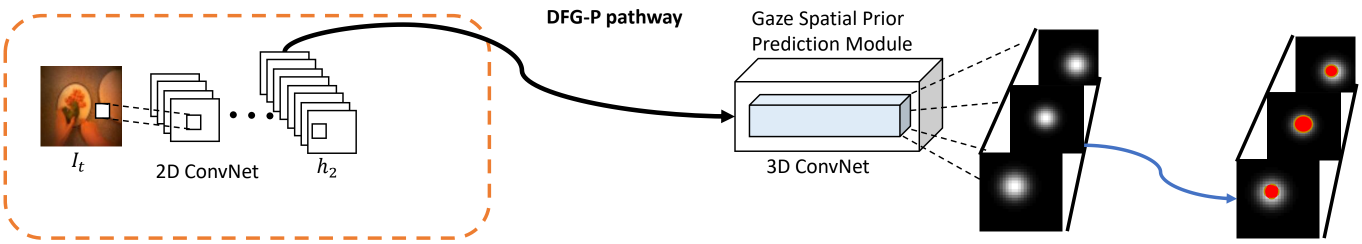


Fig. S8. Ablated model 1: we remove **DFG-G** and evaluate the predicted temporal saliency maps from **DFG-P** only.

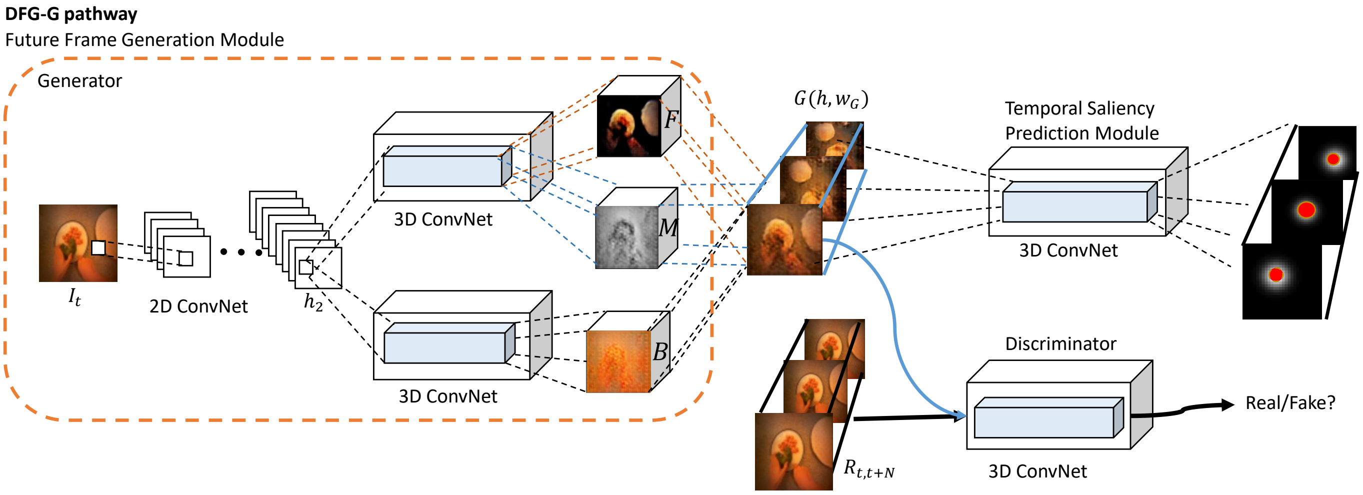


Fig. S9. Ablated model 2: we remove **DFG-P** and this is the same as our previous algorithm with only **DFG-G** [2].

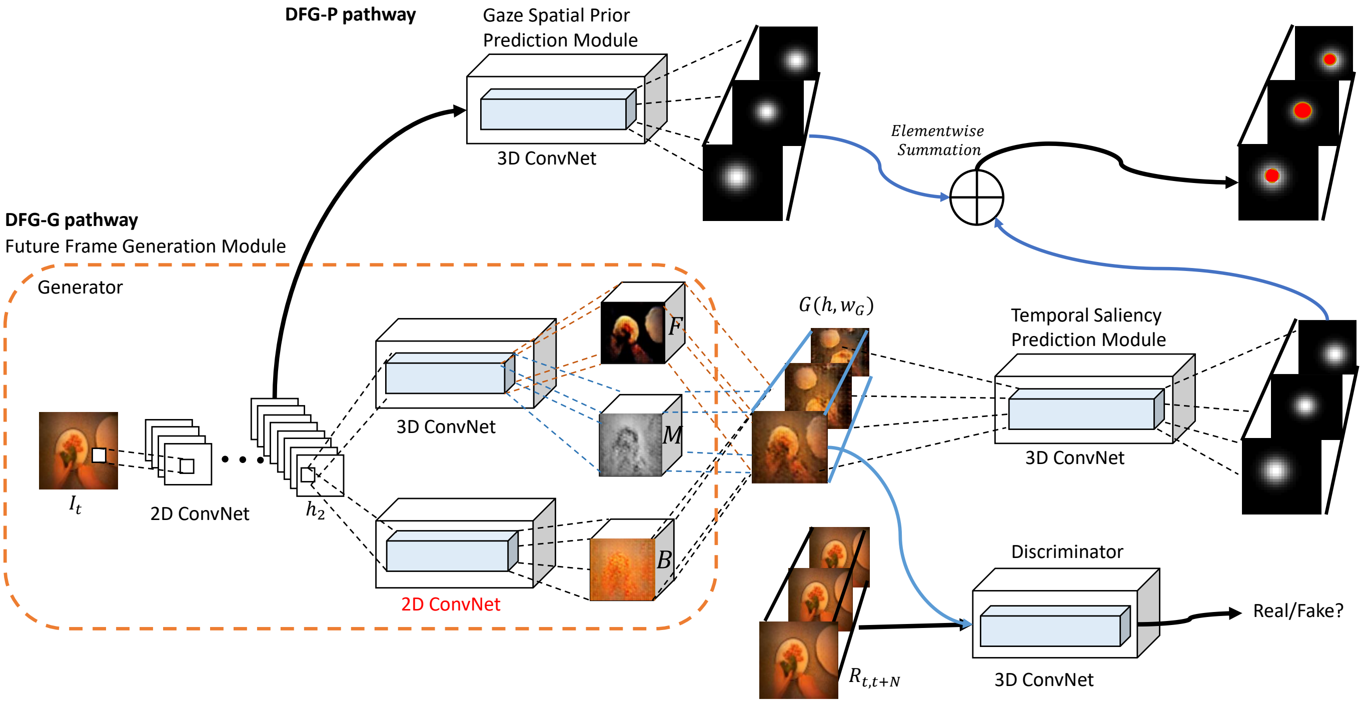


Fig. S10. Ablated model 3: we replace the two-stream 3D-CNN in **Generator** with the same structure as [3], *i.e.* the background stream is 2D-CNN which assumes the background is "static" while the foreground stream remains the same. Differences from our DFG model are highlighted in red.

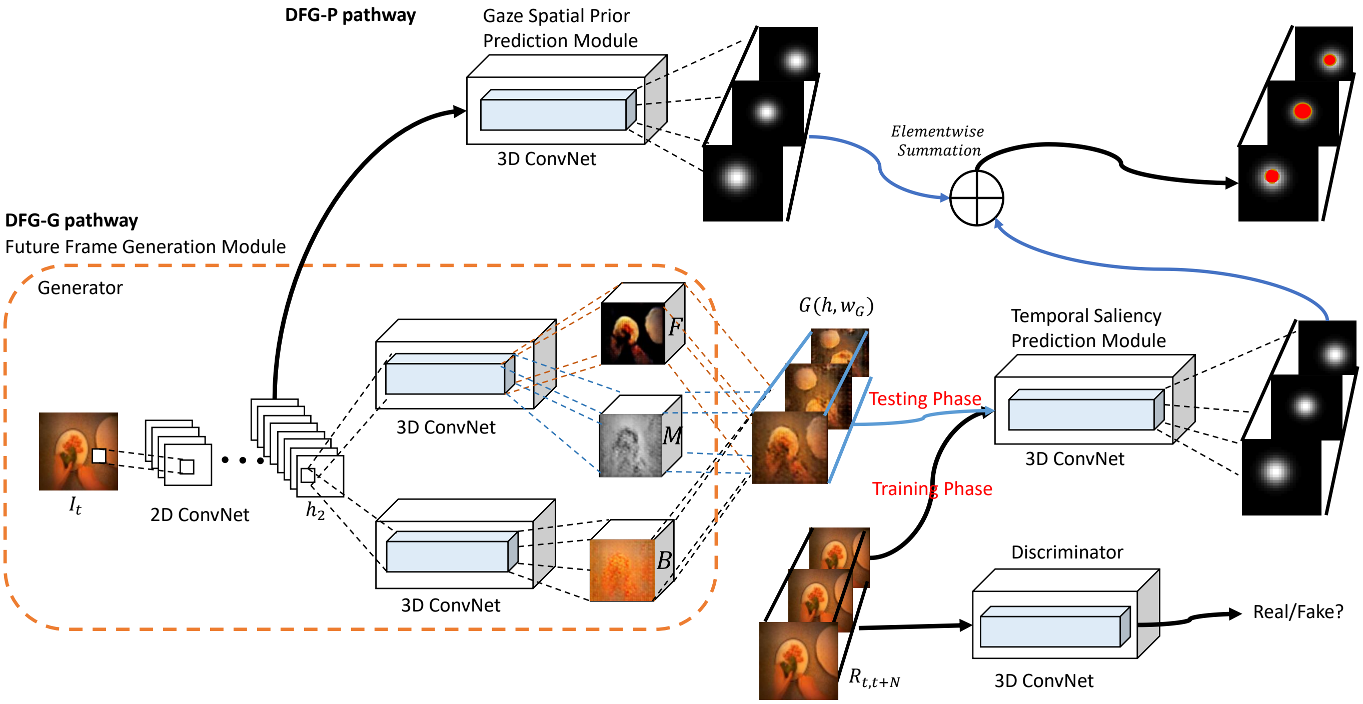


Fig. S11. Ablated model 4: we train **Temporal Saliency Prediction** directly on real frames and test it on the generated frames from **Generator**. Differences from our DFG model are highlighted in red.

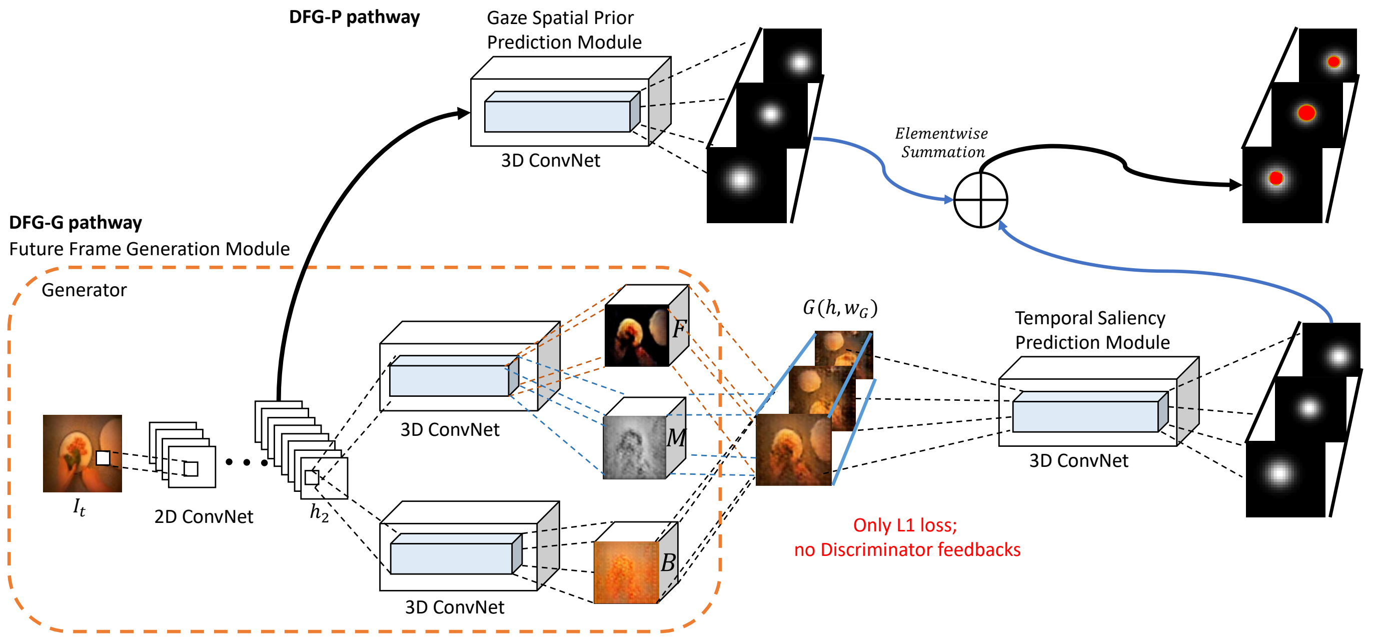


Fig. S12. Ablated model 5: we remove **Discriminator** and we only use L1 distance loss for future frame generation. Differences from our DFG model are highlighted in red.

TABLE S1
Correlation Between Number of Frames and Corresponding Performance of Our Model

Angular Average Error (AAE)					
	# 1−2	# 3−4	# 5−8	# 9−16	# 17−32
#2	10.4	−	−	−	−
#4	10.7	10.9	−	−	−
#8	10.4	10.4	10.3	−	−
#16	10.2	10.0	10.3	10.8	−
#32	8.0	8.0	8.0	8.2	8.5
Area Under the Curve (AUC)					
	# 1−2	# 3−4	# 5−8	# 9−16	# 17−32
#2	0.87	−	−	−	−
#4	0.86	0.86	−	−	−
#8	0.87	0.87	0.86	−	−
#16	0.88	0.88	0.87	0.86	−
#32	0.91	0.91	0.91	0.90	0.89

weights to tune the losses in both **Generator** and **Temporal Saliency Prediction** for the next 32 frames while maintaining the same architecture. For example, we design the weight matrix to be $[1, 1, 1, 1, 0, ..., 0]$ for gaze anticipation in the next 4 frames while ignoring the subsequent frames. In Table S1, we present the averaged metric scores of our model for gaze anticipation in the next 2, 4, 8, 16, 32 frames starting from the current frame #1. Scores for gaze anticipation in both AAE and AUC are computed every # frames indicated in columns in the testset in GTEA Dataset.

8 Visualization of Convolution Filters

[4] proposed a top 4 patch visualization approach in 2D-CNN. We extend their work to visualization of 3D-CNN. As a simplified version of their method, we parse all video frames from the test set in GTEA and record the regions with the highest filter activation in both spatial and temporal dimensions for the first and the second last convolution layer in **Temporal Saliency Prediction** in our model. Those regions are then projected back into their input video frames based on their corresponding receptive fields across both space and time dimensions where the input frames are the current frame and its subsequent 31 frames. Due to the consistency of egocentric videos between adjacent frames, we increase the diversity of the visualization by sorting the filter activation from highest to lowest and selecting these top filters where their receptive fields do not overlap with their neighboring frames by a pre-defined threshold.

9 Gaze-aided Egocentric Activity Recognition

To verify our proposed future gaze model is also useful for egocentric activity recognition, we integrate gaze information into the feedforward 3D-CNN for egocentric activity recognition. As [5] shows that 3D-CNN can be used for activity recognition, we adapt the down-scaled framework from [5] (C3D) and integrate the anticipated gaze into the network. A Gaussian mask at the gaze location for each frame, as an additional channel, is concatenated with the input frames of RGB color channels. Cross entropy loss is used for training. Since GTEAplus dataset contains rich instances per activity class as recommended by [6], we follow their evaluation settings and select the top 44 activity classes which have the most instances per class in our recognition task. Confusion matrix of the model with our anticipated gaze is shown in Figure S13. In comparison, we also use the same architecture, discard the gaze information and train the network from scratch. In addition, we provide the baseline that the same architecture with the ground truth gaze information as the upper bound. Since center bias is also

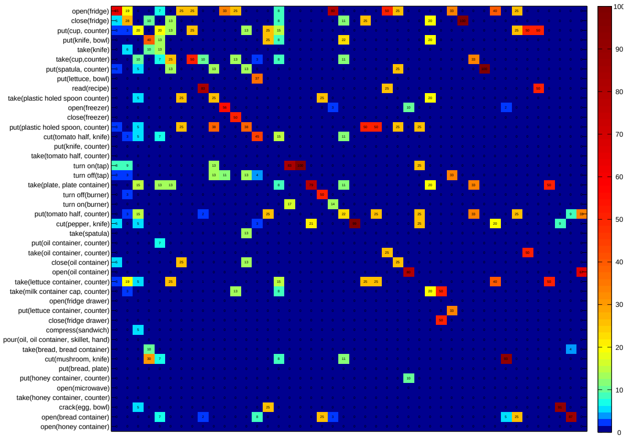


Fig. S13. Confusion matrix of 44 egocentric activity classes from GTEAplus Dataset. The 44 activity classes are selected similar as [6], [7]. The results are based upon C3D convolution architecture proposed by [5] for egocentric activity recognition with the fusion of our predicted gaze locations via one convolution layer.

TABLE S2
Accuracy of Gaze-aided Egocentric Activity Recognition

Models	Activity Recognition Rate
Guess At Random	2.3%
STIP	14.9%
Cuboids	22.7%
C3D	26.9%
C3D + center gaze	13.6%
C3D + DFG-G gaze	28.5%
C3D + our pred gaze	29.3%
C3D + ground truth gaze	33.5%

effective in gaze prediction, we create an artificial baseline where the network with the center gaze is also evaluated. Activity recognition rates are reported in Table S2.

References

[1] Y. Li, A. Fathi, and J. M. Rehg, “Learning to predict gaze in egocentric video,” in *ICCV*, 2013, pp. 3216–3223. 1

[2] Z. Mengmi, M. Keng Teck, L. Joo Hwee, Z. Qi, and F. Jiashi, “Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks,” in *CVPR, 2017*. IEEE, 2017. 4

[3] C. Vondrick, H. Pirsiavash, and A. Torralba, “Generating videos with scene dynamics,” *arXiv preprint arXiv:1609.02612*, 2016. 5

[4] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *ECCV*, 2014, pp. 818–833. 6

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” *arXiv preprint arXiv:1412.0767*, 2014. 6, 7

[6] Y. Li, Z. Ye, and J. M. Rehg, “Delving into egocentric actions,” in *CVPR*, 2015, pp. 287–295. 6, 7

[7] M. Ma, H. Fan, and K. M. Kitani, “Going deeper into first-person activity recognition,” *arXiv preprint arXiv:1605.03688*, 2016. 7