Diffusion-Refined VQA Annotations for Semi-Supervised Gaze Following: Supplementary Material

Abstract. In this supplementary material, we provide more details about the experiment settings of the main paper (Sec. S1), we describe the baseline implementation details (Sec. S2), we illustrate the details of the Grad-CAM heatmap generation procedure (Sec. S3), we present the results for each video in the video semi-supervised learning task (Sec. S4), we discuss the reason for the lower AUC of the diffusion models in the fully supervised experiments (Sec. S5), we show the result of training with 50% annotations on GazeFollow (Sec. S6) and training with 10% labels on VideoAttentionTarget using the normal train-test split (Sec. S7), we applied our methods to other gaze following models (Sec. S8), we analyzed the effect of the amount of unlabeled data by fixing the amount of labeled data (Sec. S9), we visualize outputs from the diffusion model at intermediate inference steps (Sec. S10), we provide detailed discussions on the available semi-supervised learning methods and their relations to gaze following (Sec. S11) and evaluate the raw Grad-CAM heatmaps when obtained with different extraction methods (Sec. S12).

S1 Training Details

Here we provide more details of the training parameters for the fully-supervised and semi-supervised settings on the GazeFollow dataset as well as semi-supervised training on the VideoAttentionTarget dataset.

On the GazeFollow dataset, when training the teacher models with supervised data only, the batch size was set at 48 for all models (VAT, VAT-GC, and Diffusion model). The learning rate for training VAT and VAT-GC was 2.5×10^{-4}. For the training of the diffusion model, we used a learning rate of 5×10^{-5}. During the evaluation of the fully-supervised diffusion models in the ablation studies, we averaged 10 predictions from the diffusion model for each image to account for the variability introduced by the stochastic sampling process. In semi-supervised training, we used a learning rate of 2.5×10^{-4}, with a decay factor of 0.2 at the 15th epoch when training with 5%, 10% and 20% of annotations, and 25th epoch when training with 50% of annotations. We randomly sampled 20 samples from the pool of labeled data in each batch when training with 5%, 10%, and 20% of annotations to stabilize the training, following the training procedure of Mean Teacher [33]. When refining the Grad-CAM heatmaps, we added noise at $t = 250$ when training with 10%, 20% and 50% annotations, and $t = 200$ when training with 5% annotations due to the weaker annotation prior when training the diffusion model with less labeled data.

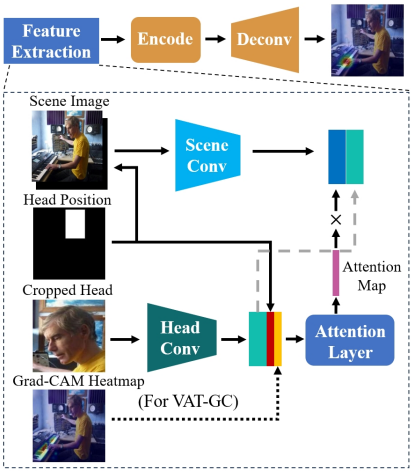

Fig. S1: Structure of the VAT and VAT-GC baselines. The model consists of a feature extraction module, an encoder, and a deconvolution module to output the gaze target heatmap. The figure shows the model's structure. For video gaze following, a Conv-LSTM layer is added between the encoder and deconvolution module. For VAT-GC, the Grad-CAM heatmap is concatenated with the extracted head feature to serve as conditional input.

In the task of semi-supervised training on videos, when training the teacher models, we tuned the learning rates for each video specifically based on the training loss, ranging from 5×10^{-6} to 5×10^{-5}. When training the student model in semi-supervised training, for each video, we set the same learning rate for all teachers. The learning rates range from 1×10^{-5} to 5×10^{-5}.

S2 Baseline Implementation Details

In this section, we provide the implementation details of the baseline teacher models. VAT [4] is a popular gaze following model that achieves close-to-human performance without using additional modalities. The structure of the model is shown in Fig.S1. The model consists of a feature extraction module, an encoder, and a deconvolution module that outputs the heatmap. The feature extraction module takes a scene image, a binary head position mask, and the cropped head of the person as input, and outputs the extracted features from two pathways for encoding the scene and gaze features. The original model consists of two output branches for gaze target heatmap prediction and in/out prediction. In our experiments, we only used the heatmap prediction branch as we were focused only on the gaze target prediction task. When training the diffusion model, we used the same feature extraction module as the VAT model to extract conditional features that are input to the denoiser U-Net.

For VAT-GC, we modified the head pathway of the VAT model to take the raw Grad-CAM heatmap as additional input, as shown in the dashed lines of

Fig.S1. The attention layer was also modified according to the new input dimension.

For VAT-MT, we used the Mean Teacher method [33] to train the VAT model on both labeled and unlabeled data. The student VAT model is trained with a combination of an L2 loss on labeled data, and a consistency loss between the outputs from the student and teacher model on all data. We used KL-Divergence as the consistency loss. We trained GCDR-MT in a similar manner by using the diffusion model as the student model.

S3 Details of Grad-CAM Heatmap Generation

In this section, we provide a detailed description of the procedure used to extract the raw Grad-CAM heatmaps from the OFA model [34]. The OFA model has a transformer-based encoder-decoder structure. Given an image with the target person overlaid by the detected bounding box and the gaze following question Q, it first uses a ResNet [7] to extract visual features, which are then flattened into N_v patch tokens: $\mathcal{V} = \{\boldsymbol{v}_i \in R^d\}_{i=1}^{N_v}$. Meanwhile, it uses byte-pair encoding (BPE) [29] to transform the text sequence of the question Q into a subword sequence and then embed them into text tokens $\mathcal{Q} = \{\boldsymbol{q}_j \in R^d\}_{j=1}^{N_q}$. The visual and text tokens are then concatenated and fed into the transformer encoder. The transformer decoder takes in the embeddings from the encoder output, and generates output tokens auto-regressively, starting with a specific [BOS] token as the query.

After getting the output tokens from the VQA model, we compute the Grad-CAM heatmap $\boldsymbol{g}$ on the decoder cross-attention weights between the last input query token, and the image patch tokens when generating the word-of-interest during the auto-regressive generation process. In most cases, the answer is just a single noun, such as *"food"* and *"plant"*. In such cases, we directly compute the Grad-CAM heatmap when the VQA model outputs these specific tokens. When the answer contains multiple words, such as *"The girl on the left"*, we use parts-of-speech (POS) tagging using spaCy [9] to find the word of interest. Specifically, each word is assigned a POS label, and only the words with a label of either noun or proper noun are selected [30]. When the answer contains multiple nouns, such as *"The boy wearing glasses"*, we select the first noun as it is the representative word in most cases. Finally, the Grad-CAM heatmaps are linearly interpolated to have the same size as the ground truth heatmaps in the training set.

S4 Detailed Results of the Video Experiment

In this section, we provide detailed results of the semi-supervised learning experiments on VideoAttentionTarget, including the exact performance of all methods, as well as the full video names and the number of annotations used in each video. The results are shown in Tab.S1.

As discussed in the main paper, our method performs the best on almost all videos. We would like to note that, in real-world applications when videos are

Table S1: Performance of different methods for each video in semi-supervised finetuning. We also show the number of annotations used for fine-tuning and the ratio of the visible annotations with respect to all annotations in the entire video. GCDR performs best on almost all videos in both metrics.

Video	# Annot. Visible	Annot. Ratio	VAT		Semi-VAT		Semi-VAT-GC		GCDR	
			Dist. ↓	AUC ↑	Dist.↓	AUC ↑	Dist. ↓	AUC ↑	Dist. ↓	AUC ↑
West World	109	11.61%	0.253	0.619	0.255	0.659	0.232	0.709	**0.186**	**0.798**
Titanic	114	3.36%	0.222	0.752	0.208	0.781	0.201	0.777	**0.182**	**0.799**
Hell's Kitchen	101	23.38%	0.239	0.732	0.243	0.774	0.224	0.835	**0.176**	**0.855**
I Wanna Marry Harry	74	9.18%	0.177	0.700	0.168	0.716	0.163	0.732	**0.159**	**0.768**
It's Always Sunny in Philadelphia	116	5.02%	0.188	0.756	0.180	0.785	0.176	0.782	**0.157**	**0.835**
Survivor	73	3.99%	0.150	0.762	0.147	0.782	**0.129**	0.829	0.147	**0.849**
Downton Abbey	108	3.54%	0.121	0.871	0.120	**0.888**	0.123	0.877	**0.118**	0.854
Jamie Oliver	70	6.30%	0.125	0.823	0.129	0.828	0.121	0.821	**0.116**	**0.838**
CBS This Morning	96	2.21%	0.099	0.844	**0.101**	0.860	0.105	0.902	0.105	**0.913**
MLB Interview	95	3.86%	0.086	0.872	0.087	0.899	0.077	0.915	**0.076**	**0.921**

collected from a new scene, users can compare the raw Grad-CAM heatmaps with the annotations from a few annotated frames. Based on the Grad-CAM heatmap quality of that specific scene, users can adjust the magnitude/timestep of the noise added to the Grad-CAM heatmap to decide the extent of information to be kept. In our method, we added noise at the 300th step instead of 250 for *I Wanna Marry Harry* and *Survivor* due to the lower quality of the Grad-CAM heatmaps for those videos. Overall, the results show that our method can also be applied to new videos, of which only a small number of frames are annotated.

S5 Diffusion Performance regarding AUC

In this section, we visually illustrate the cause of the lower performance in the AUC metrics for the diffusion baselines in Tab. 4 in the main paper. In the evaluation of the AUC metric on the GazeFollow test set, for each image, the heatmap is evaluated with a binary map where each annotation from the 10 annotators is assigned as one. Therefore, the AUC favors heatmap predictions covering a larger area so that they overlap with more annotations, even if the heatmap predictions have high responses on incorrect locations, as shown in Fig.S2. As the diffusion model learns to model the distribution of human-labeled annotation of the training data, which is a single Gaussian, the outputs of Diffusion-GC tend to overlap with fewer human annotations, even when the predicted target

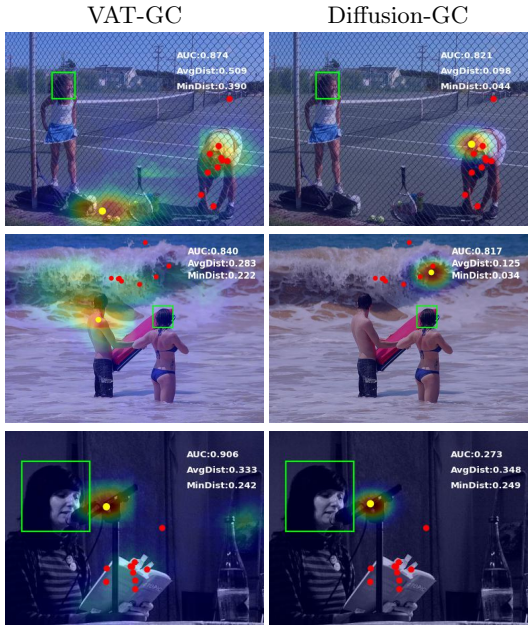

Fig. S2: Example demonstrations of the lower AUC for diffusion model outputs on GazeFollow test images. For each image, the predicted gaze target is shown as a yellow dot and the ground truth annotations are shown as red dots. The lower AUC for the diffusion model is mainly due to the smaller overlap between the output heatmaps and the annotations. The AUC is negatively affected when the annotations span a larger area, despite the predicted targets being closer to the actual gaze targets (Rows 1-2). In Row 3, although both models predict the same incorrect target, VAT-GC attains a much higher AUC due to the overlap of the group-level annotations with a weak response from the model.

locations are more accurate (Rows 1-2). In Row 3, VAT-GC and Diffusion-GC show very similar predictions with the predicted targets on incorrect locations, while VAT-GC shows much higher AUC due to the overlap of a weak response with a cluster of annotations.

These examples also serve as a clue for the smaller performance advantage in AUC compared to the distance metrics in the semi-supervised results in the main paper. As mentioned in the paper, our methods perform better in predicting the probable location of the target than in predicting the exact shape of the heatmap.

S6 Training with 50% of Annotations

In Tab. S2, we provide the semi-supervised experiment results when training with 50% annotations. Our method still outperforms the baselines. We achieved close performance to the VAT model trained with full annotations. The performance

Table S2: Semi-supervised experiments when training with 50% labels

Method	Dist. ↓		AUC ↑
	Avg.	Min.	
VAT (Supervised)	0.160	0.096	0.912
Semi-VAT	0.153	0.090	0.916
Semi-VAT-GC	0.149	0.086	0.917
VAT-MT	0.150	0.086	0.918
GCDR (Ours)	0.143	0.082	0.919
GCDR-MT (Ours)	**0.141**	**0.080**	**0.920**
VAT (100% labels)	0.137	0.077	0.921

did not outperform the VAT trained with 100% annotations because only 50% unlabeled data can be used in this scenario, which is the same amount as labeled data, while semi-supervised learning usually assumes that the unsupervised data has a much larger scale than labeled data (As shown by results of training with 5%, 10% and 20% in the paper). Based on the analysis results of the amount of unlabeled data in Tab. S6, we believe the performance of our method can outperform the model trained with 100% annotations when additional large unsupervised data is available.

S7 Training with 10% labels on VideoAttentionTarget

In the main paper, we showed the results of fine-tuning the VAT model on each of the 10 videos in VideoAttentionTarget in a semi-supervised manner. Here we adopt an ordinary train-test evaluation scheme and evaluate the performance of the VAT model trained with 10% of annotations in the training set and evaluated on the test set of VideoAttentionTarget, with the pseudo annotations of the unlabeled data generated by different methods. Tab. S3 shows that our method (GCDR) outperforms the pseudo annotation generation baselines in both metrics. Our method even slightly outperforms the VAT model trained with full labels on VideoAttentionTarget, possibly due to the stochastic nature of the diffusion model.

Table S3: Results of Training with 10% labels on VideoAttentionTarget

Method	Dist. ↓	AUC ↑
VAT (Supervised)	0.144	0.825
Semi-VAT	0.136	0.844
Semi-VAT-GC	0.141	0.856
GCDR	**0.127**	**0.865**
VAT (100% labels)	0.134	0.860

S8 Experiments with Other Gaze Following Models

In the main paper, we based our experiments on VAT. Here we show that our method is also applicable to other gaze following models. In this section, we applied our semi-supervised learning method to Miao *et al*. [24] and Lian *et al*. [19]. Note that [24] uses depth as additional input and has another output branch for patch distribution prediction (PDP) of the gaze target. When applying our method to [24], we modified the feature extraction module of the diffusion model to be the same as [24] to use depth as input. In semi-supervised training, we computed the patch-level gaze distribution following [24] from the generated pseudo heatmaps as the pseudo label for the PDP branch.

We tested the pseudo annotation generation methods with 10% of annotations. Similar to Tab.1 in the paper, we built *Semi-Miao* and *Semi-Lian* by training [24] and [19] on the labeled data to generate pseudo labels. We also built *Semi-Miao-GC* and *Semi-Lian-GC* by modifying the models to use the Grad-CAM heatmap as conditional input. *Miao-MT* was implemented by applying the consistency loss on the predicted heatmaps and patch-level distributions in [24] from the teacher and student models, while *Lian-MT* applies consistency loss between the predicted gaze directions and heatmaps from the teacher and student models. We still have two versions of our method: *GCDR* and *GCDR-MT*, except that when applied to [24], the feature extraction module of the diffusion model was the same as [24].

Table S4: Using [24] as base model

Method	Dist. ↓		AUC ↑
	Avg.	Min.	
Miao (10% labels)	0.190	0.123	0.877
Semi-Miao	0.180	0.115	0.892
Semi-Miao-GC	0.185	0.117	0.893
Miao-MT	0.217	0.146	0.854
GCDR (Ours)	0.167	0.104	0.897
GCDR-MT (Ours)	**0.163**	**0.100**	**0.900**
Miao (20% labels)	0.166	0.103	0.900

Table S5: Using [19] as base model

Method	Dist. ↓		AUC ↑
	Avg.	Min.	
Lian (10% labels)	0.216	0.145	0.875
Semi-Lian	0.203	0.133	0.886
Semi-Lian-GC	0.198	0.129	0.889
Lian-MT	0.196	0.128	0.882
GCDR (Ours)	0.178	0.112	0.895
GCDR-MT (Ours)	**0.171**	**0.106**	**0.898**
Lian (20% labels)	0.186	0.120	0.892

Tab. S4 and Tab. S5 present the results. Our method shows consistent improvements compared to the baseline methods, especially in the distance metrics. Our semi-supervised methods show comparable or better performance to the supervised model trained with 20% annotations, the same as the results in the main paper. The Mean Teacher method failed when directly applied to [24], possibly due to the instability when applying the consistency loss to outputs from two branches in [24]. This substantiates the effectiveness of our method when applied to other gaze following models, including models using additional modalities, by modifying the feature extraction module of the diffusion model to be the same as the base model for using the additional modalities.

S9 Effect of the Amount of Unlabeled data

In the main paper, we ran experiments with different amounts of annotations in GazeFollow and treated the rest of the data as unlabeled. Though this experiment setting was designed following previous semi-supervised learning works [4, 19, 26], the amount of both labeled and unlabeled data was different across settings. In this section, we specifically tested the effect of the amount of unlabeled data by fixing the number of labeled annotations as 10%, and varying the amount of unlabeled data.

Results are shown in Tab. S6. Along with the results of training with 10% labeled and 90% unlabeled data copied from Tab.1, we experimented using 10% and 50% unlabeled data in training. Our method shows consistent improvement over the baselines, and more unlabeled data leads to better performance.

Table S6: Training on 10% labeled data and varying amounts of unlabeled data.

Method	10% unsupervised			50% unsupervised			90% unsupervised		
	Dist. ↓		AUC ↑	Dist ↓		AUC ↑	Dist. ↓		AUC ↑
	Avg.	Min.		Avg.	Min.		Avg.	Min.	
VAT (Supervised)	0.202	0.133	0.869	0.202	0.133	0.869	0.202	0.133	0.869
Semi-VAT	0.199	0.130	0.872	0.196	0.129	0.874	0.195	0.128	0.875
Semi-VAT-GC	0.200	0.132	0.872	0.197	0.128	0.877	0.195	0.127	0.879
VAT-MT	0.202	0.133	0.875	0.194	0.125	0.878	0.189	0.122	0.882
GCDR (Ours)	0.185	0.119	0.880	0.181	0.117	0.883	0.179	0.115	0.886
GCDR-MT (Ours)	**0.183**	**0.117**	**0.881**	**0.176**	**0.112**	**0.886**	**0.172**	**0.108**	**0.892**

S10 Visualizations of Diffusion Model Output in each Inference Step

In our ablation study in the main paper, we presented the performance for various initial refinement timesteps, while the outputs are the outputs from the last inference step. In this section, we investigate the model's behavior by visualizing the output heatmaps at each inference step. Recall that we predict $\mathbf{h}_0$ directly at each inference step.

Fig.S3 shows the heatmap visualizations during the Grad-CAM heatmap refinement (Rows 1-2) and sampling from pure Gaussian noise (Rows 3-4). In the first step, the diffusion model usually outputs uncertain heatmaps that span a large area. During the iterative inference process, the output heatmaps become more concentrated, similar to the single Gaussian shape of human-labeled annotations. During refinement, the diffusion model outputs more concentrated heatmaps at earlier steps, due to the smaller magnitude of noise added compared to sampling from pure Gaussian noise. From (a) and (b) we observe inference

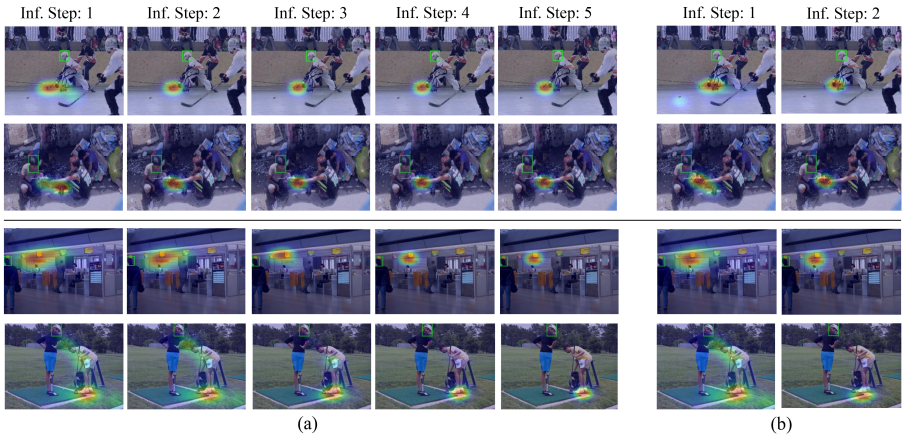

Fig. S3: Visualizations of the diffusion model output in each step. (a): 5 inference steps in total (b): 2 inference steps in total. The top 2 rows show the output during refinement of the Grad-CAM heatmaps, and the bottom 2 rows show the output when sampling from pure Gaussian noise. The diffusion model tends to generate more concentrated heatmaps during the inference process.

with 2 steps shows a similar final output with 5 steps, therefore we choose 2 inference steps to reduce inference time.

S11 Discussion on Existing Semi-Supervised Learning Works

In this section, we provide a more detailed discussion of the existing semi-supervised learning methods and their potential applicability in gaze following. As mentioned in the main paper, most semi-supervised learning methods target recognition and segmentation tasks and usually involve task-specific operations [2,11,36]. Some semi-supervised recognition methods convert the predicted multi-class distribution from the teacher models to a one-hot vector in recognition [17, 31] by selecting the class with the highest predicted confidence, while some work uses a 'sharpening' operation to obtain a softer vector similar to one-hot [2, 10], or requires the predicted class label from the student model to find a stable sample [12]. Meanwhile, some segmentation methods also compute pixel-wise one-hot pseudo-labels in segmentation by selecting the most confident class [35], or compare the similarities between the predicted classification distributions from the teacher and student models in a pixel-wise manner [11]. In addition, some earlier semi-supervised methods use Generative Adversarial Networks (GANs) or Variational AutoEncoders (VAEs) to learn a latent space in which features from different classes are better separated [14,25,28,32]. These classification-specific operations cannot be easily adapted to gaze following, as

Table S7: Results of applying "one-hot" transformation on the baseline predictions. We generated a Gaussian on the maximum response pixel on the predicted heatmaps as pseudo-annotations for the baseline methods (Semi-VAT-Gaussian and Semi-VAT-GC-Gaussian). Other results are copied from Tab.1 for reference. Experiments are performed by training with 10% annotations.

Method	Dist. ↓		AUC ↑
	Avg.	Min.	
VAT (Supervised)	0.202	0.133	0.869
Semi-VAT	0.195	0.128	0.875
Semi-VAT-Gaussian	0.576	0.489	0.777
Semi-VAT-GC	0.195	0.127	0.879
Semi-VAT-GC-Gaussian	0.195	0.127	0.863
GCDR (Ours)	0.179	0.115	0.886
GCDR-MT (Ours)	**0.172**	**0.108**	**0.892**

gaze following predicts a dense spatial heatmap for the target instead of predicting the class of the input. Even if the target heatmap is flattened to a vector by treating each pixel as one category, it is incorrect to convert the prediction to a one-hot vector as it ignores the spatial correlations between neighboring pixels compared to using the spatial heatmap as pseudo annotations.

In Tab. S7, we also tried generating a Gaussian on the maximum response pixel from the predicted heatmaps of the baseline models as pseudo-annotation, which mimics the procedures of generating softer "one-hot" labels. We didn't perform this on VAT-MT, because the Mean Teacher method enforces consistencies between the original output from the teacher and student models, while the operation of generating a Gaussian will change the output. After generating a Gaussian as the pseudo-annotations, Semi-VAT-GC shows the same performance in Dist., but with a lower AUC as the Gaussian heatmap support now spans a smaller area than the original prediction. The Semi-VAT totally fails after this operation, possibly because without the Grad-CAM heatmap prior, the maximum responses usually locate in incorrect locations, which makes the training very unstable. In contrast, the original prediction may possibly overlap with the correct target due to the larger and uncertain heatmap response, even when the predicted target is incorrect.

In addition, semi-supervised learning methods usually involve strong augmentations and enforce consistencies between the student and teacher model outputs from the strongly and weakly augmented views of the input. Some methods used CutOut [6, 31], CutMix [13], shear or translation of images [5, 37], masking out image patches [1], Mixup of images and labels [2, 3] as the strong augmentations for training the semi-supervised model. Unfortunately, none of these augmentations can be applied to gaze following which requires an intact scene image for inferring the target, where the person and target must both be in the image.

On the other hand, semi-supervised learning was also applied in certain regression tasks, such as crowd-counting. However, the related methods still mostly involve task-specific operations, including auxiliary tasks such as segmentation

of crowd on the images [22,23,38], crowd density ranking on image patches [20], training uncertainty prediction module on cropped patches [18], which makes these methods not directly applicable to gaze following.

Despite the plenty of non-applicable methods, some 'general' methods, that do not rely on task-specific operations, can be adapted to gaze following with appropriate modifications. For example, II Model [27] enforces consistency on the outputs from the shared-weights teacher and student models, whose inputs are the same image added with different noises; Temporal Ensembling [16] updates the teacher model output as the exponential moving average (EMA) of outputs from past epochs; Mean Teacher [33] updates the teacher model weights as EMA of student model weights during training. Although mostly tested on recognition tasks, the strategies of these methods can be adapted to gaze following and applied to the predicted heatmaps. Within these methods, the EMA-Teacher framework in Mean Teacher has been dominantly adopted by recent semi-supervised learning works spanning multiple tasks [1, 3, 8, 15, 21], despite usually augmented with task-specific operations for the specified tasks. Therefore, we choose Mean Teacher as the consistency regularization baseline in the experiments, and we also show that we can use Mean Teacher to enhance the diffusion model training and achieve even larger improvements.

S12 Ablations in Grad-CAM Heatmap Computation

In this section, we show the ablations of different alternatives for obtaining the Grad-CAM heatmaps from the VQA model, by evaluating the Grad-CAM heatmaps computed on the GazeFollow test set directly. We varied the decoder layer of the cross-attention map from which the Grad-CAM heatmaps are computed and tested overlaying the scene image with the head bounding box directly instead of the body bounding box obtained with a person detector. It can be seen from Tab.S8 that even raw Grad-CAM heatmaps can achieve adequate performance on the GazeFollow test set. Overlaying the image with the body bounding box showed slightly better quality than overlaying with the head bounding box, while computing Grad-CAM from layer 11 (12 layers in total in the decoder of OFA model [34]) showed the best performance.

Table S8: Quality of Grad-CAM heatmaps generated from different alternatives on GazeFollow test set.

Overlay	Decoder Layer	Dist. ↓		AUC ↑
		Avg.	Min	
Head Box	Layer 11	0.258	0.182	0.784
Body Box	Layer 10	0.262	0.190	0.775
Body Box	Layer 11	**0.254**	**0.180**	**0.791**
Body Box	Layer 12	0.283	0.209	0.785

References

1. Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., Ballas, N.: Masked siamese networks for label-efficient learning. In: European Conference on Computer Vision. pp. 456–473. Springer (2022)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems **32** (2019)
3. Cai, Z., Ravichandran, A., Favaro, P., Wang, M., Modolo, D., Bhotika, R., Tu, Z., Soatto, S.: Semi-supervised vision transformers at scale. Advances in Neural Information Processing Systems **35**, 25697–25710 (2022)
4. Chong, E., Wang, Y., Ruiz, N., Rehg, J.M.: Detecting attended visual targets in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5396–5406 (2020)
5. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
6. French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.: Semi-supervised semantic segmentation needs strong, varied perturbations. In: British Machine Vision Conference. No. 31 (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Ho, C.J., Tai, C.H., Lin, Y.Y., Yang, M.H., Tsai, Y.H.: Diffusion-ss3d: Diffusion model for semi-supervised 3d object detection. Advances in Neural Information Processing Systems **36** (2024)
9. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., et al.: spacy: Industrial-strength natural language processing in python (2020)
10. Hu, Z., Yang, Z., Hu, X., Nevatia, R.: Simple: Similar pseudo label exploitation for semi-supervised classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15099–15108 (2021)
11. Kalluri, T., Varma, G., Chandraker, M., Jawahar, C.: Universal semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5259–5270 (2019)
12. Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W.: Dual student: Breaking the limits of the teacher in semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6728–6736 (2019)
13. Kim, J., Jang, J., Park, H., Jeong, S.: Structured consistency loss for semi-supervised semantic segmentation. arXiv preprint arXiv:2001.04647 (2020)
14. Kingma, D.P., Mohamed, S., Jimenez Rezende, D., Welling, M.: Semi-supervised learning with deep generative models. Advances in neural information processing systems **27** (2014)
15. Kwon, D., Kwak, S.: Semi-supervised semantic segmentation with error localization network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9957–9967 (2022)
16. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations (2017), `https://openreview.net/forum?id=BJ6oOfqge`

17. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896. Atlanta (2013)
18. Li, C., Hu, X., Abousamra, S., Chen, C.: Calibrating uncertainty for semi-supervised crowd counting. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16685–16695. IEEE (2023)
19. Lian, D., Yu, Z., Gao, S.: Believe it or not, we know what you are looking at! In: Asian Conference on Computer Vision. pp. 35–50. Springer (2018)
20. Lin, H., Ma, Z., Hong, X., Wang, Y., Su, Z.: Semi-supervised crowd counting via density agency. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1416–1426 (2022)
21. Lin, W., Chan, A.B.: Optimal transport minimization: Crowd localization on density maps for semi-supervised counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21663–21673 (2023)
22. Liu, Y., Liu, L., Wang, P., Zhang, P., Lei, Y.: Semi-supervised crowd counting via self-training on surrogate tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 242–259. Springer (2020)
23. Meng, Y., Zhang, H., Zhao, Y., Yang, X., Qian, X., Huang, X., Zheng, Y.: Spatial uncertainty-aware semi-supervised crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15549–15559 (2021)
24. Miao, Q., Hoai, M., Samaras, D.: Patch-level gaze distribution prediction for gaze following. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 880–889 (2023)
25. Paige, B., van de Meent, J.W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., Torr, P., et al.: Learning disentangled representations with semi-supervised deep generative models. Advances in neural information processing systems **30** (2017)
26. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper/2015/file/ec8956637a99787bd197eacd77acce5e-Paper.pdf
27. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems **29** (2016)
28. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
29. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
30. Shen, R., Inoue, N., Shinoda, K.: Text-guided object detector for multi-modal video question answering. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1032–1042 (2023)
31. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
32. Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390 (2015)

33. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
34. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning. pp. 23318–23340. PMLR (2022)
35. Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X.: Semi-supervised semantic segmentation using unreliable pseudo-labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4248–4257 (2022)
36. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering (2022)
37. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems **34**, 18408–18419 (2021)
38. Zhu, P., Li, J., Cao, B., Hu, Q.: Multi-task credible pseudo-label learning for semi-supervised crowd counting. IEEE Transactions on Neural Networks and Learning Systems (2023)