Listen to Look into the Future: Audio-Visual Egocentric Gaze Anticipation

Supplementary Materials

Bolin Lai[1], Fiona Ryan[1], Wenqi Jia[1], Miao Liu[2,*], and James M. Rehg[3,*]

[1] Georgia Institute of Technology
[2] GenAI, Meta
[3] University of Illinois Urbana-Champaign
{bolin.lai,fkryan,wenqi.jia}@gatech.edu miaoliu@meta.com
jrehg@illinois.edu

This is the supplementary material for the paper titled "Listen to Look into the Future: Audio-Visual Egocentric Gaze Anticipation". We organize the content as follows:

A Comparison with Prior Audio-Visual Learning Strategies

We have specified the key differences between the egocentric action anticipation task and saliency prediction task in the second paragraph of Sec. 2 in the main paper. The experiment results also validate that our proposed spatial-temporal separable fusion strategy performs better in our task than other fusion strategies designed for saliency prediction and action recognition (please refer to Tab. 2 in the main paper). In this section, we further compare our model with typical

* Equal corresponding author.

Table 1: Comparison with typical audio-visual learning methods for audio-visual saliency prediction and recognition. If more than one fusion strategies have been tried in one method, we only show the strategy leading to the best performance.

Methods	View	Fusion	Cntr	Architecture	Task
Tavakoli *et al.* [26]	Exo	Concatenation	w/o	CNN	Saliency Prediction
Min *et al.* [22]	Exo	Correlation Analysis	w/o	CNN	Saliency Prediction
Tsiami *et al.* [27]	Exo	Bilinear	w/o	CNN	Saliency Prediction
Yao *et al.* [32]	Exo	Inner Product	w/o	CNN	Saliency Prediction
Change *et al.* [1]	Exo	Bilinear	w/o	CNN	Saliency Prediction
Jain *et al.* [12]	Exo	Bilinear	w/o	CNN	Saliency Prediction
Wang *et al.* [28]	Exo	Concatenation	w/o	CNN	Saliency Prediction
Xiong *et al.* [31]	Exo	Self-Attentation	w/o	CNN	Saliency Prediction
Nagrani *et al.* [23]	Exo	Attention Bottleneck	w/o	Transformer	Video Classification
Huang *et al.* [10]	Exo	Self-Attention	w/	Transformer	Video Classification
Gao *et al.* [6]	Exo	Linear	w/o	LSTM	Action Recognition
Kazakos *et al.* [15]	Exo	Linear	w/o	CNN	Action Recognition
Wang *et al.* [29]	Exo	Weigted Sum	w/o	CNN	Action Recognition
Xiao *et al.* [30]	Exo	Self-Attentation	w/o	CNN	Action Recognition
Liu *et al.* [19]	Exo	Linear	w/o	CNN	Action Recognition
Senocak *et al.* [25]	Exo	Linear	w/o	CNN	Action Recognition
Praveen *et al.* [24]	Exo	Self-Attention	w/o	CNN	Emotion Recognition
Chudasama *et al.* [2]	Exo	Self-Attention	w/o	Transformer	Emotion Recognition
CSTS (Ours)	**Ego**	**Spatial-Temporal Separable**	**w/**	**Transformer**	**Gaze Anticipation**

audio-visual learning methods for saliency prediction and recognition tasks in terms of model design.

In Tab. 1, all prior methods are designed for exocentric videos (*i.e.*, third-person videos) that have a *fixed* camera viewpoint through all frames. Though various fusion approaches are used in these methods, they fuse audio-visual embeddings *jointly* in time and space. In contrast, the egocentric gaze anticipation task has the unique challenges of *moving* viewpoint together the *latency* between the audio stimuli and human reactions. To address these challenges, our model uses a novel spatial-temporal separable fusion strategy which has not been studied in prior work. The experiments in Tab. 2 of the main paper shows that our method achieves the best performance in egocentric gaze anticipation task compared with prior audio-visual learning strategies. In addition, using contrastive learning to boost audio-visual representations in a specific task is still an understudied area. Huang *et al.* [10] use inter- and intra- contrastive loss to learn aligned audio and visual embeddings. However, they straightforwardly apply contrastive loss on the *raw* embeddings right after the encoders. In our model, we innovatively propose to adopt contrastive loss on the embeddings after fusion layers (*i.e.*, post-fusion contrastive learning). We also validate its advantage in Tab. 3 of the main paper. These key differences consolidate our contributions and clearly distinguish our model from other audio-visual learning methods.

Table 2: Zero-shot experiments on Aria dataset. All baselines and our model are trained only on Ego4D training set. We consider F1 score as the *primary* metric in our experiments. The green row refers to our model, and the best results are highlighted with **boldface**. See Sec. C.1 for further discussion.

Methods	F1 Score	Recall	Precision
GazeMLE [17]	44.0	59.0	35.0
AttnTransit [11]	43.1	57.5	34.5
I3D-R50 [5]	41.5	77.2	28.4
MViT [4]	44.1	59.7	35.0
GLC [16]	46.9	72.8	34.6
DFG [34]	39.3	**80.4**	26.0
DFG+ [33]	43.1	76.4	30.0
CSTS	**50.8**	62.2	**42.9**

B More Dataset Details

The Ego4D [9] eye-tracking subset is collected in social settings (*i.e.*, social interaction benchmark) and totals 31 hours of egocentric videos from 80 participants. All videos have a fixed 30 fps frame rate and spatial resolution of 1088×1080, and audio streams are recorded with a sampling rate of 44.1kHz. We use the train/test split released in [16] in our experiments, *i.e.*, 15310 video segments for training and the other 5202 video segments for testing.

The Aria [21] dataset contains 143 egocentric videos (totaling 7.3 hours) collected with Project Aria glasses. It covers a variety of indoor everyday activities including cooking, exercising and spending time with friends. All videos have a fixed 20 fps frame rate and spatial resolution of 1408×1408. A sliding window is used to trim long videos into 5-second video segment with a stride of 2 seconds. We use 107 videos (10456 segments) for training and 36 videos (2901 segments) for testing. We will release our split to facilitate future studies in this direction.

Note that Ego4D and Aria are the two largest public datasets that provide all necessary data and labels (*i.e.*, egocentric videos, aligned audio streams and eye-tracking data) for egocentric audio-visual gaze anticipation.

C Additional Experiment Results

C.1 Experiments about Model Generalization Capability

To validate the generalization capability of our model, we compare our model with prior state-of-the-art models in a zero-shot setting. Specifically, We train our model and all baselines with Ego4D training set and test them with Aria test set. Note that the Aria data is invisible to all models during training. The results are presented in Tab. 2. Our model outperforms the best egocentric gaze anticipation model (DFG+) by +7.7% and also exceeds the strongest baseline (GLC) by +3.9% in F1 score (primary metric). The remarkable improvement

Table 3: Comparison with prior state-of-the-art models on egocentric gaze estimation. The green row refers to our model. The best results are highlighted with **boldface**. See Sec. C.2 for further discussion.

Methods	Ego4D			Aria		
	F1 Score	Recall	Precision	F1 Score	Recall	Precision
Center Prior	14.9	21.9	11.3	28.9	21.7	43.1
GazeMLE [17]	35.4	49.7	27.5	58.7	63.4	54.7
AttnTransit [11]	36.4	47.6	29.5	59.2	60.2	58.3
I3D-R50 [5]	37.5	52.5	29.2	60.9	69.5	54.2
MViT [4]	40.9	57.4	31.7	61.7	**71.2**	54.5
GLC [16]	43.1	57.0	34.7	63.2	67.4	59.5
CSTS	**43.7**	**58.0**	**35.1**	**64.5**	69.6	**60.1**

suggests that, with our novel fusion and contrastive learning approaches, our model is able to generalize better to other unseen data, which is critical for applying it to real-world problems.

C.2 Experiments on Egocentric Gaze Estimation

In addition to egocentric gaze anticipation, we also evaluate the advantage of our model in another gaze modeling problem – egocentric gaze estimation. Instead of forecasting *future* gaze, egocentric gaze estimation requires gaze target prediction in the *current* video frames. We use the same experiment setup from the recent state-of-the-art method [16].

As demonstrated in Tab. 3, the prior work [16] has shown the superiority of using a transformer-based architecture for egocentric gaze estimation. By incorporating the audio modality, CSTS surpasses the backbone MViT [4] (vision-only counterpart) by +2.8% on both Ego4D on Aria in terms of F1 score. These results indicate the audio modality also makes important contributions to the performance on egocentric gaze estimation. Furthermore, our model outperforms GLC [16] by +0.6% and +1.3% on Ego4D and Aria respectively, achieving a new state-of-the-art performance for this problem. However, our method has a smaller performance improvement on the gaze estimation task compared to gaze anticipation. The possible reason is that the audio stream has a stronger connection with future gaze targets than current gaze behaviors because of the natural latency between the audio stimuli and human reactions.

C.3 Additional Experiments on Contrastive Learning

In our model, we propose to use the audio-visual representations obtained after fusion (*i.e.* u_v and u_a) to calculate contrastive loss (*i.e.*, post-fusion contrastive learning). As a comparison, we also implement a baseline by feeding the raw embeddings from the encoders (*i.e.* $\phi(x)$ and $\psi(a)$) to the contrastive loss which is

Table 4: Study of different strategies for contrastive loss implementation. Post Cntr refers to our proposed post-fusion contrastive learning strategy, and the green row refers to the complete CSTS model. The best results are highlighted with **boldface**. See Sec. C.3 for further discussion.

Methods	Ego4D			Aria		
	F1 Score	Recall	Precision	F1 Score	Recall	Precision
STS + Vanilla Contr	39.0	53.7	30.6	59.1	66.5	53.1
STS + S-Contr	38.5	53.5	30.0	59.0	66.3	53.1
STS + T-Contr	38.9	54.0	30.5	59.0	66.7	53.0
STS + Cross Contr	38.9	**54.4**	30.2	59.3	66.8	53.3
STS + Post Contr	**39.7**	53.3	**31.6**	**59.9**	**66.8**	**54.3**

denoted as **Vanilla Contr**. To further investigate the contribution of contrastive learning, we also conduct experiments with three additional strategies:

Cross Contr. In our final model (CSTS), we use the new visual representation $u_v = u_{v,s} \otimes u_{v,t}$ and the new audio representation $u_a = \psi(a) \otimes u_{a,t}$ as input to the contrastive loss. In Cross Contr, we still use u_v yet replace u_a by reweighting the audio representation $u_{a,s}$ after the spatial fusion with weight $u_{a,t}$ from the temporal fusion, *i.e.* $u_a^* = u_{a,s} \otimes u_{a,t}$, as input to the contrastive loss. Please refer to Fig. 2 in the main paper for the meaning of each notation.

S-Contr. We use the output from the spatial fusion module ($u_{v,s}$,$u_{a,s}$) to calculate the contrastive loss.

T-Contr. We use the output from the temporal fusion module ($u_{v,t}$,$u_{a,t}$) to calculate the contrastive loss.

We implement all contrastive learning baselines above on our proposed model architecture and fusion strategy (*i.e.*, STS). The results are summarized in Tab. 4. Both S-Contr and T-Contr lag behind or perform on par with Vanilla Contr. One possible reason is that conducting contrastive learning using features obtained from only one fusion branch may compromise the representation learning of the other branch. Additionally, Cross Contr works on-par with Vanilla Contr on Ego4D but performs better on Aria. It also consistently outperforms S-Contr and T-Contr. This result validates our claim that implementing contrastive loss with reweighted representations from both spatial and temporal fusion leads to more gains for egocentric gaze anticipation. Moreover, our proposed strategy (reweighting the raw audio embedding $\psi(a)$ rather than the fused embedding after spatial fusion) outperforms Cross Contr. This is because in Cross Contr $u_{a,s}$ is derived from spatial fusion, where each audio token is fused with 64 visual tokens in the spatial fusion branch resulting in the dilution of audio features. All results further demonstrate the benefits of our proposed contrastive learning strategy.

C.4 Additional Visualization

We showcase more qualitative comparisons with all the baselines for egocentric gaze anticipation in Fig. 1. We observe CSTS makes the most accurate predictions. We also illustrate some typical failure cases in Fig. 2. In the first example, our model makes an accurate prediction in the first frame but fails at the following time steps due to the gaze movement. In the second example, the the camera view and gaze target move from the left to the right. This drastic change causes the mistake in our model's predictions. Similar failures also happen in the predictions of all baselines. Notably, existing deep models tend to only successfully anticipate steady gaze fixations or small gaze movements in the near future, and can not effectively capture large gaze shifts. This is the a common limitation shared by many existing works of future anticipation [13] in egocentric videos.

D More Implementation Details

D.1 Implementation Details of Our Model

Architecture. Inspired by [7], we use a light-weight audio encoder composed of four self-attention blocks from MViT [4]. The model architecture is further detailed in Tab. 5. We initialize the video encoder with Kinetics-400 pretraining [14] and initialize the audio encoder using Xavier initialization [8]. The resulting video embeddings $\phi(x)$ have a dimension of $T = 4, H = 8, W = 8, D = 768$, and the resulting audio embeddings $\psi(a)$ have a dimension of $T = 4, M = 64, D = 768$. We follow [18] to map audio-visual representation vectors to dimension $D' = 256$ for the contrastive loss. The output from the decoder is a downsampled heatmap which is upsampled to match the input size using trilinear interpolation. Following [16], we add intermediate features from each video encoder block to the corresponding decoder block output via skip connections to compensate for the loss of low-level textures.
Training. We set both temperature factor $\mathcal{T}$ of contrastive loss and re-weight parameter α as 0.05. Follow [16, 17], we use a Gaussian distribution with kernel size of 19 centered on the gaze location in each frame as the ground truth gaze heatmap during training. The model is trained with AdamW [20] optimization for 15 epochs. The momentum and weight decay are set as 0.9 and 0.05. The initial learning rate is 10^{-4} which decreases with the cosine learning rate decay strategy. The model is trained with a batch size of 8 across 4 GPUs.

D.2 Implementation Details of Baseline Fusion Strategies

We compare with multiple different audio-visual fusion strategies in main paper Tab. 2. The details of each baseline are listed as follows:
Linear. We reshape the video embedding and audio embedding to the shape $\hat{N} \times D$. We concatenate the two reshaped embeddings (resulting in the dimension of $\hat{N} \times 2D$) and input it to two linear layers. The dimension of the output is $\hat{N} \times D$ and we reshape it back to $T \times H \times W \times D$ which is fed into the decoder.

Bilinear. We reduce the length of video tokens and audio tokens to 256 using a linear layer respectively. Then we input the resulting video and audio tokens into a bilinear layer. The output is fed into the decoder for gaze forecasting.
Concat. We reshape the audio embedding $\psi(a) \in \mathbb{R}^{T\times H\times W\times D}$ to the same dimension as the video embedding $\phi(x) \in \mathbb{R}^{T\times H\times W\times D}$ and concatenate them along the channel to obtain an audio-visual representation with dimension of $T \times H \times W \times 2D$. This representation is fed into the decoder for gaze forecasting.
Vanilla SA. In this baseline, we flatten the video embedding and audio embedding into a list of tokens and thereby obtain $T \times (N + M)$ tokens in total, where $N = H \times W$. Then we input all tokens to a standard self-attention layer followed by multiple linear layers to perform fusion in the spatial and temporal dimensions simultaneously. We split the output into a new visual embedding incorporating audio information with dimension of $T \times N \times D$ and a new audio embedding incorporating visual information with dimension of $T \times M \times D$. The new visual embedding is input into the decoder.
STS. This is a baseline using the same fusion strategy as our method but without using the contrastive loss for training.

E Limitation and Future Work

In this paper, we propose a novel contrastive spatial-temporal separable fusion model for audio-visual egocentric gaze anticipation. Our method is validated on the Ego4D [9] and Aria [21] datasets. Our method has larger performance improvement on the Aria dataset comparing with Ego4D dataset. We believe this is because the multi-person social interaction setting from Ego4D dataset incurs additional challenges for audio representation learning, like multiple people and speakers present. Our current model design did not explicitly address this challenging nature of multi-speaker social interactions. Another limitation is that our model fails to anticipate the drastic gaze movements (see the failure cases in Fig. 2). In addition, in this work we do not explore the spatial geometry context provided by multi-channel audio signals. Our approach and experiments sugggest several important future research directions:

- The proposed CSTS model can be applied to other video understanding tasks related to the audio modality, such as action recognition, action localization, and video question answering. We hope to further investigate our proposed approach on these problem settings.
- A model explicitly designed for audio-visual representation learning in multi-person, multi-speaker environments merits further investigation.
- A model that learns better temporal representations for anticipating large gaze shifts remains to be explored.
- The visualization of correlation weights in the spatial fusion module indicates the potential of our model for weakly-supervised/self-supervised sound localization and active speaker detection, which can be investigated in further work.

F Code and License

The usage of the Aria dataset is under the Apache 2.0 License[1], and the usage of the Ego4D dataset is under the license agreement[2]. Our implementation is built on top of [3], which is under the Apache License[3]. Our code and the train/test split on Aria dataset will be available at: `https://bolinlai.github.io/CSTS-EgoGazeAnticipation/`.

References

1. Chang, Q., Zhu, S.: Temporal-spatial feature pyramid for video saliency detection. Cognitive Computation (2021)
2. Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., Onoe, N.: M2fnet: Multi-modal fusion network for emotion recognition in conversation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4652–4661 (2022)
3. Fan, H., Li, Y., Xiong, B., Lo, W.Y., Feichtenhofer, C.: Pyslowfast. `https://github.com/facebookresearch/slowfast` (2020)
4. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6824–6835 (2021)
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
6. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 35–53 (2018)
7. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10457–10467 (2020)
8. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
9. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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. pp. 18995–19012 (2022)
10. Huang, P.Y., Sharma, V., Xu, H., Ryali, C., Li, Y., Li, S.W., Ghosh, G., Malik, J., Feichtenhofer, C., et al.: Mavil: Masked audio-video learners. Advances in Neural Information Processing Systems **36** (2024)
11. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. In: Proceedings of the European conference on computer vision (ECCV). pp. 754–769 (2018)

1 `https://github.com/facebookresearch/vrs/blob/main/LICENSE`
2 `https://ego4d-data.org/pdfs/Ego4D-Licenses-Draft.pdf`
3 `https://github.com/facebookresearch/SlowFast/blob/main/LICENSE`

12. Jain, S., Yarlagadda, P., Jyoti, S., Karthik, S., Subramanian, R., Gandhi, V.: Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3520–3527. IEEE (2021)
13. Jia, W., Liu, M., Rehg, J.M.: Generative adversarial network for future hand segmentation from egocentric video. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII. pp. 639–656. Springer (2022)
14. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
15. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5492–5501 (2019)
16. Lai, B., Liu, M., Ryan, F., Rehg, J.: In the eye of transformer: Global-local correlation for egocentric gaze estimation. British Machine Vision Conference (2022)
17. Li, Y., Liu, M., Rehg, J.: In the eye of the beholder: Gaze and actions in first person video. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
18. Lin, K.Q., Wang, A.J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. Advances in Neural Information Processing Systems (2022)
19. Liu, Y., Tan, Y., Lan, H.: Self-supervised contrastive learning for audio-visual action recognition. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 1000–1004. IEEE (2023)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Lv, Z., Miller, E., Meissner, J., Pesqueira, L., Sweeney, C., Dong, J., Ma, L., Patel, P., Moulon, P., Somasundaram, K., Parkhi, O., Zou, Y., Raina, N., Saarinen, S., Mansour, Y.M., Huang, P.K., Wang, Z., Troynikov, A., Artal, R.M., DeTone, D., Barnes, D., Argall, E., Lobanovskiy, A., Kim, D.J., Bouttefroy, P., Straub, J., Engel, J.J., Gupta, P., Yan, M., Nardi, R.D., Newcombe, R.: Aria pilot dataset. `https://about.facebook.com/realitylabs/projectaria/datasets` (2022)
22. Min, X., Zhai, G., Zhou, J., Zhang, X.P., Yang, X., Guan, X.: A multimodal saliency model for videos with high audio-visual correspondence. IEEE Transactions on Image Processing **29**, 3805–3819 (2020)
23. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Advances in neural information processing systems **34**, 14200–14213 (2021)
24. Praveen, R.G., de Melo, W.C., Ullah, N., Aslam, H., Zeeshan, O., Denorme, T., Pedersoli, M., Koerich, A.L., Bacon, S., Cardinal, P., et al.: A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2486–2495 (2022)
25. Senocak, A., Kim, J., Oh, T.H., Li, D., Kweon, I.S.: Event-specific audio-visual fusion layers: A simple and new perspective on video understanding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2237–2247 (2023)
26. Tavakoli, H.R., Borji, A., Rahtu, E., Kannala, J.: Dave: A deep audio-visual embedding for dynamic saliency prediction. arXiv preprint arXiv:1905.10693 (2019)

27. Tsiami, A., Koutras, P., Maragos, P.: Stavis: Spatio-temporal audiovisual saliency network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4766–4776 (2020)
28. Wang, G., Chen, C., Fan, D.P., Hao, A., Qin, H.: From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15119–15128 (2021)
29. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12695–12705 (2020)
30. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740 (2020)
31. Xiong, J., Wang, G., Zhang, P., Huang, W., Zha, Y., Zhai, G.: Casp-net: Rethinking video saliency prediction from an audio-visual consistency perceptual perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6441–6450 (2023)
32. Yao, S., Min, X., Zhai, G.: Deep audio-visual fusion neural network for saliency estimation. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1604–1608. IEEE (2021)
33. Zhang, M., Ma, K.T., Lim, J.H., Zhao, Q., Feng, J.: Anticipating where people will look using adversarial networks. IEEE transactions on pattern analysis and machine intelligence **41**(8), 1783–1796 (2018)
34. Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., Feng, J.: Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4372–4381 (2017)

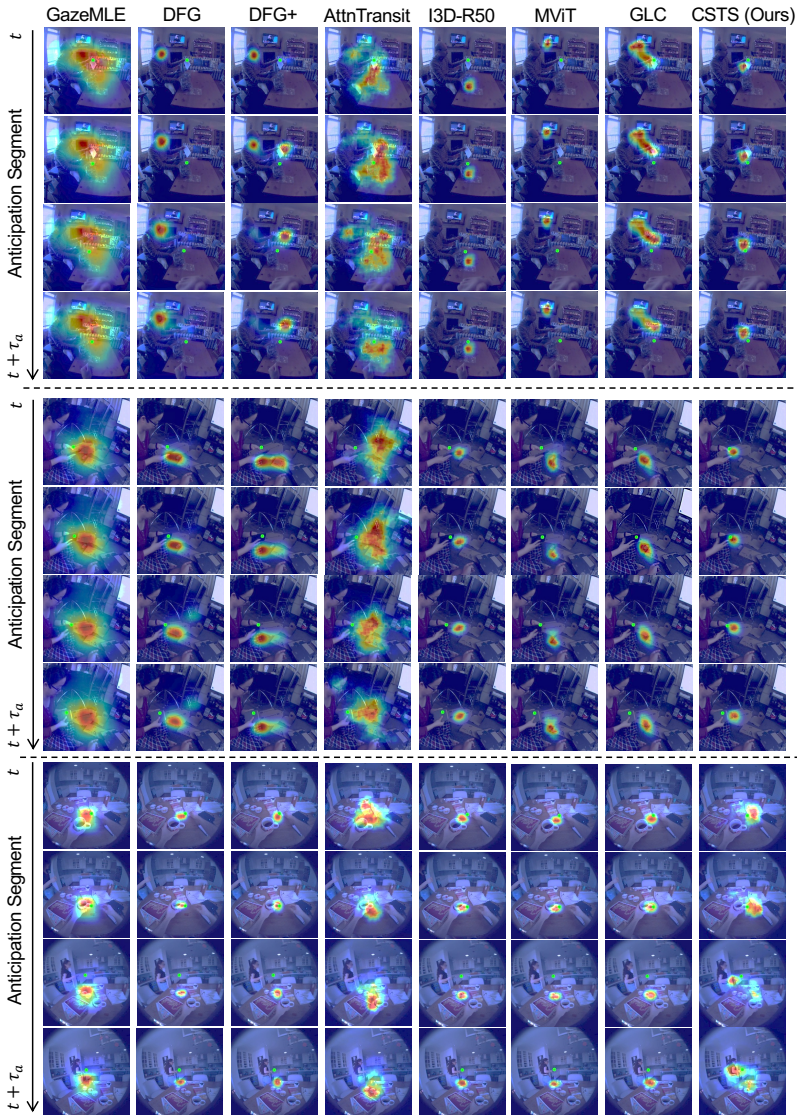

Fig. 1: Additional egocentric gaze anticipation results from our model and other baselines. Green dots indicate the ground truth gaze location. The first two examples are from the Ego4D dataset, and the last example is from the Aria dataset.

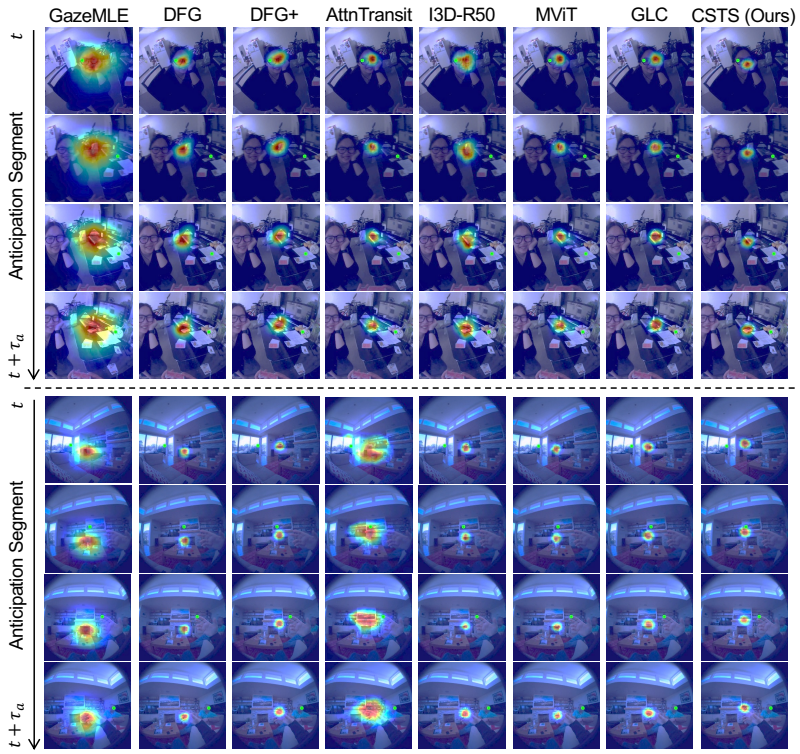

Fig. 2: Failure cases of our model and baselines. Green dots indicate the ground truth gaze location. The first example is from the Ego4D dataset, and the second example is from the Aria dataset.

	Stages	Operators	Output Size
Video Encoder $\phi(x)$	video frames	-	$8 \times 256 \times 256 \times 3$
	video token embedding	$Conv(3 \times 7 \times 7,\ 96)$ $stride\ 2 \times 4 \times 4$	$4 \times 64 \times 64 \times 96$
	tokenization	flattening	$(4 \times 64 \times 64) \times 96$
	video encoder block1	$\begin{bmatrix} MSA(96) \\ MLP(384) \end{bmatrix} \times 1$	$(4 \times 64 \times 64) \times 192$
	video encoder block2	$\begin{bmatrix} MSA(192) \\ MLP(768) \end{bmatrix} \times 2$	$(4 \times 32 \times 32) \times 384$
	video encoder block3	$\begin{bmatrix} MSA(384) \\ MLP(1536) \end{bmatrix} \times 11$	$(4 \times 16 \times 16) \times 768$
	video encoder block4	$\begin{bmatrix} MSA(768) \\ MLP(3072) \end{bmatrix} \times 2$	$(4 \times 8 \times 8) \times 768$
Audio Encoder $\psi(a)$	audio spectrograms	-	$8 \times 256 \times 256 \times 1$
	audio token embedding	$Conv(3 \times 7 \times 7,\ 96)$ $stride\ 2 \times 4 \times 4$	$4 \times 64 \times 64 \times 96$
	tokenization	flattening	$(4 \times 64 \times 64) \times 96$
	audio encoder block1	$\begin{bmatrix} MSA(96) \\ MLP(384) \end{bmatrix} \times 1$	$(4 \times 4096) \times 192$
	audio encoder block2	$\begin{bmatrix} MSA(192) \\ MLP(768) \end{bmatrix} \times 1$	$(4 \times 1024) \times 384$
	audio encoder block3	$\begin{bmatrix} MSA(384) \\ MLP(1536) \end{bmatrix} \times 1$	$(4 \times 256) \times 768$
	audio encoder block4	$\begin{bmatrix} MSA(768) \\ MLP(3072) \end{bmatrix} \times 1$	$(4 \times 64) \times 768$
Fusion Modules	conv1	$Conv(768 \times 1 \times 8 \times 8,\ 768)$ $stride\ 1 \times 1 \times 1$	$4 \times 1 \times 768$
	in-frame self-attention $\sigma(\cdot)$	$\begin{bmatrix} MSA(768) \\ MLP(3072) \end{bmatrix} \times 1$	$4 \times (64+1) \times 768$
	conv2	$Conv(768 \times 1 \times 8 \times 8,\ 768)$ $stride\ 1 \times 1 \times 1$	$4 \times 1 \times 768$
	conv3	$Conv(768 \times 1 \times 8 \times 8,\ 768)$ $stride\ 1 \times 1 \times 1$	$4 \times 1 \times 768$
	cross-frame self-attention $\pi(\cdot)$	$\begin{bmatrix} MSA(768) \\ MLP(3072) \end{bmatrix} \times 1$	$8 \times 1 \times 768$
	reweighting	$u_{v,s} \otimes u_{v,t}$	$8 \times 64 \times 768$
	reweighting	$\psi(a) \otimes u_{a,t}$	$8 \times 64 \times 768$
Decoder	decoder block1	$\begin{bmatrix} MSA(1536) \\ MLP(3072) \end{bmatrix} \times 1$	$(4 \times 16 \times 16) \times 768$
	decoder block2	$\begin{bmatrix} MSA(768) \\ MLP(1536) \end{bmatrix} \times 1$	$(4 \times 32 \times 32) \times 384$
	decoder block3	$\begin{bmatrix} MSA(384) \\ MLP(768) \end{bmatrix} \times 1$	$(4 \times 64 \times 64) \times 192$
	decoder block4	$\begin{bmatrix} MSA(192) \\ MLP(384) \end{bmatrix} \times 1$	$(8 \times 64 \times 64) \times 96$
	head	$Conv(1 \times 1 \times 1,\ 1)$ $stride\ 1 \times 1 \times 1$	$8 \times 64 \times 64 \times 1$

Table 5: Architecture of the proposed model. Convolutional layers are denoted as $Conv(kernel\ size,\ output\ channels)$. The number of input channels in multi-head self-attention is shown in the parenthesis of MSA. The dimension of the hidden layer in multi-layer perceptron is listed in parenthesis of MLP. conv1 is the convolutional layer in the spatial fusion module. conv2 and conv3 are convolutional layers in the temporal fusion module.