Enhancing 3D Gaze Estimation in the Wild using Weak Supervision with Gaze Following Labels

Supplementary Material

What is expected? The supplementary material consists of datasets details, experiments details, and extended experiments analysis mentioned in the main paper. In addition, videos of qualitative examples of our method on VideoAttentionTarget further demonstrate the robustness in challenging real-world scenarios.

A. Datasets Details

A.1. Datasets

Gaze360 (G360). [28] is video 3D gaze datasets. It is collected in both indoor and outdoor environments in unconstrained setting, which contains 3D gaze of 238 subjects with a wide-range head pose and gaze direction. G360 is recorded at 8FPS. In all of our experiments, we always used the same training set as [28] with 126928 samples. For the test set, we followed the split of [28] where G360 Full corresponds to "All 360°" (the entire test set) with 25969 samples, G360 180 corresponds to "Front 180°" (gaze within 90°) with 20322 samples, and G360 40 to "Front Facing" (gaze within 20°) with 3995 samples. In addition to those splits, we consider G360 Back (gaze above 90°) [8] with 5647 samples and finally G360 Face (all detected faces) with 16031 samples, which is used in many constrained gaze studies [1, 7, 9–11, 18, 55, 61]. When we refer to G360 Face 180 (15895 samples), it corresponds to the detected face with a gaze within 90°, a subset of G360 180, the same for G360 Face 40 with 3687 samples. We used the validation set described in [28] with 17038 samples.

GFIE. [25] is a video 3D gaze dataset collected indoors with 71799 frames from 61 subjects (27 male and 34 female). It is an unconstrained dataset with a wide range of head poses. It was collected for gaze following task; using a complex calibrated laser setup, they can infer the 3D gaze from the eye to the visual target direction. They recorded people doing various indoor activities at 30 fps. We follow the data splits described in [25], 59217 for training, 6281 for validation, and 6281 for testing.

MPSGaze (MPS). [59] is a modified 3D gaze datasets that has been automatically generated using ETH-Xgaze [63] eyes. They apply a blending technique on people from the Widerface [56] dataset to put eyes with a known 3D gaze from ETH on heads with similar head poses. This dataset is diverse, with more than 10k identities and challenging poses, appearances, and lighting conditions. However, the blending process reduces the quality of the visual appearance, and it contains only near frontal head poses and no back view. We used the same training and test split with 24282 samples in training and 6277 samples in testing. No validation is defined in this work.

EYEDIAP (EDIAP). [19] is a 3D gaze video dataset. It includes videos from 16 subjects (30 fps), using either screen targets (CS, DS subset EDIAP) or 3D floating balls (FT subset EDIAP-FT) as gaze targets. It is a constrained setup with mainly frontal head poses. Following [12, 54], we used the evaluation set under screen target session (CS, DS, namely EDIAP) with 16674 samples from 14 subjects.

MPIIFaceGaze (MPII). [60] is a 3D gaze image dataset collected from 15 subjects in a screen-based gaze target setup, resulting in a constrained dataset with mostly frontal head pose. We follow the standard evaluation protocol [12, 54, 60], which selects 3000 images from each subject to form an evaluation set for a total of 45000 samples.

GazeFollow (GF). [43] is a 2D gaze image dataset annotated on in the wild dataset for the gaze the following task. The 2D target label corresponds to where a given person is looking at in the image. It is a diverse dataset that includes various head poses, appearances, scenes, and lighting conditions. Overall, it has around 130K annotated person-target instances in 122K images.

A.2. Video Processing

As mentioned in the main section, for video clip input, our approach predicts the 3D gaze from an 8-frame video clip. However, video datasets have different frame rates, which can impact the gaze prediction. In this work, since G360 has a lower frame rate, we resample EYEDIAP and GFIE to match G360's frame rate of 8 fps.

A.3. Gaze Representation

Working with different 3D gaze datasets requires a unified way to define and represent the 3D gaze vector. Usually, in constrained gaze estimation, studies use data normalization to map the input image to a normalized space where a virtual camera is used to warp the face patch out of the original input image according to the 3D head pose [63]. Thus, the gaze is expressed in this virtual camera coordinate defined by the 3D head pose.

However, in unconstrained settings, it is not possible to get access to a robust and reliable 3D head pose; thus, we follow the gaze representation of Gaze360 [28] in the "Eye

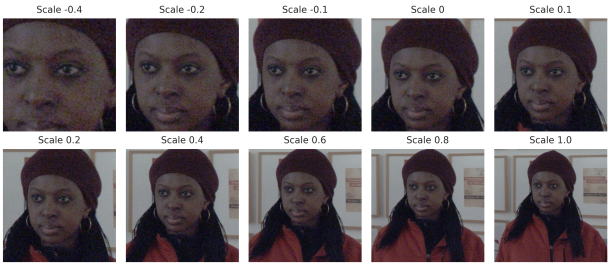


Figure S1. Input head crop using different scales. In our work, a scale of -0.1 is used and proved to be effective in both constrained and frontal face setting Sec. C.1

coordinate system". The practical interpretation of the eye coordinate system is that the positive x-axis points to the left, the positive y-axis points up, and the positive z-axis points away from the camera, *i.e.* [-1,0,0] is a gaze looking to the right or [0,0,-1] straight into the camera from the camera's point of view, irrespective of subjects position in the world. The origin of the gaze vector is the middle of the eyes, except for MPS and MPII, where the gaze origin is the average of 3D eyes and mouth landmarks resulting in an origin located at the middle of the nose, and for GF, we used the center of the head bounding box as the origin.

B. Experiments Details

Metric. We follow the test split described in the state-of-the-art method and explained in Sec. A.1. As a metric, we use the standard angular error in degrees between the predicted and ground truth gaze prediction [19, 28, 60, 63]. Previous methods reporting video evaluation used a 7-frame video clip and predict the middle frame gaze direction. Since our approach outputs eight gaze directions from an 8-frame video clip, for a fair comparison, we use the 4th gaze prediction of an 8-frame video clip to compute the evaluation metric.

Training. We used the same setup in all the experiments to be as fair as possible. All the models are trained for a minimum of 20 epochs. We used an early stopping on the validation set with a patience of 10 epochs. We use the AdamW optimizer [37] with a learning rate of 1e-4 and a cosine annealing schedule with a 5 epochs linear warmup (from 2e-5 to 1e-4). For evaluation, we report the performance of the best model defined by the best angular error on the validation set.

Data augmentation. Data augmentation is crucial for robust gaze estimation in the wild. In this work, we used standard data augmentation techniques. First, we applied jittering during the head crop to introduce slight variations in scale and aspect ratio, which reduces the model's sensitivity to noisy or imprecise head bounding boxes. Next, color jittering was applied by adjusting brightness, contrast, and saturation, making the model more resilient to diverse lighting conditions commonly encountered in real-world scenarios. Since gaze labels, such as those in the GF 2D dataset, may exhibit bias toward one side, we applied horizontal flipping to the images while appropriately adjusting the gaze direction, ensuring more balanced training data in the yaw gaze direction. These augmentations collectively improved the model's ability to handle variations in data and enhance its generalization to unseen environments.

Method	Training Dataset	MPII Img	EDIAP Img	EDIAP Vid
PureGaze [13] (Res18)	G360I Face	9.3	9.2	-
Liu *et al.* [34] (Res18)	G360I Face	7.7	9.0	-
Liu *et al.* [34] (Res50)	G360I Face	8.3	7.5	-
RAT [4] (Res18)	G360I Face	7.6	7.1*	-
RAT [4] (Res50)	G360I Face	7.7	7.1*	-
CDG [54] (Res50)	G360I Face	7.0	**7.3**	-
Supervised (GaT)	G360I&V	7.43	8.88	8.28
ST-WSGE (GaT)	G360I&V+GF	**6.43**	8.87	8.19

Table S1. **Comparision with state-of-the-art on constrained domain generalization benchmarks**. All these methods [4, 13, 34, 53, 54] use a face crop as input and are trained on the detected face subset of Gaze360. Our method is trained and tested on head crop which makes it more general but more challenging for frontal gaze estimation. * In [4] they used only 6400 sample for EDIAP but we follow [12, 13, 54] with 16674 samples.

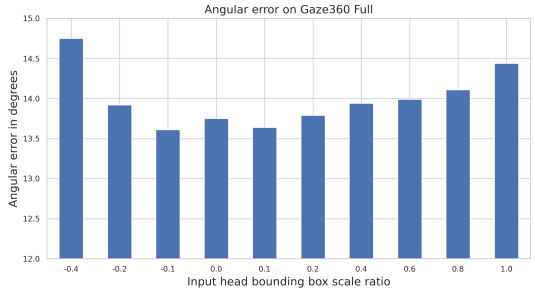
C. Additional Experiments

C.1. Effect of Head Crop Size

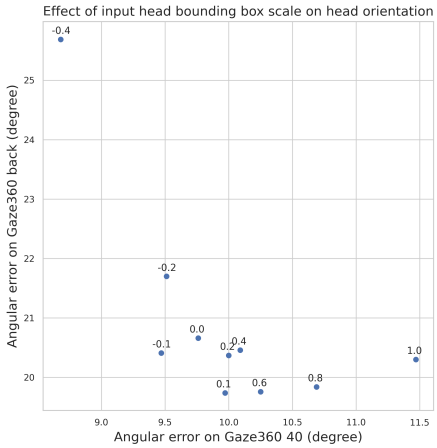
As mentioned by Chen *et al.* [8], the input head crop scale impacts the 3D gaze estimation. We find that the effect on the prediction depends on the head orientation. Fig. S1 illustrates the different inputs with different head crop scales. As shown in Fig. S2b, a smaller head crop tighter to the face improves 3D gaze estimation on frontal head poses, while a larger head crop improves gaze on the non-frontal head pose. Indeed, as shown in Fig. S1, a tighter crop increases the eye resolution in the image and a larger crop provides more context about the head orientation and upper body orientation, which gives a strong prior for the gaze direction when eyes are not visible. In the context of gaze estimation in the wild, a scale of -10% is part of the Pareto front as illustrated in Fig. S2b and is also the best on the G360 Full image as shown in Fig. S2a. Therefore, it is a reasonable trade-off between frontal and back view performance. We use it for all our experiments.

C.2. Constrain Gaze Evaluation

The objective of this work is to improve unconstrained gaze estimation in the wild. As seen in Sec. C.1, compared to a tight face crop a larger crop improves gaze in challenging head pose. Therefore, a larger crop is more suited to our objective. In contrast, some methods specialize in frontal



(a) Effect of head bounding box scale as input on the 3D gaze angular error on G360 Full test set. A scale ratio of 0.1 corresponds to a 10% bounding box scale.



(b) Effect of head bounding box scale on the angular error with respect to G360 Back and G360 40 test subset.

Figure S2. **Effect of head crop size.**

gaze estimation and rely on tight face crops, which provide better resolution for the eye regions. While this is not a fully fair comparison, we compare our approach to these constrained methods for generalization on constraint benchmarks. Note that for the constrained methods, models are trained and tested only on a subset of detected faces (G360 Face), while in our approach the model is trained on G360 Full.

As shown in Tab. S1, on MPII, the supervised GaT lags behind the best method by 6%. On EDIAP, GaT is 21% behind the best method in image evaluation but narrows the gap to 13% when evaluated on videos. Then, when using our ST-WSGE learning framework including GF labels, we observe an important improvement on MPII with state-of-the-art angular error of 6.43 compared to 7 from CDG. On EDIAP the improvement is marginal. Compared to EDIAP, MPII has more diversity in lighting conditions and environment. GF doesn't contain a lot of frontal gaze direction but has a broad diversity of environments. Therefore, the improvement on MPII should come from the additional diversity that GF brings but this is not useful for EDIAP prediction. While constrained methods excel in frontal settings, they fail in unconstrained scenarios. Our approach, which achieves state-of-the-art performance in unconstrained environments (G360, GFIE) while remaining competitive in constrained settings (MPII, EDIAP), proves to be a versatile and robust solution for gaze estimation in the wild.

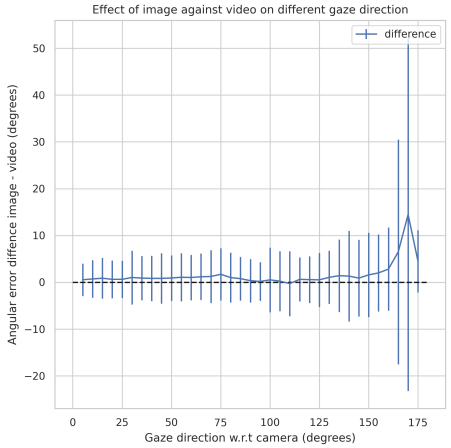


Figure S3. **Image vs video predictions, where does it help?**. GaT trained on G360I&V and tested on G360 Full image and video. The difference between image and video angular error with respect to the ground truth gaze directions from the camera ([0,0,-1]). The mean and standard deviation are displayed for each 10° bin. Positive values indicate better performance in video prediction compared to image prediction.

C.3. Qualitative Analysis

When does temporal context contribute most effectively? As seen in the main paper, video prediction consistently outperforms image prediction. To understand the significance of temporal context in gaze estimation, we examined cases with large angular errors between image and video predictions. Several key observations emerged. As illustrated in Fig. S4 in the first two rows, temporal context proves valuable during blinks, as it allows the model to interpolate gaze direction when the eyes are closed. If the head pose is not informative, temporal context helps disambiguate between blinking and looking down since the eyes are not visible, as shown in row 1. Additionally, when individuals are viewed entirely from behind (rows 6-7), video inferences provide a more consistent gaze direction in relation to time. Thus, there is less jittering and it might improve the prediction accuracy. In rows 4-5, the head and eye motion can be used in video prediction to improve the gaze direction. Finally, it can help in case of occlusion, as seen in row 3.

Furthermore, we explore the impact of image- and video-based prediction with respect to gaze direction. Indeed, we expected more improvement when people are from the back since additional head motion cues can be useful for gaze estimation. In the results, video prediction on G360 Back clearly improves image prediction. In addition, in Fig. S3, we plot the difference between image and video prediction angular error for different gaze directions. If we look at the trend, video prediction seems to be better, especially for gaze over 150°, but given the standard deviation, it might not be a statistically significant observation. A more detailed analysis by considering only cases where there is a head motion can better highlight the impact of video prediction.

What are the limitations of temporal context for gaze? We investigate prediction made on the VideoAttentionTarget [15] (VAT) videos using our ST-WSGE framework and GaT model. VAT is a challenging dataset with real-world scenarios, various appearances, and diverse gaze distribution, making it well-suited for assessing our approach. Our qualitative analysis reveals two limitations of video-based inference compared to image-based inference using our model. The first limitation arises in cases of rapid head rotation, as illustrated in Fig. S5, temporal context may be misused, leading to predictions that do not align with the actual gaze. It might be because no rapid head motion is present in the G360 training sets. The second aspect involves cases of "gaze recentering", where the gaze direction returns to its initial position following a shift. This behavior can occur very rapidly, within just 3-4 frames. Due to the smoothing effect in the temporal modeling, the predicted gaze may not exhibit the same amplitude as the actual movement. Indeed, this behavior is not present in the G360 dataset, and the use of videos sampled at 8 frames per second may limit the ability to capture fine-grained gaze dynamics. However, such behavior is better captured during image-based inference. This highlights a trade-off: while video-based inference provides smoother and more robust predictions, image-based inference offers greater accuracy but can result in jittery outputs. To mitigate the lack of natural gaze behavior we apply our ST-WSGE framework using 2D gaze video data from VAT. Unfortunately, since current benchmarks don't contain natural gaze behavior, the results don't show quantitative improvement. Further research to evaluate this aspect is needed.

In which scenarios does ST-WSGE with GazeFollow labels provide the most benefit? We demonstrated the advantages of ST-WSGE with GazeFollow labels across various benchmarks, both within- and cross-datasets. But in which scenarios does it outperform supervised methods trained solely on G360? To address this question, we analyze predictions made in real-world scenarios using the VideoAttentionTarget (VAT) dataset [15]. Our findings reveal that ST-WSGE achieves the most notable improvements in cases of extreme head poses, particularly when the head is facing downward, as shown in Fig. S6. It is also more robust to appearance diversity like hair partially occluding the face or varying skin tones. It also helps in difficult lighting conditions and low-resolution inputs. Additionally, we include a video (provided in the supplementary materials) displaying predictions on VAT with an explanation, enabling a direct comparison between the two methods and a clearer visualization of our approach's performance on real-world data.

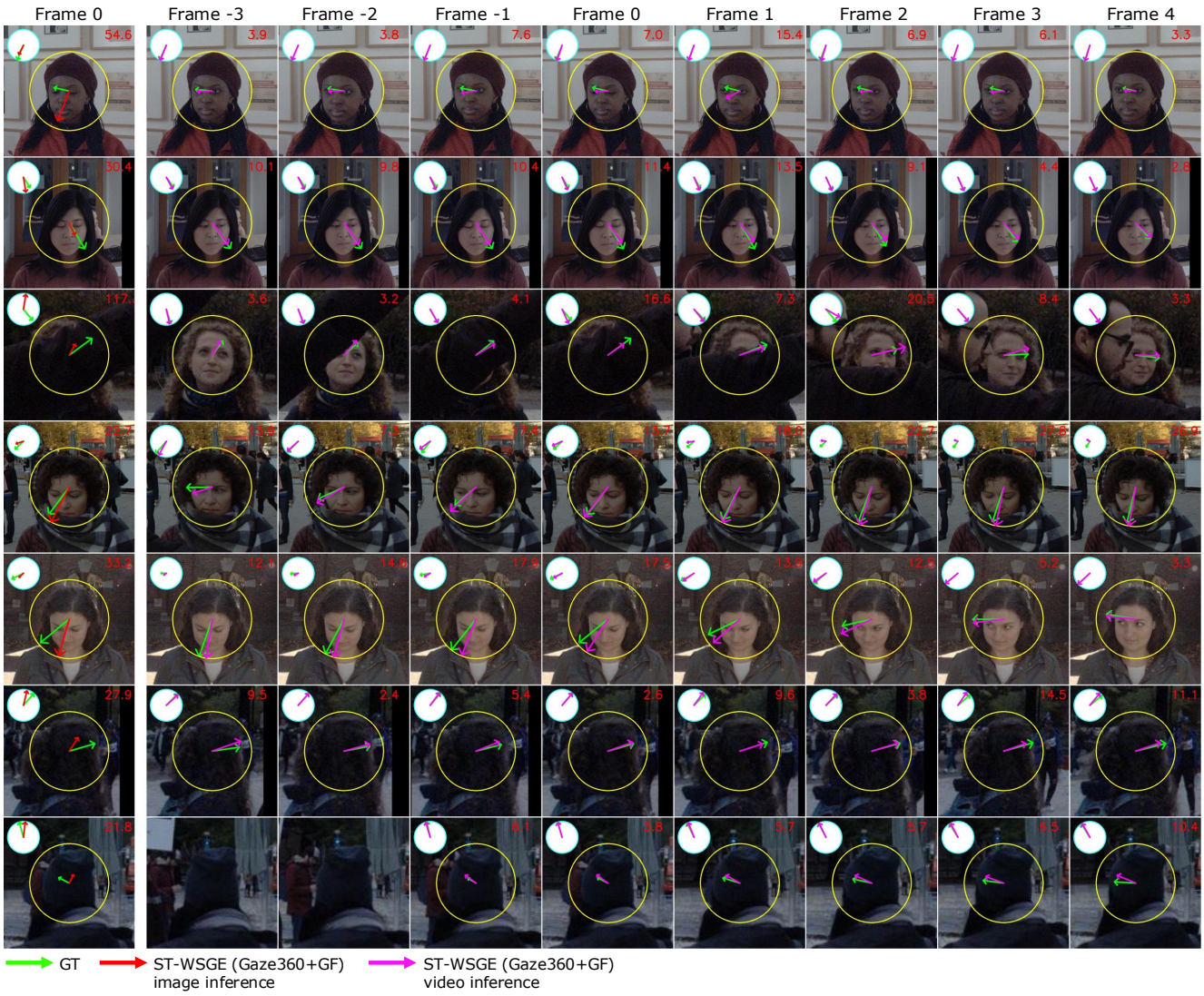


Figure S4. **Illustration of image against video prediction.** Comparison between single-image (frame 0) and video predictions (frame -3 to 4). We use our ST-WSGE learning framework with GaT trained on G360 and GF. All examples are from G360 test set. Rows 1-2 illustrate eye blinks, Row 3 shows an example of occlusion, Rows 4-5 demonstrate frontal head/eyes motion, and Rows 6-7 depict back view prediction. In the last row, the first two frames are not part of the test subset. Arrows in red represent image predictions, and arrows in magenta are video predictions. The angular error between groundtruth and prediction is displayed in red at the top right corner. The circles in the images represent unit disks where 3D gaze vectors are projected onto the image plane (x,y in yellow) and a top view (x,z in blue)

Figure S5. **Illustration of image and video prediction in case of rapid head motion.** We use our ST-WSGE learning framework with GaT trained on G360 and GF. All examples are from VideoAttentionTarget [15] (VAT). Arrows in red represent image predictions, and arrows in magenta are video predictions. The circles in the images represent unit disks where 3D gaze vectors are projected onto the image plane (x,y in yellow) and a top view (x,z in blue). Note that since VAT has a frame per second (fps) of 24 and G360 has a fps of 8, we show the temporal context used for video inference corresponding to 8 fps.

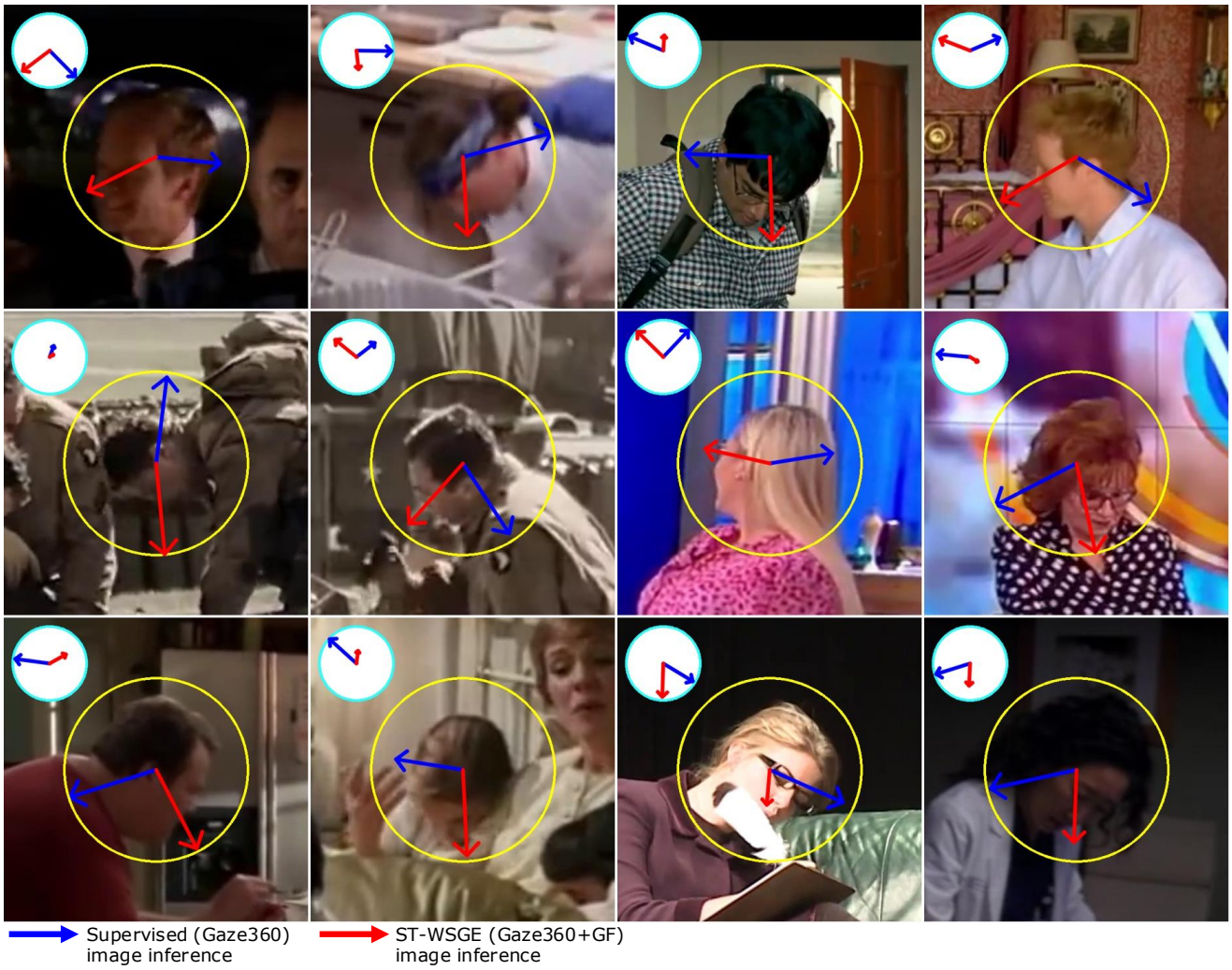


Figure S6. **Illustration of supervised against ST-WSGE learning framework with GazeFollow label.** We use in both experiments our GaT model. All examples are from VideoAttentionTarget [15] (VAT). Arrows in blue represent image predictions with supervised GaT trained on G360, and arrows in red are image predictions with ST-WSGE GaT trained on G360 and GF. The circles in the images represent unit disks where 3D gaze vectors are projected onto the image plane (x,y in yellow) and a top view (x,z in blue).