Diffusion-Refined VQA Annotations for Semi-Supervised Gaze Following

Qiaomu Miao[1], Alexandros Graikos[1], Jingwei Zhang[1], Sounak Mondal[1], Minh Hoai[2], and Dimitris Samaras[1]

[1] Stony Brook University, Stony Brook, USA
[2] The University of Adelaide, Adelaide, Australia
{qiamiao,agraikos,jingwezhang,somondal,minhhoai,samaras}@cs.stonybrook.edu

Abstract. Training gaze following models requires a large number of images with gaze target coordinates annotated by human annotators, which is a laborious and inherently ambiguous process. We propose the first semi-supervised method for gaze following by introducing two novel priors to the task. We obtain the first prior using a large pretrained Visual Question Answering (VQA) model, where we compute Grad-CAM heatmaps by 'prompting' the VQA model with a gaze following question. These heatmaps can be noisy and not suited for use in training. The need to refine these noisy annotations leads us to incorporate a second prior. We utilize a diffusion model trained on limited human annotations and modify the reverse sampling process to refine the Grad-CAM heatmaps. By tuning the diffusion process we achieve a trade-off between the human annotation prior and the VQA heatmap prior, which retains the useful VQA prior information while exhibiting similar properties to the training data distribution. Our method outperforms simple pseudo-annotation generation baselines on the GazeFollow image dataset. More importantly, our pseudo-annotation strategy, applied to a widely used supervised gaze following model (VAT), reduces the annotation need by 50%. Our method also performs the best on the VideoAttentionTarget dataset. Code is available at https://github.com/cvlab-stonybrook/GCDR-Gaze.git

Keywords: Gaze following · Semi-supervised learning · Diffusion Model

1 Introduction

Human gaze behavior is a vital cue for understanding human cognitive processes [74,75] and for applications such as human-machine interaction [1,13,47], social interaction analysis [5,42], and human intention interpretation [56,60]. In contrast to eye-tracking glasses which are intrusive and mostly restricted to laboratory environments, the gaze following task [9,14,53] predicts a person's gaze target in an image in-the-wild, by predicting a target heatmap from an input scene image with the person's head bounding box.

The effectiveness of gaze following methods depends on the quantity and quality of training data with annotated gaze targets. Annotating gaze targets is

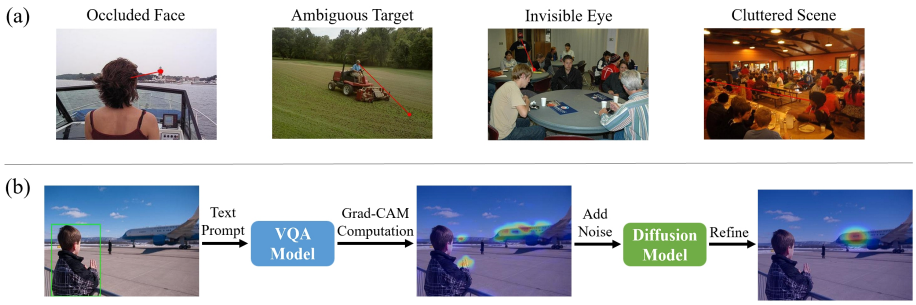

Fig. 1: (a) **Gaze following annotation challenges.** Annotating gaze is a laborious task with inherent ambiguities. (b) **Pseudo annotations for gaze following**. We generate pseudo annotations by first computing Grad-CAM heatmaps from a pre-trained VQA model, and then refining the noisy heatmaps with a diffusion model.

a laborious and ambiguous process. For an image and a subject in the image, an annotator must consider the subject's head orientation, gaze angle, and follow the gaze direction to determine where the point of gaze falls on an object or surface within the image. This can be very challenging, as the subject may be looking away from the camera, their eyes may be occluded, the image may be cluttered, and there may be multiple plausible targets, as shown in Fig. 1(a).

It is much more efficient to utilize large quantities of unlabeled data, along with a limited amount of labeled data, and leverage semi-supervised learning to improve model performance. Semi-supervised learning has been successful in tasks such as image recognition and semantic segmentation [2, 27, 30, 67, 73]. In the context of gaze following, semi-supervised learning can be particularly useful when developing models for a specific scene, such as a psychology lab conducting social interaction studies or a supermarket studying customer gaze behaviors, by requiring only a small fraction of the collected data to be annotated.

This is the first semi-supervised gaze following method, to the best of our knowledge. Semi-supervised recognition and segmentation methods are typically not directly applicable to gaze following due to task differences. As gaze following models predict a target heatmap and require an intact scene image as input, methods that operate on the predicted multi-class distribution [30, 65], perform strong data augmentation (e.g.,CutOut [15, 62]) cannot easily be adapted to gaze following. Additionally, for semi-supervised methods that use a teacher model to pseudo-annotate the unlabeled data [33, 34, 67], it is unclear how to design a gaze following teacher model to generate good quality pseudo-annotations.

To generate high-quality pseudo-gaze labels, our semi-supervised method combines the power of pre-trained large vision-language (VL) models with an annotation refinement method. Due to their extensive training sets, VL models have naturally acquired a wealth of knowledge, including the knowledge for inferring human gaze targets. Specifically, we "prompt" a pre-trained Visual Question Answering (VQA) model [44, 52, 78] with an appropriate question to ob-

tain the text description of the gaze target of a subject in the image. We then use Grad-CAM [6, 57], to ground the VQA model's description to spatial locations, referred to as Grad-CAM heatmaps. We want to turn these heatmaps into pseudo-annotations for training gaze following models.

There are technical challenges in utilizing Grad-CAM heatmaps. First, Grad-CAM heatmaps do not precisely overlap with the gaze targets because they are dispersed and noisy, as shown in Fig. 1(b). Second, some Grad-CAM heatmaps do not provide useful prior knowledge and may highlight incorrect gaze target locations. Hence, the noisy Grad-CAM heatmaps should be refined into clean pseudo-annotations, with a trade-off between retaining the Grad-CAM heatmap information and the pre-trained prior on the annotations available.

We choose a diffusion model as the annotation prior of our refinement method. We are inspired by the recent use of diffusion models as inverse problem solvers and image editors that modify the reverse sampling process with a conditional input [43,48,64]. The goal is to generate an output image that retains the semantics of the degraded input (here the Grad-CAM heatmap), while having similar properties (i.e., geometry and gaze context) to the training data distribution.

We begin by training a diffusion model to capture the distribution of human-labeled annotations. The trained diffusion model then generates refined annotations by running the reverse sampling process, initialized from Grad-CAM heatmaps with an appropriate level of Gaussian noise. The added noise serves to smooth artifacts and noisy activations in the Grad-CAM heatmaps; the magnitude of the noise dictates how much of the information is preserved.

In summary, we propose the first semi-supervised gaze following method, in which we introduce two novel priors to the task: (1) Prior knowledge from pre-trained VQA models for initial annotation generation. (2) A novel diffusion-based annotation prior to refine the noisy VL annotations into reliable pseudo-labels.

Our method outperforms pseudo-annotation generation baselines on the GazeFollow image dataset [53]. More importantly, our pseudo-annotation strategy, applied to the widely used VAT model [9], outperforms the fully supervised model trained with double the amount of annotations when only 5% and 10% labels are available. Our method also performs the best on the VideoAttentionTarget dataset [9], where we adapt a pre-trained gaze following model to new videos using only around 100 (2–23%) annotated frames.

2 Prior Work

Gaze Following was first introduced in [53] together with the GazeFollow dataset and a model composed of two separate pathways for encoding gaze orientation and scene saliency information, also adopted by later work [8, 9, 14, 38]. Chong *et al.* [8] considered out-of-frame targets and also extended the task to videos [9], proposing the VideoAttentionTarget dataset and the VAT model. Later work has improved gaze following models by leveraging monocular depth estimations [14,25,29,68] and human poses [4,20]. Tu *et al.* [71] proposed a transformer model for gaze following, while [28, 45] added training losses for numer-

ical coordinate regression and patch-level gaze distribution prediction. Recent works investigate 3D gaze following [24, 26], gaze following for children [66], and object-aware gaze target prediction [69]. All these methods are fully-supervised. A concurrent work investigated semi-supervised gaze following using a saliency prediction model [50]. This approach does not leverage the strong priors from pretrained VL models, and its performance falls significantly behind our method.

Semi-supervised Learning methods mostly adopt a teacher-student framework, which can be categorized into self-training [7, 19, 34, 62] and consistency regularization [15, 22, 33, 67]. Self-training methods generate pseudo annotations from teacher models trained with labeled data, while consistency regularization applies a consistency loss between the student and teacher model outputs, updating the teacher model gradually during training. Due to task differences, most semi-supervised recognition or segmentation methods are not directly applicable to gaze following [15, 30, 37, 49, 62]. However, some general methods without task-specific operations [33, 67] are applicable with appropriate modifications (see Supplementary). In our method, we adopt a self-training pipeline using the VQA prior and diffusion-based annotation prior, and show that we can get further improvements by using applicable consistency regularization methods (e.g. Mean Teacher [67]) to enhance the diffusion model training.

Vision-Language Models have prospered after the development of large language models (LLMs) [10,11,40,41,70], and have succeeded in tasks such as visual grounding, visual question answering (VQA), and image captioning [31,35,39,59, 72]. They also do well in low-shot generalization tasks by "prompting" [44,46,51], where task instructions are given to a pretrained model to generate outputs useful for other tasks. Grad-CAM [58] visualizations *localize* the linguistic input effect on the visual input [36, 77] thus achieving a degree of interpretability.

Diffusion Models. Denoising diffusion probabilistic models (DDPM) [23, 61] give state-of-the-art results on image synthesis [12, 17, 18, 32, 54]. They are also used in inverse problems [64], image editing [43], and adversarial purification [48]. These works treat images with added noise as intermediate steps of the reverse process, showing that the noise magnitude affects the information retained from the input. Recent work adopts diffusion models in semi-supervised learning. [22] trains a diffusion model for 3D object detection, whereas [76] trains a diffusion model for image generation with images pseudo-labeled by existing semi-supervised models. In contrast, we use diffusion model as annotation prior, to refine the initial noisy annotations into high-quality pseudo annotations.

3 Proposed Method

3.1 Overview

Our approach is illustrated in Fig. 2a. The goal is to leverage the knowledge priors from a VQA model to improve gaze following with minimal manual annotation. To achieve this, we utilize Grad-CAM heatmaps of a VQA model that is prompted with gaze-related questions.

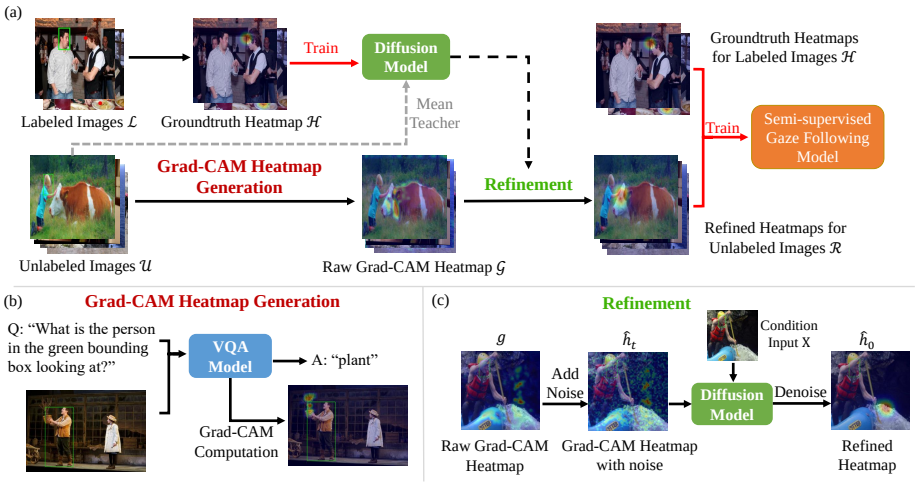

Fig. 2: (a) Overall pipeline. We compute Grad-CAM heatmaps for unlabeled images and train the diffusion model with a small human-labeled set (or with unlabeled images using Mean Teacher). The diffusion model refines the Grad-CAM heatmaps into pseudo-annotations. Both the pseudo-annotations and the human-labeled set are used to train a gaze following model. **(b) Grad-CAM heatmap generation.** Given an image with an overlaid person bounding box, we "prompt" a pretrained VQA model with a gaze question and compute the Grad-CAM heatmap from the answer. **(c) Grad-CAM refinement.** We perturb the Grad-CAM heatmaps with Gaussian noise and pass through the reverse diffusion process to generate the refined pseudo-annotations.

However, the Grad-CAM heatmaps need refinement to be useful for training supervision. For this, we train a diffusion model on the groundtruth heatmaps from labeled data to learn to generate heatmaps from the human-label distribution. We use the model to refine the "noisy" Grad-CAM into high-quality pseudo annotations. We train existing gaze following models on both groundtruth heatmaps from labeled data and refined heatmaps from unlabeled data.

Suppose we only have a small number of images that are labeled with gaze target locations $\mathcal{L} = \{(\boldsymbol{X}_i, \boldsymbol{p}_i)\}_{i=1}^{N_l}$ and a larger number of unlabeled images $\mathcal{U} = \{\boldsymbol{X}_j\}_{j=1}^{N_u}$. $\boldsymbol{X}_i$ indicates an input triplet to the gaze following models, which includes a scene image $\boldsymbol{I}_i \in R^{3\times H\times W}$, the cropped head image $\boldsymbol{I}_h^i \in R^{3\times H\times W}$, and the head location binary mask $\boldsymbol{M}_h^i \in R^{H\times W}$. We represent the annotated gaze coordinate of the "gazing" person as $\boldsymbol{p}_i = [p_i^x, p_i^y]$. For each $\boldsymbol{p}_i$, a ground truth heatmap $\mathbf{h}^i \in R^{H'\times W'}$ is generated by placing a Gaussian of fixed variance centered at $\boldsymbol{p}_i$, as is done in previous gaze following frameworks [9, 14, 38].

3.2 Grad-CAM heatmap extraction

We use OFA [72], a transformer-based large pretrained VL model as the VQA model. Fig. 2b illustrates the procedure for Grad-CAM heatmap generation.

Given an image $\boldsymbol{I}$ and the head bounding box of a person $\boldsymbol{l}_h$, we use Mask R-CNN [21] for person detection and find the bounding box that maximally overlaps with $\boldsymbol{l}_h$. We overlay this bounding box on top of the input image and give it as input to the VQA model along with a "prompt" in the form of a question: "*What is the person in the green bounding box looking at?*".

OFA has an encoder and a decoder. It generates a sequence of words as the answer. We compute Grad-CAM heatmaps $\mathcal{G} = \{\boldsymbol{g}_j\}_{j=1}^{N_u}$ on the decoder cross-attention weights between the input query and the image patch tokens based on [6] after selecting the noun from the answer. Details are in the supplementary.

In our experiments, we found that Grad-CAM heatmaps are beneficial for gaze following. Naively using these heatmaps as additional input to the current gaze following model results in a significant performance boost. However, in real applications, computing Grad-CAM heatmaps at test-time can be impractical due to the memory and time costs of accessing large VL models. Instead we use Grad-CAM heatmaps offline as pseudo-annotations for training a gaze following model. As discussed, using noisy Grad-CAMs directly as pseudo labels can hurt final model performance. Thus, we propose to utilize a diffusion model to refine the initial Grad-CAM heatmaps into more suitable pseudo labels for training.

3.3 Diffusion Model Training

As there is no available diffusion model for the gaze following task, we first train a diffusion model on the labeled data to generate gaze heatmaps $\mathbf{h} \sim p(\mathbf{h}|\mathbf{X})$, following the general training procedure of DDPM [23]. For ease of understanding, we provide a synopsis of Gaussian diffusion models.

The diffusion model consists of a forward process that gradually corrupts the input by adding Gaussian noise, and a reverse process that iteratively denoises the noisy input. The forward process is defined by a noise schedule α_t as:

$$q(\mathbf{h}_t|\mathbf{h}_{t-1}) := \mathcal{N}(\mathbf{h}_t; \sqrt{\alpha_t}\mathbf{h}_{t-1}, (1-\alpha_t)\mathbf{I}) \tag{1}$$

$$q(\mathbf{h}_{1:T}|\mathbf{h}_0) := \prod_{t=1}^{T} q(\mathbf{h}_{t-1}|\mathbf{h}_t). \tag{2}$$

where $\mathbf{h}_0$ is the ground truth heatmap $\mathbf{h}$, $\mathbf{h}_1, \ldots, \mathbf{h}_{T-1}$ are intermediate latent variables that represent noisy versions of $\mathbf{h}_0$, and $\mathbf{h}_T$ is the terminal state which corresponds to a unit Gaussian distribution: $\mathbf{h}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The reverse process uses Gaussian transitions with fixed covariance σ_t and gradually denoises the data, starting from $\mathbf{h}_T$. Since $q(\mathbf{h}_{t-1}|\mathbf{h}_t)$ is intractable, it is approximated with a neural network:

$$p_\theta(\mathbf{h}_{t-1} \mid \mathbf{h}_t) := \mathcal{N}(\mathbf{h}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{h}_t, t), \sigma_t\mathbf{I}). \tag{3}$$

A useful property of the forward process is that it allows sampling of any intermediate $\mathbf{h}_t$ given $\mathbf{h}_0$, and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$:

$$q(\mathbf{h}_t|\mathbf{h}_0) = \mathcal{N}(\mathbf{h}_t; \sqrt{\bar{\alpha}_t}\mathbf{h}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{4}$$

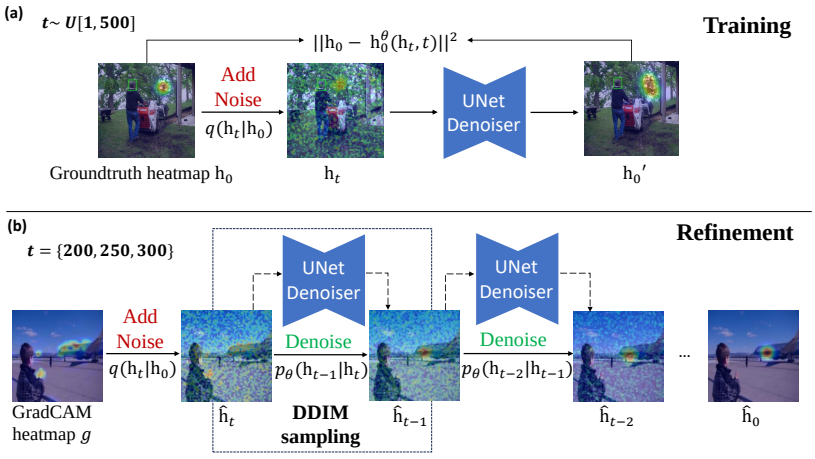

Fig. 3: Diffusion model training and refinement. (a) The diffusion model is trained on supervised data with noise added to the ground truth heatmap at random time steps. (b) During refinement, we add noise at a specific time step to the Grad-CAM heatmap. We treat this heatmap as an intermediate step input during the reverse process. Heatmaps are overlayed on the original images for illustration purposes. The conditional feature extraction for the diffusion model is omitted for simplicity.

which is utilized in learning the reverse process. For that, we adopt the $\mathbf{x}_0$-parameterization as in [3, 32] and directly predict the final $\mathbf{h}_0$, leading to the following minimization objective, where $\mathbf{h}_0^\theta$ is parameterized by a U-Net [55] :

$$\mathbb{E}_{\mathbf{h}_0,t}\left[\|\mathbf{h}_0 - \mathbf{h}_0^\theta(\mathbf{h}_t, t)\|^2\right], \tag{5}$$

Fig. 3(a) shows the training procedure. At each iteration, we sample a random time step t from $U[1, T]$ to generate a noisy input $\mathbf{h}_t$ using Eq. (4). We feed $\mathbf{h}_t$ to the diffusion model for single-step denoising, and optimize with the loss of Eq. (5). We condition the diffusion model with the gaze feature $\mathbf{c}$ extracted from the input triplet $\boldsymbol{X}$ (input image, head crop, and head location mask) to output a gaze heatmap grounded on the gazing person; the parametrization of the U-Net denoiser becomes $\mathbf{h}_0^\theta(\mathbf{h}_t, t, \mathbf{c})$. The feature extractor follows the structure of the VAT model [9], which takes the input triplet $\boldsymbol{X}$ and outputs the extracted features concatenated from two pathways for encoding the scene and gaze features. Details of the VAT structure are shown in Supplementary. Following [16], we add the extracted features as conditional features to each layer in the U-Net denoiser with the necessary transposed convolution layers to match the feature size of U-Net. The diffusion model can also be further improved by training with unlabeled data using Mean Teacher (MT) [67], a general-purpose semi-supervised learning method. As a consistency regularization method, MT enforces consistencies between the outputs from the teacher and student models fed with differently perturbed input. The teacher model has the same structure as the student model and is updated with an Exponential Moving Average (EMA)

of the student model weights during training. In our case, we treat the diffusion model as the student model and apply different color jittering and head bounding box jittering to introduce different perturbations to the teacher and student model input. During training, the diffusion model is trained with the loss on labeled data and an additional consistency loss between the output heatmaps from the teacher and student diffusion models on all data. We demonstrate in our experiments that this further improves the gaze following results.

3.4 Heatmap Refinement using Diffusion Models

After training on labeled images $\mathcal{L}$, the learned diffusion process models an annotation prior over ground truth annotations $p(\mathbf{h}|\mathbf{X})$. During the semi-supervised training process, the trained diffusion model is applied to the unsupervised data $\mathcal{U}$ to produce the refined pseudo annotations.

The refinement process is shown in Fig. 3(b). Given a noisy Grad-CAM heatmap $\boldsymbol{g}$ of an unsupervised input sample $\boldsymbol{X}$, we first add an appropriate amount of noise using the forward diffusion process $\hat{\mathbf{h}}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{g} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ according to Eq. (4). The added noise smoothes out the noise artifacts of the Grad-CAM heatmaps, while preserving the information of the highlighted regions. Then, by iteratively applying the denoising steps in Eq. (3), we get samples of the final heatmap $\hat{\mathbf{h}}_0$ that are geometrically and contextually similar to the human-labeled heatmap $\mathbf{h}_0$ while retaining gaze target information from the noisy Grad-CAM input. During the denoising process, we used DDIM [63] to sample the next step input using the predicted $\mathbf{h}_0$ from the U-Net.

The amount of information retained depends on the choice of t (*i.e.*, the magnitude of added noise), as proven in [43, 48]. A larger magnitude of noise corrupts the Grad-CAM heatmaps more, decreasing the similarity between the diffusion output and the Grad-CAM priors. Similar to [43] which achieves a good trade-off between *realism* and *faithfulness* in image editing with an intermediate level of noise, we also found that initializing from an intermediate timestep in the diffusion process achieved the best trade-off between the Grad-CAM information and the model's learned distribution $p(\mathbf{h}|\mathbf{X})$. With this formulation, the diffusion model can incorporate the high-quality Grad-CAM heatmaps, while ignoring the Grad-CAM priors when they are highly noisy or highlight unlikely gaze target locations. We provide analyses of the added noise magnitude in Sec. 4.3.

3.5 Semi-supervised Training

In semi-supervised training, we use the refined heatmaps $\mathcal{R}$ as pseudo labels for the unsupervised data $\mathcal{U}$. We train the gaze following model with the set of ground-truth heatmaps $\mathcal{H}$ from the labeled data $\mathcal{L}$ and the refined heatmaps $\mathcal{R}$ from the unlabeled data $\mathcal{U}$. We use the Mean Squared Error (MSE) loss between the predicted heatmap $\boldsymbol{h'}$ and the target label $\mathbf{h}$. $\mathbf{h}$ is either a human annotation (labeled data) or a pseudo annotation (unlabeled data). We minimize:

$$\mathcal{L} = \frac{1}{|\mathcal{H} \cup \mathcal{R}|} \sum_{\mathbf{h} \in \mathcal{H} \cup \mathcal{R}} \mathcal{L}_{mse}(\boldsymbol{h'}, \boldsymbol{h}). \tag{6}$$

4 Experiments

4.1 Setting and Implementation Details

Datasets. GazeFollow [53], the largest real image dataset, contains the annotated gaze targets of 130,339 people in 122,143 images. 4782 people are used for testing, and the rest are used in training. For the test set, there are 10 annotations per person-image pair to account for the ambiguity of the gaze target, whereas the training set contains a single annotation for each pair. **VideoAttentionTarget** [9] is a video dataset of 50 different shows collected from YouTube. Each person-image pair in both training and test set has only one annotation.

Evaluation metrics. We employ evaluation metrics suggested by previous work [9,38,53]. The distance metric (**Dist.**) refers to the normalized L_2 distance between the predicted gaze point (point with the highest heatmap response) and the ground truth location. In GazeFollow reports both average (**AvgDist**) and minimum distances (**MinDist**). Area Under Curve (**AUC**) evaluates the concordance of predicted heatmaps with ground truth [9]. For GazeFollow, ground truth is the 10 annotations for each test set image, whereas in VideoAttentionTarget it is a thresholded Gaussian centered at the single given annotation.

Implementation Details. We used a diffusion model with a linear noise schedule and T=500 steps. The heatmap size was 64×64. We used a batch size of 80 and a learning rate of 2.5×10^{-4} in the semi-supervised learning experiments. For training the diffusion model on labeled data, we used a learning rate of 5×10^{-5}. When refining Grad-CAM heatmaps, we used DDIM [63] for inference.

4.2 Semi-supervised Training Results

We build our gaze following experiments around the popular, publicly available VAT model [9]. To showcase the impact of our pseudo annotation generation method, we did not use additional modalities, processing, or supervisions, such as depth input, pose estimations, and object-level annotations [4, 14, 20, 69] in all experiments for the baseline methods and the diffusion model.

Following the standard experimental settings in semi-supervised learning [15, 34, 67], we considered different amounts of annotations from the GazeFollow training set, namely 5%, 10%, and 20%, and treated the remaining data as unlabeled. The trained student model is evaluated on the test set in all cases.

We used VAT as the student model and compared our method of generating pseudo annotations with several alternatives[3]: 1) *Semi-VAT*: the pseudo-annotations for the unlabeled data were generated by the VAT model trained with labeled images. 2) *Semi-VAT-GC*: We created VAT-GC by adding Grad-CAM heatmaps as conditional inputs to the VAT model and modifying the VAT model accordingly. VAT-GC was trained on the labeled data to generate pseudo annotations for the unlabeled data. 3) *VAT-MT*: We use the Mean Teacher (MT)

[3] Implementation details of the baselines are provided in the supplementary material.

Table 1: Results of semi-supervised training methods with different ratios of labeled data on GazeFollow dataset. In the top and bottom rows, we show the performances of VAT trained with supervised data only or with full training data, to show the potential lower and upper limits. Best numbers are marked as bold. Our methods outperform supervised VAT trained with double the amount of annotations.

Method	5% labels			10% labels			20% labels		
	Dist. ↓		AUC ↑	Dist ↓		AUC ↑	Dist. ↓		AUC ↑
	Avg.	Min.		Avg.	Min.		Avg.	Min.	
VAT (Supervised)	0.230	0.161	0.835	0.202	0.133	0.869	0.182	0.116	0.892
Semi-VAT	0.222	0.152	0.846	0.195	0.128	0.875	0.176	0.110	0.896
Semi-VAT-GC	0.217	0.149	0.846	0.195	0.127	0.879	0.178	0.112	0.895
VAT-MT	0.219	0.150	0.850	0.189	0.122	0.882	0.174	0.108	0.898
GCDR (Ours)	0.201	0.135	0.863	0.179	0.115	0.886	0.166	0.103	0.902
GCDR-MT (Ours)	**0.194**	**0.128**	**0.870**	**0.172**	**0.108**	**0.892**	**0.162**	**0.098**	**0.904**
VAT(100% labels) [9]				0.137	0.077	0.921			

method [67] to train the VAT model. The first two baselines are self-training methods, while the 3rd belongs to the consistency regularization category.

On the other hand, we also build two versions of our method: 1) *GradCAM-Diffusion-Refinement (GCDR)*: our method of using a diffusion model trained with labeled data to refine the 'noisy' Grad-CAM heatmaps and use them as pseudo-annotations. 2) *GCDR-MT*: we used the Mean-Teacher method to train the diffusion model with both labeled and unlabeled data, and this enhanced diffusion model was used to refine the Grad-CAM heatmaps.

Semi-supervised training results are in Tab. 1. Our two proposed methods consistently outperform the three semi-supervised baselines in all three annotation scenarios. When training with 5% and 10% annotations, the performance of *GCDR-MT* surpasses the supervised VAT trained with double the amount of annotations. Results of training with 50% are in the supplementary.

Our methods show more prominent improvements in Dist. compared to AUC. The AUC on GazeFollow evaluates the concordance of the predicted heatmap with the 10 annotations on each test image, while Dist. evaluates the L_2 distance between the predicted gaze point and the annotations. Therefore, the results mean that we did better in predicting the probable target location than predicting the exact shape of the heatmap. The refined heatmaps from the diffusion model closely resemble the spatial distribution of a human-labeled annotation, which is a single Gaussian (Fig. 4). In contrast, the baseline methods predict the target location less accurately, while outputting less certain and larger heatmaps which tend to overlap more with the group-level annotations, thus favoring the AUC (see supplementary material).

Fig. 4 visualizes the pseudo annotations predicted by the teacher models. In the top three rows, our model retains positional priors from the Grad-CAM heatmaps with almost identical structure to ground truth heatmaps. On the other hand, when the raw Grad-CAM heatmaps are inaccurate or very noisy (Row 4), our method can ignore the Grad-CAM heatmaps.

Fig. 4: Visualizations of pseudo heatmaps generated by different teachers. Our method generates the cleanest pseudo annotations while retaining the Grad-CAM heatmaps priors (Rows 1–3). When the initial Grad-CAM heatmap responds strongly to unlikely locations or is completely noisy, our method can also ignore it (Row 4).

4.3 Ablation Studies

In the ablation studies shown in Tab. 2, we tested the contribution of the various components of our method. We also tested different parameters of the diffusion sampling process in Tab. 3. All ablation studies were performed using 10% of the GazeFollow labels and the teacher models were trained only on supervised data (without MT) to simplify training. Additionally, in Tab. 4, we analyze the effect of the VQA priors by training the baselines in fully-supervised settings with/without Grad-CAM heatmaps.

Refinement methods In *No Grad-CAM*, the diffusion model samples pseudo annotations from Gaussian noise without using Grad-CAM heatmaps. Without the VQA model priors, there is a significant performance decrease. *No Refinement* directly uses the raw Grad-CAM heatmaps as pseudo annotations without refinement, which also leads to a significant performance decrease due to the noisy nature of the Grad-CAM heatmaps. *Argmax Refinement* generates a Gaussian heatmap from the maximum point of the Grad-CAM heatmap as pseudo annotation. Despite the improvement over *No refinement* due to the "cleaner" annotation pattern, performance is still far behind our method, because some of the Grad-CAM heatmaps are low-quality (Fig. 4), so Gaussian ends up on outliers or incorrect locations. On the contrary, our method can ignore the Grad-CAM heatmaps in these cases. In *Direct Mapping*, we trained a U-Net to learn

Table 2: Ablations of ways for generating the pseudo annotations.

Method	Dist. ↓ Avg.	Dist. ↓ Min.	AUC ↑
No Grad-CAM	0.191	0.125	0.867
No Refine	0.237	0.166	0.833
Argmax Refine	0.207	0.139	0.869
Direct Mapping	0.190	0.122	0.878
Proposed GCDR	**0.179**	**0.115**	**0.886**

Table 3: Ablations of parameters of the diffusion model refining process.

Noise.t	Inf. Steps	Dist. ↓ Avg.	Dist. ↓ Min.	AUC ↑
300	2	0.186	0.120	0.877
200	2	0.184	0.117	0.883
250	5	0.180	**0.115**	0.883
250	2	**0.179**	**0.115**	**0.886**

a "direct mapping" from Grad-CAM heatmaps to groundtruth heatmaps which we then applied to the unlabeled data. This shows a large performance decrease, because it cannot capture the full distribution that a diffusion model learns.

All the above ablations validate the importance of both the Grad-CAM heatmaps priors and the priors that the diffusion model learns.

Added Noise Magnitude Effects In this section, we analyze the effect of the added noise magnitude on the diffusion output. As shown in Tab. 3, adding noise at the 250th step achieved the best trade-off between the VQA prior and the pretrained annotation prior. As a reminder, the later the timestep we add the noise, the larger its magnitude (the more the model ignores the Grad-CAM).

We further visualized the outputs from the same diffusion model when adding noise at different time steps in Fig. 5. In the top 2 rows, the Grad-CAM heatmaps highlight correct locations, while in the bottom 2 rows, they are noisy or highlight incorrect locations. In all these cases, when adding noise at the 100th step, the outputs are located on the highlighted region of the Grad-CAM heatmap. When adding noise at the 400th step, the large magnitude of added noise leads to outputs similar to sampling directly from Gaussian noise without using the Grad-CAM heatmaps. Adding noise at step 250 appears to be the best trade-off.

Table 4: Results of using Grad-CAM heatmaps as a direct input. When Grad-CAM heatmaps are directly input into the gaze inference process, both the VAT and the diffusion model show significant improvement in performance.

Method	Auxiliary Input	5% labels			10% labels			20% labels		
		Dist. ↓ Avg.	Dist. ↓ Min.	AUC ↑	Dist ↓ Avg.	Dist ↓ Min.	AUC ↑	Dist. ↓ Avg.	Dist. ↓ Min.	AUC ↑
VAT	None	0.230	0.161	0.835	0.202	0.133	0.869	0.182	0.116	0.892
VAT-GC	Grad-CAM	0.210	0.144	**0.848**	0.188	0.121	**0.884**	0.173	0.107	**0.898**
Diffusion	None	0.230	0.160	0.768	0.203	0.135	0.847	0.189	0.124	0.849
Diffusion-GC	Grad-CAM	**0.199**	**0.131**	0.803	**0.177**	**0.112**	0.846	**0.167**	**0.103**	0.870

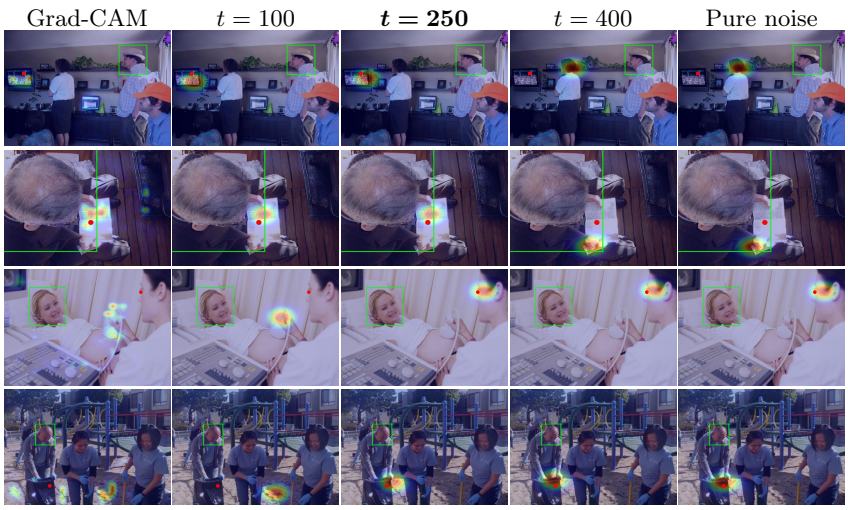

Fig. 5: Diffusion model output for noise added at different timesteps. Red dots represent the ground truth annotation. Adding noise at earlier steps generates outputs on high Grad-CAM response regions. Adding noise at later steps generates outputs similar to sampling from pure noise. Noise at step 250 is the best trade-off.

The Effect of the Grad-CAM Heatmaps In this section, we test how much guidance signal is provided by the VQA priors. This cannot be directly tested in the semi-supervised scenario, so we tested by using the Grad-CAM heatmaps as a direct input to a fully supervised gaze following model. We trained the teacher models of Sec. 4.2 using different amounts of labeled data only, and introduce a new baseline: *Diffusion-GC*, where the diffusion model trained on labeled images was used to refine the Grad-CAM heatmaps computed directly on the test set. Note that *VAT-GC* and *Diffusion-GC* require the Grad-CAM heatmaps of the test images during inference, which may not be feasible in an online setting as it necessitates access to a large vision-language model.

Tab. 4 presents the results. *VAT-GC* outperforms VAT in all cases, and *Diffusion-GC* performs best in the distance metrics. Both the VAT and diffusion model show large performance boosts when Grad-CAM heatmaps are directly used as input, which offers evidence that the VQA Grad-CAM heatmaps bring strong prior knowledge. As illustrated in Sec. 4.2, the lower AUC for *Diffusion* and *Diffusion-GC* can be attributed to the AUC on GazeFollow being evaluated with group-level annotations and favoring heatmaps spanning larger areas. We analyzed this in more detail in the supplementary material.

4.4 Semi-supervised Gaze Following for Video

In this experiment, we finetune a pretrained image gaze following model to specific video scenes in a semi-supervised manner. We experiment on 10 TV shows from the VideoAttentionTarget dataset [9], comprising a total of 31,978 gaze

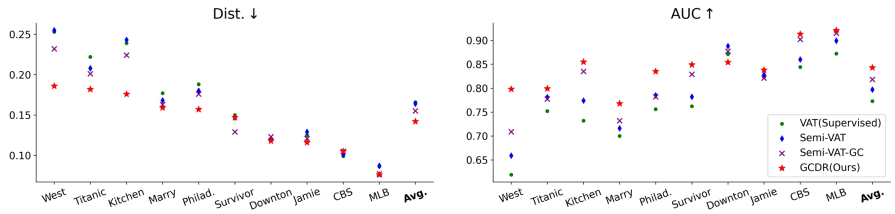

Fig. 6: Applying gaze following to videos in a semi-supervised manner. GCDR achieves the overall best performance (Dist. and AUC averaged across videos: GCDR: 0.142, 0.843; Semi-VAT-GC: 0.155, 0.818; Semi-VAT: 0.164, 0.797; VAT: 0.166, 0.773).

annotations. For each video, we randomly select a clip that contains about 100 annotated frames (2–23% of the entire video, mostly below 10%) as groundtruth data. The remainder of the videos are regarded as unlabeled. Following [9], we started with the image version of the teacher model pre-trained on the GazeFollow dataset and extended it temporally through an additional ConvLSTM network. The model was then fine-tuned on the given annotations and the generated pseudo-annotations for each video. The experiment simulates a real-world scenario where only a few frames of a specific scene are annotated.

Fig. 6 presents the video adaptation experiments results on each of the 10 videos. Our method outperforms other pseudo-annotation generation methods in both metrics on almost all videos. When the baseline models do not perform well, the improvement is more pronounced (2%–5% in both metrics, more details in supplementary). These findings demonstrate that our method can be used effectively in video gaze following, needing to initially label only a few frames.

5 Conclusion

We have proposed the first approach for generating high-quality pseudo-labels for the semi-supervised gaze following task. We leverage the priors from large VL models by computing Grad-CAM heatmaps from a pretrained VQA model that is prompted with a gaze following question. The Grad-CAM heatmaps offer strong guidance to the gaze target, but can be noisy. This led to a novel diffusion-based refinement method that refines these initial pseudo annotations with an annotation prior. Our approach works well on both image and video gaze following tasks with significant savings in the annotation effort. We hope our method will lead to the collection of larger gaze following datasets with annotation efforts similar to current datasets. We plan to apply diffusion-based refinement to "noisy" annotations generated by VL models, in new semi-supervised tasks.

Acknowledgements

This project was partially supported by US National Science Foundation Awards IIS-1763981, IIS-2123920, DUE-2055406, and a gift from Adobe.

References

1. Admoni, H., Scassellati, B.: Social eye gaze in human-robot interaction: a review. Journal of Human-Robot Interaction **6**(1), 25–63 (2017) 1
2. Athiwaratkun, B., Finzi, M., Izmailov, P., Wilson, A.G.: There are many consistent explanations of unlabeled data: Why you should average. ICLR (2019) 2
3. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems **34**, 17981–17993 (2021) 7
4. Bao, J., Liu, B., Yu, J.: Escnet: Gaze target detection with the understanding of 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14126–14135 (June 2022) 3, 9
5. Cañigueral, R., Hamilton, A.F.d.C.: The role of eye gaze during natural social interactions in typical and autistic people. Frontiers in psychology **10**, 560 (2019) 1
6. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 397–406 (2021) 3, 6
7. Chen, B., Jiang, J., Wang, X., Wan, P., Wang, J., Long, M.: Debiased self-training for semi-supervised learning. Advances in Neural Information Processing Systems **35**, 32424–32437 (2022) 4
8. Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In: Proceedings of the European conference on computer vision (ECCV). pp. 383–398 (2018) 3
9. 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) 1, 3, 5, 7, 9, 10, 13, 14
10. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023) 4
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 4
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021) 4
13. Drewes, H.: Eye gaze tracking for human computer interaction. Ph.D. thesis, lmu (2010) 1
14. Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., Zhai, G.: Dual attention guided gaze target detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11390–11399 (2021) 1, 3, 5, 9

15. 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) 2, 4, 9
16. Giannone, G., Nielsen, D., Winther, O.: Few-shot diffusion models. In: NeurIPS 2022 Workshop on Score-Based Methods (2022) 7
17. Graikos, A., Malkin, N., Jojic, N., Samaras, D.: Diffusion models as plug-and-play priors. In: Thirty-Sixth Conference on Neural Information Processing Systems (2022), https://arxiv.org/pdf/2206.09012.pdf 4
18. Graikos, A., Yellapragada, S., Le, M.Q., Kapse, S., Prasanna, P., Saltz, J., Samaras, D.: Learned representation-guided diffusion models for large-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8532–8542 (June 2024) 4
19. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. Advances in neural information processing systems **17** (2004) 4
20. Gupta, A., Tafasca, S., Odobez, J.M.: A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 5041–5050 (2022) 3, 9
21. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 6
22. 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) 4
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 4, 6
24. Horanyi, N., Zheng, L., Chong, E., Leonardis, A., Chang, H.J.: Where are they looking in the 3d space? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2677–2686 (2023) 4
25. Hu, Z., Yang, D., Cheng, S., Zhou, L., Wu, S., Liu, J.: We know where they are looking at from the rgb-d camera: Gaze following in 3d. IEEE Transactions on Instrumentation and Measurement **71**, 1–14 (2022) 3
26. Hu, Z., Yang, Y., Zhai, X., Yang, D., Zhou, B., Liu, J.: Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8907–8916 (2023) 4
27. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. In: 29th British Machine Vision Conference, BMVC 2018 (2019) 2
28. Jin, T., Lin, Z., Zhu, S., Wang, W., Hu, S.: Multi-person gaze-following with numerical coordinate regression. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 01–08. IEEE (2021) 3
29. Jin, T., Yu, Q., Zhu, S., Lin, Z., Ren, J., Zhou, Y., Song, W.: Depth-aware gaze-following via auxiliary networks for robotics. Engineering Applications of Artificial Intelligence **113**, 104924 (2022) 3
30. 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) 2, 4
31. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021) 4

32. Lai, Z., Duan, Y., Dai, J., Li, Z., Fu, Y., Li, H., Qiao, Y., Wang, W.: Denoising diffusion semantic segmentation with mask prior modeling. arXiv preprint arXiv:2306.01721 (2023) 4, 7
33. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=BJ6oOfqge 2, 4
34. 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) 2, 4, 9
35. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022) 4
36. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021) 4
37. Li, Y., Pan, Q., Wang, S., Peng, H., Yang, T., Cambria, E.: Disentangled variational auto-encoder for semi-supervised learning. Information Sciences **482**, 73–85 (2019) 4
38. 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) 3, 5, 9
39. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=w0H2xGHlkw 4
40. Lyu, W., Zheng, S., Ma, T., Chen, C.: A study of the attention abnormality in trojaned berts. arXiv preprint arXiv:2205.08305 (2022) 4
41. Lyu, W., Zheng, S., Pang, L., Ling, H., Chen, C.: Attention-enhancing backdoor attacks against BERT-based models. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023), https://openreview.net/forum?id=L7IW2foTq4 4
42. Massé, B., Ba, S., Horaud, R.: Tracking gaze and visual focus of attention of people involved in social interaction. IEEE transactions on pattern analysis and machine intelligence **40**(11), 2711–2724 (2017) 1
43. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2021) 3, 4, 8
44. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: The Eleventh International Conference on Learning Representations (2022) 2, 4
45. 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) 3
46. Mondal, S., Yang, Z., Ahn, S., Samaras, D., Zelinsky, G., Hoai, M.: Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1441–1450 (2023) 4
47. Morimoto, C.H., Mimica, M.R.: Eye gaze tracking techniques for interactive applications. Computer vision and image understanding **98**(1), 4–24 (2005) 1
48. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. In: ICML. Proceedings of Machine Learning Research, vol. 162, pp. 16805–16827. PMLR (2022) 3, 4, 8

49. 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) 4
50. Peng, C., Celiktutan, O.: Visual saliency guided gaze target estimation with limited labels 4
51. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15691–15701 (2023) 4
52. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2
53. 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 1, 3, 9
54. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10674–10685. IEEE (2022) 4
55. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (3). Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015) 7
56. Sakita, K., Ogawara, K., Murakami, S., Kawamura, K., Ikeuchi, K.: Flexible cooperation between human and robot by interpreting human intention from gaze information. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566). vol. 1, pp. 846–851. IEEE (2004) 1
57. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017) 3
58. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017) 4
59. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022) 4
60. Singh, R., Miller, T., Newn, J., Velloso, E., Vetere, F., Sonenberg, L.: Combining gaze and ai planning for online human intention recognition. Artificial Intelligence **284**, 103275 (2020) 1
61. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach, F.R., Blei, D.M. (eds.) Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. JMLR Workshop and Conference Proceedings, vol. 37, pp. 2256–2265. JMLR.org (2015), http://proceedings.mlr.press/v37/sohl-dickstein15.html 4

62. 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) 2, 4
63. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020) 8, 9
64. Song, Y., Shen, L., Xing, L., Ermon, S.: Solving inverse problems in medical imaging with score-based generative models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=vaRCHVj0uGI 3, 4
65. Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390 (2015) 2
66. Tafasca, S., Gupta, A., Odobez, J.M.: Childplay: A new benchmark for understanding children's gaze behaviour. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20935–20946 (2023) 4
67. 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) 2, 4, 7, 9, 10
68. Tonini, F., Beyan, C., Ricci, E.: Multimodal across domains gaze target detection. In: Proceedings of the 2022 International Conference on Multimodal Interaction. pp. 420–431 (2022) 3
69. Tonini, F., Dall'Asen, N., Beyan, C., Ricci, E.: Object-aware gaze target detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21860–21869 (2023) 4, 9
70. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 4
71. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2192–2200. IEEE (2022) 3
72. 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) 4, 5
73. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering (2022) 2
74. Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., Hoai, M.: Predicting goal-directed human attention using inverse reinforcement learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020). https://doi.org/10.1109/CVPR42600.2020.00027 1
75. Yang, Z., Mondal, S., Ahn, S., Xue, R., Zelinsky, G., Hoai, M., Samaras, D.: Unifying top-down and bottom-up scanpath prediction using transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1683–1693 (2024) 1
76. You, Z., Zhong, Y., Bao, F., Sun, J., Li, C., Zhu, J.: Diffusion models and semi-supervised learners benefit mutually with few labels. Advances in Neural Information Processing Systems **36** (2024) 4
77. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. In: International Conference on Machine Learning. pp. 25994–26009. PMLR (2022) 4

78. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) 2