MTGS: A Novel Framework for Multi-Person Temporal Gaze Following and Social Gaze Prediction

Anshul Gupta **Samy Tafasca** **Arya Farkhondeh** **Pierre Vuillecard** **Jean-Marc Odobez**
Idiap Research Institute, Martigny, Switzerland
École Polytechnique Fédérale de Lausanne, Switzerland
{agupta, stafasca, afarkhondeh, pvuillecard, odobez}@idiap.ch

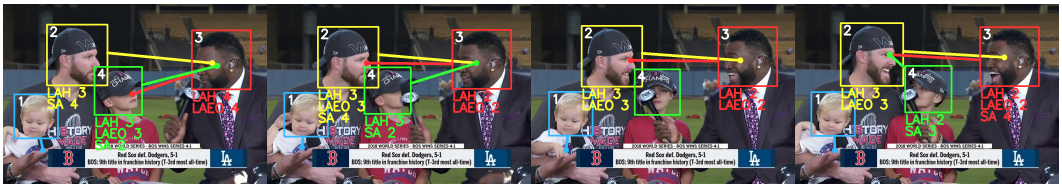


Figure 1: Results of our multi-person and temporal transformer architecture for joint gaze following and social gaze prediction, namely Looking at Humans (LAH), Looking at Each Other (LAEO), and Shared Attention (SA). For each person, the social gaze predictions are listed with the associated person ID (*e.g.* in frame 1, person 2 is in SA with person 4). More qualitative results can be found in the supplementary G.

Abstract

Gaze following and social gaze prediction are fundamental tasks providing insights into human communication behaviors, intent, and social interactions. Most previous approaches addressed these tasks separately, either by designing highly specialized social gaze models that do not generalize to other social gaze tasks or by considering social gaze inference as an ad-hoc post-processing of the gaze following task. Furthermore, the vast majority of gaze following approaches have proposed models that can handle only one person at a time and are static, therefore failing to take advantage of social interactions and temporal dynamics. In this paper, we address these limitations and introduce a novel framework to jointly predict the gaze target and social gaze label for all people in the scene. It comprises (i) a temporal, transformer-based architecture that, in addition to frame tokens, handles person-specific tokens capturing the gaze information related to each individual; (ii) a new dataset, VSGaze, built from multiple gaze following and social gaze datasets by extending and validating head detections and tracks, and unifying annotation types. We demonstrate that our model can address and benefit from training on all tasks jointly, achieving state-of-the-art results for multi-person gaze following and social gaze prediction. Our annotations and code will be made publicly available.

1 Introduction

Social interaction plays a pivotal role in our daily lives and is influenced by an array of behavioral elements, encompassing not only verbal communication but also non-verbal cues like gestures (53) or body language. In particular, the ability to decode people's gaze, including communicative behaviors like eye contact and shared attention on a particular object, is highly related to our capacity to connect with or learn from others (44). In contrast, the absence or impairment of these skills is often indicative of developmental disorders such as autism (42). Thus, designing social gaze prediction algorithms

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-0500

and systems has attracted considerable attention from different communities, ranging from medical diagnosis to human-robot interactions (42; 43; 36).

In this work, we investigate whether we can build a unified framework to infer from video data the *gaze target* and *social gaze label* in *one stage* for *all people* in the scene. This requires: (i) A new architecture capable of jointly modeling these tasks, (ii) A large-scale dataset with annotations for all the tasks. Specifically, we focus on the LAH, LAEO and SA social gaze tasks (illustrated in Fig. 1).

Methods for social gaze prediction in the literature adopt one of two approaches. The first one focuses on the design of dedicated networks to process pairs of head crops and potentially other scene information (31; 30; 11; 5; 12; 46). While their specialization makes them effective, they offer little room for generalization to other gaze-related tasks. The second one first address the gaze following tasks (40), defined as predicting the 2D location people's gaze targets, and then uses the predicted gaze points to infer social gaze through ad-hoc post-processing schemes. For instance, combining gaze following heatmaps from multiple people to predict shared attention (9).

However, gaze following itself is a challenging task. Besides geometric aspects, the task requires understanding and establishing a correspondence between top-down information related to the person's state, activity, cognitive intent, and bottom-up saliency related to the scene context (salient items like objects or talking people). Furthermore, gaze following performance, as measured by distance, does not always translate to similar social semantic performance (*e.g*. when evaluating if the predicted gaze point falls on a person's head or not (48)).

Motivation. Existing methods for gaze following suffer from several drawbacks. First, most of the methods perform prediction for a single person (40; 41; 9; 14; 17; 48), requiring multiple inference passes on the same image to process multiple people in the scene. In contrast, a multi-person gaze following architecture processes the image only once and has to capture salient items for all people in the scene, while maintaining the ability to infer the gaze target of each individual. This is inherently more complex and challenging. Another drawback of the single-person formulation is that it does not explicitly model people's interactions, thereby preventing the possibility of jointly inferring people's gaze target and social gaze attributes. Secondly, the majority of proposed models for gaze following are static, using only a single image at a time. This is partly due to the absence of large and diverse video datasets, and the difficulty of leveraging large-scale static ones like GazeFollow (41). This is a limitation, as temporal information can capture head and gaze coordination patterns (43) which can help gaze direction inference, especially when the eyes are not completely visible (34).

Finally, none of these methods have investigated learning the gaze following and social gaze prediction tasks jointly. Thus, it remains a research question whether such formulation can improve performance by having social cues inform the gaze following task and vice-versa, or if performance would degrade as we try to accommodate multiple tasks, datasets, and people within the same framework.

Contributions. Given these motivations, we propose a new, unified framework for gaze following and social gaze prediction with the following contributions:

- A novel temporal and multi-person architecture for gaze following and social gaze prediction (Sec. 3). Our approach posits people as specific tokens that can interact with each other and the scene content (*i.e.* frame tokens). This token-based multi-person representation allows for the modeling of (i) temporal information at multiple levels (from 2D gaze direction to 2D gaze target level), (ii) the joint prediction of the gaze target and social gaze label.
- A new dataset, VSGaze, that unifies annotation types across multiple gaze following and social gaze datasets (Sec. 4.1.1).
- New social gaze protocols and metrics for better evaluating semantic gaze following performance (Sec. 4.3).

Results. In our experiments (Sec. 5), we show that our architecture achieves state-of-the-art results for multi-person gaze following, while also performing competitively against single-person models. It is also able to leverage our proposed VSGaze dataset to jointly tackle gaze following and social gaze prediction, achieving competitive performance compared to methods trained on individual tasks. In particular, our experiments further demonstrate that the performance *benefits from this joint prediction,* i.e. *adding the social loss, improves gaze following performance, and vice-versa*. Finally, the new social gaze metrics provide complementary information to the standard distance-based metrics, helping assessing model performance from the social interaction perspective.

It is worth noting that our architecture is easily extendable and allows for the integration of auxiliary person-specific information that can influence the final predictions. In the supplementary F, we explore this aspect by integrating people's speaking status in the person tokens to improve the results.

2 Related work

Gaze Following. Typical methods for this task exploit a two-branch architecture: one for processing the scene and the other for processing the person of interest (40; 9; 14; 17; 22; 23; 26). They have distinguished themselves by the addition of other relevant modalities like depth (14; 17; 23), pose (17), and objects (20), or by potentially leveraging scene geometry (19; 23; 14). However, only a few efforts have addressed the multi-person case. (22) first proposed a simple architecture relying on a scene backbone to get a person-agnostic image representation that is subsequently fused with the head crop features of each individual obtained using another backbone. While this reduces computation, the model does not account for person interactions, as each head is processed separately. In another direction, (52; 51) rely on a transformer-based architecture to perform multi-person gaze following. Their methods borrow from DETR (6), taking the image as input and simultaneously predicting the head bounding box and gaze target for every person in the scene. While these methods can implicitly model person interactions, an important limitation is that they compute performance on detected heads which are matched to the ground truth. Given that both the head detection and matching steps are error-prone, it precludes comparing their results to others. We provide examples in the supplementary (Fig. 8) where (51) makes incorrect head detections.

Temporal gaze estimation. Temporal information has proven effective for 3D gaze estimation. Previous research developed models to learn from various inputs, including face, eyes, and facial landmarks using a multi-stream recurrent CNN (37); eyes and visual stimuli or raw RGB frames from in-the-wild settings using convolutional RNNs or LSTMs (38; 25); and the temporal coordination of gaze, head, and body orientations using LSTMs (34). However, the use of such methods for the gaze following task in arbitrary scenes has been underexplored. The only exceptions are (9) who introduced a convolutional LSTM block at the bottleneck of the heatmap prediction architecture, and (32) who leveraged temporal attention over aggregated frame-level features. However, both approaches only showed a slight improvement compared to their static versions, highlighting the challenge of exploiting temporal information for this task. Conceptually, the methods did not model 2D gaze direction dynamics, and can not be extended for multi-person gaze inference.

Social gaze prediction. Several research papers in the literature are dedicated to the study of looking at each other (LAEO) and shared attention (SA) tasks. For LAEO, most methods rely on processing the head crops to obtain some gaze directional information, and then combining it with 2D or inferred 3D geometric information to predict the LAEO label (31; 11; 30; 5). Drawbacks include processing pairs of persons independently and, as they only process heads and do not address gaze following, lacking global image context and not extending easily to other social tasks like SA. Similarly to (52; 51), a recent paper (15) proposed an encoder-decoder transformer architecture to predict heads and the LAEO labels, and while achieving good results, suffers from the same drawbacks.

Regarding shared attention, the first method to address it in the wild was (12), which framed the problem as 2 tasks: the binary classification of whether SA occurs in a frame, and the inference of the location of the SA target object. Their method combined predicted 2D gaze cones of people in the scene with a heatmap of object region proposals, while others (46) directly inferred SA from the raw image. Since then, several methods leveraged combining gaze following heatmap predictions of all people (9; 52) and improved performance. Nevertheless, the above task formulation (12) used by all papers suffers from two main issues: (i) it cannot distinguish between multiple SA instances if they occur in the same frame; (ii) it does not determine which specific people are sharing attention. Our work solves both problems by framing the task as a binary classification between pairs of people. This formulation is more natural and has the benefit of extending to other social gaze tasks.

Finally, none of the previous works performed both social gaze prediction tasks. There are three notable and interesting exceptions. First, (13) who addressed the inference of gaze communication activities (atomic and events, including LAEO and SA) using a graph-based approach with 12 dimensional tokens. However, as it predicts a single gaze 'state' for each person, it is problematic as it does not allow for simultaneous LAEO and SA. It also does not identify the other person involved in the social gaze interaction. The second is (7) who addressed dyadic communication by proposing a gaze following style 2-branch architecture processing order-dependent pairs of people. However, inference using (7) is inefficient because the model needs $\frac{N_p!}{(N_p-2)!}$ forward passes to consider all pairwise relationships for a given scene with N_p people. Also both methods do not address the gaze following task. More recently, (16) did address all tasks using a graph approach, but there was no temporal modeling nor joint training on all tasks.

 https://doi.org/10.52202/079017-0500

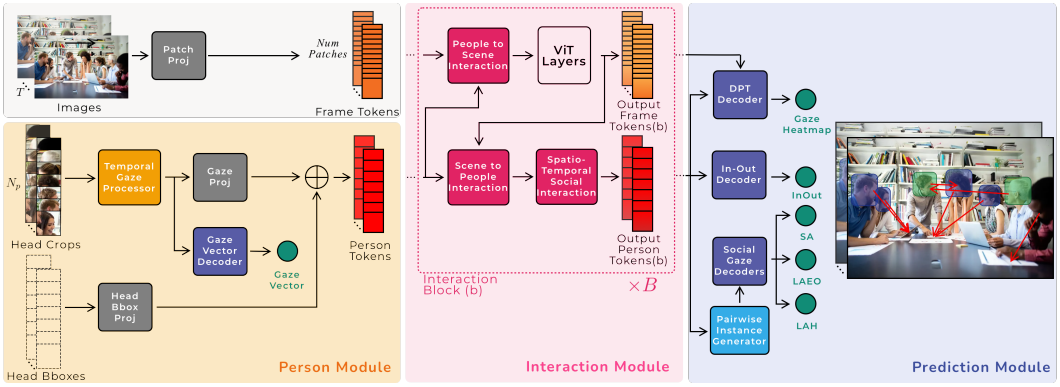


Figure 2: Proposed architecture for multi-person temporal gaze following and social gaze prediction. See approach overview in Section 3.

3 Architecture

As motivated in Section 1, designing a temporal architecture for gaze following is challenging given the lack of large scale video datasets. Since the broader scene tends to remain relatively static for a short temporal window, we focused on modeling person-level temporal information to simplify the learning problem. In particular, our architecture processes person and scene tokens through separate transformers, with a temporal transformer for processing the person tokens. At the same time, it facilitates interactions between the scene and person tokens via cross-attention.

Approach overview. Our approach is illustrated in Fig. 2. It takes as input a sequence of $t = 1 \dots T$ frames, as well as the head bounding box tracks $\mathbf{h}^{\text{box}}_{i,1:T}$ and corresponding head crops $\mathbf{h}^{\text{crop}}_{i,1:T}$ which are assumed to have been extracted for each of the $i = 1 \dots N_p$ persons. The outputs are the sequence of gaze heatmaps $\mathcal{A}_{i,1:T}$ and in-out gaze labels $\mathbf{o}_{i,1:T}$ for each person i, as well as the sequence of per-frame pair-wise social gaze labels for each task and pair $i, j \in \{1 \dots N_p\}$: $\mathbf{e}_{i \to j,1:T}$ for LAH, $\mathbf{e}_{i \leftrightarrow j,1:T}$ for LAEO, and $\mathbf{c}_{i,j,1:T}$ for SA.

The model proceeds as follows. First, each frame t is processed by a standard ViT tokenizer to produce the set of patch-wise frame tokens $\mathbf{f}_t$, resulting in a sequence of frame tokens $\mathbf{f}_{1:T}$. In parallel, the Person Module processes the sequence of head crops from each person i using the Temporal Gaze Processor, and the resulting sequence outputs are then tokenized at each frame along with the bounding box locations to produce the sequence of person token $\mathbf{p}_{i,1:T}$. Secondly, the Interaction Module jointly processes the frame and person tokens, iteratively updating them at each time step through person-scene cross-attention interaction components and scene ViT self-attention, and in time through person spatio-temporal social interaction components. Finally, the Prediction Module processes at each time step the resulting frame and person tokens (from multiple blocks) to infer the sequence of gaze heatmaps and in-out gaze labels for each person, as well as pair-wise social gaze labels. We detail the three modules in the next sections.

3.1 Person Module

This module aims to model person-specific information relating to gaze and head location.

Temporal Gaze Processor. It aims to capture all gaze-related information (direction, dynamics). First, individual head crops $\mathbf{h}^{\text{crop}}_{i,t}$ are processed by a Gaze Backbone $\mathcal{G}_{\text{stat}}$ to produce gaze embeddings according to $\mathbf{g}^{\text{stat}}_{i,t} = \mathcal{G}_{\text{stat}}(\mathbf{h}^{\text{crop}}_{i,t})$. Then, to model the gaze dynamics of a person, we rely on a Temporal Gaze Encoder $\mathcal{G}_{\text{temp}}$ to process the sequence $\mathbf{g}^{\text{stat}}_{i,1:T}$ of gaze embeddings plus learnable temporal position embeddings $\mathbf{x}_{1:T}$ and obtain their temporal counterparts: $\mathbf{g}^{\text{temp}}_{i,1:T} = \mathcal{G}_{\text{temp}}(\mathbf{g}^{\text{stat}}_{i,1:T} + \mathbf{x}_{1:T})$. $\mathcal{G}_{\text{temp}}$ is implemented as a single Transformer layer with self-attention. Finally, to supervise the learning of relevant gaze embeddings, we attach a Gaze Vector Decoder that predicts a person's 2D gaze vector at each time step, $\mathbf{g}^{\text{v}}_{i,t} = \mathcal{G}_{\text{vec}}(\mathbf{g}^{\text{temp}}_{i,t})$, where $\mathcal{G}_{\text{vec}}$ is implemented as a 2-layer MLP.

Person tokenization. The person tokens are obtained by projecting the temporal gaze embeddings and normalized 4d head box locations using learnable linear layers ($\mathcal{P}_{\text{gaze}}$ and $\mathcal{P}_{\text{box}}$ respectively) to tokens of same dimension than frame token, and adding them together:

https://doi.org/10.52202/079017-0500

$$\mathbf{p}_{i,t} = \mathcal{P}_{\text{gaze}}(\mathbf{g}_{i,t}^{\text{temp}}) + \mathcal{P}_{\text{box}}(\mathbf{h}_{i,t}^{\text{box}}). \quad (1)$$

3.2 Interaction Module

The Interaction module aims at modeling the exchange of information between persons and the scene at each time step, as well as the spatio-temporal social interactions between people. One important goal of this process is to align the person and frame token representations so that (i) *person-specific* gaze heatmaps can be predicted from the set of output frame tokens and each person output token; (ii) in-out gaze and social gaze prediction can be made from the person tokens.

To do so, we designed the module to consist of B blocks, each comprising Person-Scene Interaction and Spatio-Temporal Social Interaction components. The input to the first block is the set of person tokens $\mathbf{p}_{1:N_p,1:T}$ from the Person Module, and the frame tokens $\mathbf{f}_{1:T}$. Each block then processes the set of output[1] person tokens $\mathbf{p}^{\text{o,b-1}}_{1:N_p,1:T}$ and output frame tokens $\mathbf{f}^{\text{o,b-1}}_{1:T}$ from the previous block, and returns updated tokens after a series of self/cross-attention layers through the components.

Person-Scene Interaction. This component models the interactions between people and the scene and can capture inferring gaze to scene objects or body parts like hands or exploit some global context. It is inspired by ViT-Adaptor (8) which has shown good performance for dense prediction tasks when relying on pretrained models and small amounts of data for the target task. It proceeds in 3 steps:

(i) People-to-Scene Encoder $\mathcal{I}^{\text{b}}_{\text{ps}}$: it updates the frame tokens with person information relevant to gaze by processing the frame tokens $\mathbf{f}^{\text{o,b-1}}_t$ and frame-level person tokens $\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}$ according to $\mathbf{f}^{\text{p,b}}_t = \mathcal{I}^{\text{b}}_{\text{ps}}(\mathbf{f}^{\text{o,b-1}}_t, \mathbf{p}^{\text{o,b-1}}_{1:N_p,t})$. It is implemented as a single Transformer layer with cross-attention, where $\mathbf{f}^{\text{o,b-1}}_t$ generate the queries and $\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}$ generate the keys and values.

(ii) The updated frame tokens $\mathbf{f}^{\text{p,b}}_t$ pass through the standard set of of ViT layers $\mathcal{V}_b$ to process the scene information, resulting in the output frame tokens for the block b: $\mathbf{f}^{\text{o,b}}_t = \mathcal{V}_b(\mathbf{f}^{\text{p,b}}_t)$.

(iii) Scene-to-People Encoder $\mathcal{I}^{\text{b}}_{\text{sp}}$: it updates the person tokens so that they capture location information related to the salient items they are probably looking at. It works by processing the frame-level person tokens $\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}$ and obtained frame tokens $\mathbf{f}^{\text{o,b}}_t$ according to: $\mathbf{p}^{\text{s,b}}_{1:N_p,t} = \mathcal{I}^{\text{b}}_{\text{sp}}(\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}, \mathbf{f}^{\text{o,b}}_t)$. It is also implemented as a single Transformer layer with cross-attention, where the set $\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}$ generates the queries and $\mathbf{f}^{\text{o,b}}_t$ generates the keys and values.

Spatio-temporal Social Interaction. This component allows the sharing of information between people and of the alignment of their representations for social gaze prediction. This also include modeling the temporal evolution of individual tokens. To achieve this, a Social Encoder $\mathcal{I}^{\text{b}}_{\text{pp}}$ first processes and updates the frame-level person tokens $\mathbf{p}^{\text{s,b}}_{1:N_p,t}$ to capture interactions between people at each frame, according to: $\mathbf{p}^{\text{p,b}}_{1:N_p,t} = \mathcal{I}^{\text{b}}_{\text{pp}}(\mathbf{p}^{\text{s,b}}_{1:N_p,t})$. It is followed by a Temporal Person Encoder $\mathcal{I}^{\text{b}}_{\text{pt}}$ that processes the updated person token sequences $\mathbf{p}^{\text{p,b}}_{i,1:T}$ of each person i and updates them to capture temporal patterns of attention, resulting in the output person tokens for the block: $\mathbf{p}^{\text{o,b}}_{i,1:T} = \mathcal{I}^{\text{b}}_{\text{pt}}(\mathbf{p}^{\text{p,b}}_{i,1:T})$. Both $\mathcal{I}^{\text{b}}_{\text{pp}}$ and $\mathcal{I}^{\text{b}}_{\text{pt}}$ are implemented as a single Transformer layer with self-attention.

3.3 Prediction Module

The Prediction Module processes the set of output person $\{\mathbf{p}^{\text{o,b}}_{1:N_p,1:T}\}$ and frame $\{\mathbf{f}^{\text{o,b}}_{1:T}\}$ tokens from all Interaction Module blocks to predict the person-specific gaze heatmaps and in-out labels, as well as the pair-wise social gaze labels.

Gaze Heatmap Prediction. Here, we follow the model introduced in (49) which takes inspiration from the DPT decoder (39) for dense prediction tasks, and adapts it to handle multiple heatmap predictions from the same ViT outputs. This is performed *by conditioning the decoding on each person's token*. The standard DPT decodes the image features from multiple layers of a ViT in a Feature Pyramid Network (27) style. It works by fusing at block level b the feature maps from level $b+1$ after an upsampling stage, and the feature maps computed by a reassemble stage from the ViT output of block b. We aim to apply this approach to the frame tokens $\{\mathbf{f}^{\text{o,b}}_{1:T}, b = 1:B\}$,

[1]Note that the superscript o, p and s do not represent indices, but intermediate token updates within a block b.

https://doi.org/10.52202/079017-0500

but conditioned on a specific person. In our model, this is achieved through a modification in the reassemble stage, in which the image feature maps produced by the standard reassemble stage are multiplied at every location (using a Hadamard product) with the projected person token $\mathbf{p}_{i,t}^{\text{o,b}}$ of that same block level. The gaze heatmap $\mathcal{A}_{i,t}$ for each person at each frame is thus obtained as:

$$\mathcal{A}_{i,t} = \mathcal{D}(\{(\mathbf{f}_t^{\text{o,b}}, \mathbf{p}_{i,t}^{\text{o,b}}), b = 1 : B\}) \tag{2}$$

where $\mathcal{D}$ denotes this conditional DPT. See (49) and supplementary H for details.

Social Gaze Prediction. This decoder processes the person tokens from all B Interaction Module blocks to predict the social gaze label for every pair of people in every frame. In practice, the B tokens $\{\mathbf{p}_{i,t}^{\text{o},1} \cdots \mathbf{p}_{i,t}^{\text{o},B}\}$ corresponding to a single person in a frame are linearly projected and concatenated to produce a multi-scale person token $\mathbf{p}_{i,t}^{\text{ms}}$. Then, to predict a social gaze label, pairs of these tokens are concatenated and processed by the decoders E for LAH and C for SA (illustrated through the Pairwise Instance Generator in Fig. 2). Their outputs are the predicted LAH score $\mathbf{e}_{i\to j,t}$ for person i looking at j, and the predicted SA score $\mathbf{c}_{i,j,t}$ for i, j.

$$\mathbf{e}_{i\to j,t} = E(\mathbf{p}_{i,t}^{\text{ms}}, \mathbf{p}_{j,t}^{\text{ms}}) \text{ and } \mathbf{c}_{i,j,t} = C(\mathbf{p}_{i,t}^{\text{ms}}, \mathbf{p}_{j,t}^{\text{ms}}). \tag{3}$$

E and C are implemented as 3-layer MLPs with residual connections. For LAEO, both people i, j need to be looking at each other for a positive label, and either one can be looking away for a negative label. Hence, we simply compute the LAEO score $\mathbf{e}_{i\leftrightarrow j,t}$ as $\min(\mathbf{e}_{i\to j,t}, \mathbf{e}_{j\to i,t})$.

In-Out Prediction. This decoder $\mathcal{O}$ processes the multi-scale person tokens $\mathbf{p}_{i,t}^{\text{ms}}$ to predict at every frame whether people are looking inside the frame or outside the frame, as $\mathbf{o}_{i,t} = \mathcal{O}(\mathbf{p}_{i,t}^{\text{ms}})$, where $\mathcal{O}$ is implemented as a 5-layer MLP with residual connections.

3.4 Losses

The total loss $\mathcal{L}$ is a linear combination of the gaze heatmap loss $\mathcal{L}_{\text{HM}}$, gaze vector loss $\mathcal{L}_{\text{VEC}}$, social gaze losses $\mathcal{L}_{\text{LAH}}, \mathcal{L}_{\text{SA}}$ and the in-out loss $\mathcal{L}_{\text{IO}}$:

$$\mathcal{L} = \lambda_{\text{HM}}\mathcal{L}_{\text{HM}} + \lambda_{\text{VEC}}\mathcal{L}_{\text{VEC}} + \lambda_{\text{LAH}}\mathcal{L}_{\text{LAH}} + \lambda_{\text{SA}}\mathcal{L}_{\text{SA}} + \lambda_{\text{IO}}\mathcal{L}_{\text{IO}} \tag{4}$$

$\mathcal{L}$ is applied at each time step per person for $\mathcal{L}_{\text{HM}}, \mathcal{L}_{\text{VEC}}, \mathcal{L}_{\text{IO}}$, and per pair for $\mathcal{L}_{\text{LAH}}, \mathcal{L}_{\text{SA}}$. All losses are standard: $\mathcal{L}_{\text{HM}}$ is defined as the pixel-wise MSE loss between the GT and predicted heatmaps, $\mathcal{L}_{\text{VEC}}$ as the cosine loss, and the social gaze and in-out losses as binary cross-entropy losses. Since LAEO is inferred from LAH predictions (Sec 3.3), we do not have any LAEO loss.

4 Experiments

4.1 Datasets

We perform experiments on multiple gaze following and social gaze datasets.

GazeFollow (41) is a large-scale static dataset for gaze following, featuring 122K images. Most images are annotated for a single person with their head bounding box and gaze target point. The test set contains gaze point annotations by multiple annotators. Despite lower quality images and annotations, given its rich diversity, it remains a good dataset to use for pre-training.

VideoAttentionTarget (VAT) (9) is a video dataset, annotated with head bounding boxes, gaze points, and inside vs outside frame gaze for a subset of the people in the scene. It contains 1331 clips collected from 50 shows on YouTube.

ChildPlay (48) is a recent video dataset for gaze following, annotated with head bounding boxes, gaze points, and a label indicating 7 non-overlapping gaze classes including inside frame, outside frame and gaze shifts. It contains 401 clips from 95 YouTube videos, and features children playing and interacting with other children and adults.

VideoCoAtt (12) is a video dataset for shared attention estimation, containing 380 videos or 492k frames from TV shows. When a shared attention behavior occurs (i.e. about 140k frames), the relevant frames are annotated with the bounding box of the target object, as well as the head bounding boxes of the people involved.

UCO-LAEO (31) is a video dataset for LAEO estimation, annotated with head bounding boxes, and a label indicating whether two heads are LAEO. It contains 22,398 frames from 4 TV shows.

Dataset	Gaze Points	LAH	LAEO	SA
GazeFollow (41)	118k	27k/493k	0	0
VAT (9)	109k	74k/729k	13k/461k	16k/94k
ChildPlay (48)	217k	59k/682k	7k/351k	4k/55k
VideoCoAtt (12)	367k	290k/1551k	0	400k/918k
UCO-LAEO (31)	21k	21k/36k	10k/54k	0
VSGaze	714k	444k/2998k	30k/866k	420k/1067k

Table 1: Person-wise gaze point and pair-wise social gaze annotation (positive/negative) statistics for our datasets. VSGaze unifies annotation types across VAT, ChildPlay, VideoCoatt and UCO-LAEO.

Two other interesting datasets with annotations for multiple social gaze behaviours are VACATION (13) and GP-static (7). However, VACATION does not provide annotations for all social gaze behaviours when they occur simultaneously (examples in supplementary Fig. 7). This annotation scheme is problematic and is linked to their method design as described in Sec. 2. On the other hand, GP-static only considers dyadic interactions and is not publicly available.

4.1.1 VSGaze dataset

A limitation of the above datasets is that they only contain annotations for gaze following or specific social gaze tasks. Hence, we propose the **V**ideo dataset with **S**ocial gaze and **Gaze** following annotations or **VSGaze** dataset. VSGaze extends head track annotations and unifies annotation types across VAT, ChildPLay, VideoCoAtt and UCO-LAEO. This allows for joint training of gaze following and social gaze, and provides new tasks and metrics for evaluating performance on the component datasets. The construction of VSGaze is described below.

Extending Head Track Annotations. As each dataset contains annotations for only a subset of people in the scene, we detect all missing heads using the pre-trained Yolov5 head detection model (24) used by Tafasca *et al*. (48), and track them using the ByteTrack algorithm (54). We further manually verified the accuracy of the obtained tracks. This step is vital to obtain positive and negative social gaze pairs. Consider a scene with 3 people, i, j, k, where i is looking at j. If only i is annotated, the positive LAH pair $i \rightarrow j$ would be missed. The negative LAH pair $i \rightarrow k$ would also be missed.

Unifying Gaze Following and Social Gaze Annotations. Given the extended set of head bounding box annotations, as well as existing gaze following and social gaze annotations, we then process these annotations to obtain gaze following and social gaze labels across all the datasets. The gaze target for UCO-LAEO and VideoCoAtt is set as the center of the LAEO and SA target bounding box respectively. LAH pairs are obtained by checking if the gaze target falls inside another person's head box. For LAEO we check the reverse as well. SA pairs are obtained by checking if the gaze targets for both people fall in the same head box. We provide the detailed protocol in the supplementary B.

Annotation statistics are summarized in Table 1. Overall, VideoCoatt is the largest source of annotations except for LAEO. We also see that the pair-wise annotations are skewed towards negative cases. The statistics further provide insight into the content of the datasets. As VAT, VideoCoAtt and UCO-LAEO contain clips from TV shows, there are many more instances of looking at other people and at each other. On the other hand for ChildPlay, LAH mainly occurs when the supervising adult looks at a child and there is limited LAEO.

4.2 Training and Validation

We follow standard practice (9; 14; 17) by first training the static version of our model (i.e. with no temporal attention $\mathcal{G}_{\text{temp}}$, $\mathcal{I}_{\text{pt}}^{\text{b}}$) on GazeFollow. It is trained for 20 epochs with a learning of rate of 1×10^{-4}. The resulting weights then serve as initialization for our proposed temporal model. We freeze the ViT $\mathcal{V}$ and train the temporal model on VSGaze for another 20 epochs with a learning rate of 3×10^{-6}. For validation, we use the provided splits for UCO-LAEO, VideoCoAtt and ChildPlay, and the splits proposed by Tafasca *et al*. (48) for GazeFollow and VAT. For all our experiments, we use $T = 5$ frames with a temporal stride of 3. To allow for batch training, we randomly sample up to $N_p = 4$ people in a scene (padding in case there are less). During testing, we evaluate per sample and consider all people. The Interaction Module consists of $B = 4$ blocks , interacting with $\mathcal{V}$ at layers $\{2, 5, 8, 11\}$. Additional implementation details are provided in the supplementary E.

https://doi.org/10.52202/079017-0500

Method	PP	Dist. ↓	AP_{IO} ↑	$F1_{LAH}$ ↑	$F1_{LAEO}$ ↑	AP_{SA} ↑
$Chong_S$* (9)	✓	0.121	0.918	0.778	0.562	0.288
$Chong_T$* (9)	✓	0.130	**0.956**	0.764	0.529	0.331
Gupta* (17)	✓	0.119	0.929	0.784	0.590	0.335
Ours-noGF	✗	-	-	0.738	0.579	<u>0.515</u>
Ours-noSoc	✓	<u>0.111</u>	<u>0.945</u>	<u>0.802</u>	<u>0.598</u>	0.339
Ours	✗	**0.107**	0.940	0.795	0.590	**0.576**
Ours-PP	✓	**0.107**	0.940	**0.812**	**0.603**	0.352

Table 2: Comparison against gaze following methods on VSGaze. All models were trained on VSGaze. PP indicates social gaze predictions from post-processing gaze following outputs (✓) vs predictions from decoders (✗). Best results are in bold, second best results are underlined.

4.3 Evaluation

Gaze Following. We use the standard metrics:

- *AUC*: for GazeFollow, the predicted heatmap is compared against a binary GT map with value 1 at annotated gaze point positions, to compute the area under the ROC curve.
- *Distance (Dist.)*: the arg max of the heatmap provides the gaze point. We can then compute the L2 distance between the predicted and GT gaze point on a 1×1 square. For GazeFollow, we compute Minimum (Min.) and Average (Avg.) distance against all annotations.
- *In-Out AP (AP_{IO})*: it is the Average Precision (AP) of the In-Out gaze prediction scores.

Social Gaze. We propose an evaluation protocol for LAH, as well as a new pair-wise evaluation protocol for SA as motivated in Sec. 2.

Social gaze decoders: For the LAH task we compute an F1 score. A sample is an individual person i, and at inference, it is assigned the person $\hat{j}$ for which the $\mathbf{e}_{i \to j,t}$ is the largest. Hence, for a GT positive case, the prediction will be considered as a true positive if $\hat{j}$ matches the GT target AND $\mathbf{e}_{i \to \hat{j},t}$ is above 0.5. Otherwise, the prediction is a false negative. Similarly, a GT negative is a true negative if $\mathbf{e}_{i \to \hat{j},t}$ is below 0.5, otherwise it is a false positive. For LAEO, as with LAH, we compute an F1 score. A sample is a pair of people i, j, and for person i, we consider $\mathbf{e}_{i \leftrightarrow \hat{j},t}$ with the highest score. We then set $\mathbf{e}_{i \leftrightarrow j,t}$ to 0 $\forall j \neq \hat{j}$ before computing the performance. We also do the reverse for j, i. For SA, we compute a standard AP score by considering a sample as a pair of people, and thresholding predicted scores at different values to compute the area under the Precision-Recall curve.

Gaze following methods: Social gaze predictions are obtained by post-processing the predicted gaze points. For LAH, we check whether the predicted gaze point for a person falls inside the target person's head box. For LAEO we check the reverse as well. We compute F1 scores for both. For SA, we check if the distance between two people's predicted gaze points is within a set of thresholds and compute an AP score.

5 Results

5.1 Comparison against the State-of-The Art

We compare against recent SoTA methods addressing either social gaze tasks or gaze following. In addition, for fairness and to evaluate the impact of the VSGaze dataset, we also re-trained on this dataset the static image based models of Chong (9) ($Chong_S$*) and Gupta (17) (Gupta*), as well as the temporal model of (9) ($Chong_T$*), the only temporal gaze following model with available code.

VSGaze. The results on VSGaze are given in Table 2. Note that regarding our approach, for social gaze, we compute the scores by leveraging either the predictions from the respective task decoders (Ours), or by post-processing the gaze following outputs of our model (Ours-PP).

Compared to the baselines, we observe that our model achieves the best performance for all tasks except for in-out gaze prediction. In particular, we achieve significant gains in the distance and AP_{SA} metrics when leveraging the predictions from the SA decoder. The latter highlights the importance of modeling SA as a classification task compared to post-processing gaze following outputs, which struggles to capture whether the gaze points for a pair of people falls on the same semantic item.

In addition, we note that better gaze following performance does not always translate to better social gaze performance. For instance, although $Chong_S$* has a better distance score compared to $Chong_T$*, it performs worse for shared attention. This effect is even more pronounced on ChildPlay (Supp C), and suggests the benefit of considering social gaze metrics for better characterizing the performance

Method	Multi	AUC↑	Avg.Dist.↓	Min.Dist.↓
Fang (14)	✗	0.922	0.124	0.067
Tonini (50)	✗	0.927	0.141	-
Jin (23)	✗	0.920	0.118	0.063
Bao (4)	✗	0.928	0.122	-
Hu (21)	✗	0.923	0.128	0.069
Tafasca (48)	✗	0.936	0.125	0.064
Chong$_S$ (9)	✗	0.921	0.137	0.077
Gupta (17)	✗	0.933	0.134	0.071
Jin (22)	✓	0.919	0.126	0.076
Ours-static	✓	0.929	**0.116**	**0.059**

(a) Results on GazeFollow (41).

Method	Multi	Dist.↓	AP_{IO} ↑
Fang (14)	✗	0.108	0.896
Tonini (50)	✗	0.129	-
Jin (23)	✗	0.109	0.897
Bao (4)	✗	0.120	0.669
Hu (21)	✗	0.118	0.881
Tafasca (48)	✗	0.109	0.834
Chong$_T$ (9)	✗	0.134	0.853
Gupta (17)	✗	0.134	0.864
Jin (22)	✓	0.134	**0.880**
Ours	✓	**0.105**	0.869
Ours†	✓	**0.105**	0.869

(b) Results on VAT (9).

Method	Multi	Dist.↓	AP_{IO} ↑
Tafasca (48)	✗	0.107	0.986
Gupta (17)	✗	0.113	0.983
Ours	✓	0.117	**0.994**
Ours†	✓	**0.113**	0.993

(c) Results on ChildPlay (48).

Method	Dist.↓	AP_{LAEO} ↑
Jiminez (31)	-	0.795
Doosti (11)	-	0.762
Jiminez (30)	-	0.867
Ours	0.023	0.963
Ours†	**0.019**	**0.974**

(d) Results on UCO-LAEO (31).

Table 3: Comparison against task specific methods fine-tuned on individual datasets. Best multi-person results are in bold, overall best results are underlined. Multi indicates multi-person (✓) vs single-person (✗) gaze following methods. Ours is initialized from training on GazeFollow, while Ours† is initialized from training on VSGaze.

of gaze following models, especially its semantic performance. In the supplementary C, we provide a breakdown of performance on each of the component datasets of VSGaze.

State-of-the-art comparison: fine tuning on individual datasets. Table 3 compares our model against task specific methods. For GazeFollow, we use our static model (Ours-static) that was trained on GazeFollow and used to initialize our model trained on VSGaze. For the video datasets, as SoTA methods were trained (or finetuned) on individual datasets, for fairness we also fine-tune our model on these datasets, investigating two initialization alternatives: either from the model trained on GazeFollow (Ours), or from the model trained on VSGaze (Ours†). Note that we are unable to compare against previous results for VideoCoatt due to our new pair-wise evaluation protocol that better captures SA performance (Sec 4.3).

On GazeFollow and VAT, our model outperforms the only other comparable multi-person gaze following model of Jin (22). It also achieves competitive or better results to single-person methods, even those leveraging auxiliary modalities such as depth (14; 50; 4; 23; 21; 48). Importantly, on the social LAEO task, we set the new state of the art on UCO-LAEO, far outperforming methods designed specifically for LAEO (31; 30; 11).

We also note that fine-tuning using the VSGaze model initialization can improve results compared to the standard protocol of fine-tuning after training on GazeFollow (ex. distance on ChildPlay and AP_{LAEO} on UCO-LAEO). This suggests that training on VSGaze can leverage the complementary knowledge provided by the different tasks and datasets, which follows observations made in other works addressing multi-task training (10).

5.2 Analysis

Impact of Architecture. Comparing the performance of our model with no social gaze losses (Ours-noSoc) against the baselines (Table 2), we see that it already performs on par or better than them while being much more efficient as it processes the image only once for all people in the scene. It also serves as a strong gaze following baseline to compare performance against.

Impact of Social Gaze Loss. Our architecture can further benefit from the social gaze losses, showing improved gaze following performance and social gaze prediction (Ours and Ours-PP, Table 2). In particular, we observe significant gains for the SA compared to Ours-noSoc. Interestingly, the addition of the social gaze losses also better aligns the gaze following outputs for social gaze prediction. Comparing Ours-PP and Ours-noSocial, we see that performance for all social gaze tasks is improved.

Impact of Gaze Following Loss. We additionally train our model without the standard gaze following losses: heatmap, gaze vector and in-out (Ours-noGF, Table 2). Across VSGaze, we see that

https://doi.org/10.52202/079017-0500

performance for all social gaze tasks drops, which indicates that the gaze following and social gaze losses provide complementary information, and using both can give improved performance.

Impact of VSGaze. When comparing the performance of the models trained on VSGaze (Supp Table 4) against their versions fine-tuned on individual datasets (Table 3), we see that the fine-tuned models always perform better. This is because fine-tuning allows the models to learn dataset specific priors (ex. more LAH cases in VAT, Table 1). For instance, on VAT, Gupta* has a distance score of 0.138 when trained on VSGaze, compared to a score of 0.134 when directly fine-tuned on VAT. Also, our model has a distance score of 0.112 when trained on VSGaze, and a score of 0.105 when fine-tuned on VAT. This highlights the challenge in leveraging multiple datasets: while we may expect better performance by having more data, the different priors and statistics bring additional difficulties. Nevertheless, our model trained on VSGaze is able to achieve strong performance across all datasets.

Impact of temporal information. Comparing the static and temporal versions of our model trained on VSGaze, we observe improvements in performance for shared attention (0.555 vs 0.576, Tab. 5 in supplementary), and similar or slightly improved performance for other metrics. This is in contrast with Chong$_T$, which often has lower performance than Chong$_S$ (for distance, LAH and LAEO metrics in Table 2). These results follow observations from prior work (Section 2) and highlight the challenge in leveraging temporal information for gaze following. We provide a detailed analysis in supplementary D.

The supplementary (overview in A) provides additional ablations and discussions.

6 Conclusion

We propose a new framework for multi-person, temporal gaze following and social gaze prediction comprising of a novel architecture and dataset. Through a series of experiments, we show that our model can effectively learn from a mix of video-based datasets with different statistics, to perform gaze following and social gaze prediction without sacrificing performance on any of them. The trained model can then be further fine-tuned on individual datasets to improve performance towards a specific scenario or task.

We hope that our proposed framework opens new directions for modeling people's gaze behavior, and are keen to see applications of the proposed dataset and social gaze losses in other methods. In the future, we intend to further investigate the benefits of temporal information and other auxiliary signals, including new ways of incorporating them into the architecture. We also plan to expand the VSGaze dataset to include more samples and annotations.

Acknowledgement. This research has been supported by the AI4Autism project (Digital Phenotyping of Autism Spectrum Disorders in Children, grant agreement number CRSII5 202235/1) of the the Sinergia interdisciplinary program of the SNSF.

References

[1] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)

[2] Association., A.P.: Diagnostic and statistical manual of mental disorders (5th ed.) (2013), https://doi.org/10.1176/appi.books.9780890425596

[3] Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: Multimae: Multi-modal multi-task masked autoencoders. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII. pp. 348–367. Springer (2022)

[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. pp. 14126–14135 (2022)

[5] Cantarini, G., Tomenotti, F.F., Noceti, N., Odone, F.: Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty (2021)

[6] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 213–229. Springer (2020)

[7] Chang, F., Zeng, J., Liu, Q., Shan, S.: Gaze pattern recognition in dyadic communication. In: Proceedings of the 2023 Symposium on Eye Tracking Research and Applications. pp. 1–7 (2023)

[8] Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. In: The Eleventh International Conference on Learning Representations (2022)

[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)

[10] Ci, Y., Wang, Y., Chen, M., Tang, S., Bai, L., Zhu, F., Zhao, R., Yu, F., Qi, D., Ouyang, W.: Unihcp: A unified model for human-centric perceptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17840–17852 (2023)

[11] Doosti, B., Chen, C.H., Vemulapalli, R., Jia, X., Zhu, Y., Green, B.: Boosting image-based mutual gaze detection using pseudo 3d gaze. Proceedings of the AAAI Conference on Artificial Intelligence **35**(2), 1273–1281 (May 2021). https://doi.org/10.1609/aaai.v35i2.16215, `https://ojs.aaai.org/index.php/AAAI/article/view/16215`

[12] Fan, L., Chen, Y., Wei, P., Wang, W., Zhu, S.C.: Inferring shared attention in social scene videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6460–6468 (2018)

[13] Fan, L., Wang, W., Huang, S., Tang, X., Zhu, S.C.: Understanding human gaze communication by spatio-temporal graph reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

[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 (CVPR). pp. 11390–11399 (June 2021)

[15] Guo, H., Hu, Z., Liu, J.: Mgtr: End-to-end mutual gaze detection with transformer. In: Proceedings of the Asian Conference on Computer Vision. pp. 1590–1605 (2022)

[16] Gupta, A., Tafasca, S., Chutisilp, N., Odobez, J.M.: A unified model for gaze following and social gaze prediction. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG) (2024)

[17] 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. pp. 5041–5050 (2022)

[18] Gupta, A., Vuillecard, P., Farkhondeh, A., Odobez, J.M.: Exploring the zero-shot capabilities of vision-language models for improving gaze following. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 615–624 (2024)

[19] 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 (2022)

[20] Hu, Z., Zhao, K., Zhou, B., Guo, H., Wu, S., Yang, Y., Liu, J.: Gaze target estimation inspired by interactive attention. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8524–8536 (2022)

[21] Hu, Z., Zhao, K., Zhou, B., Guo, H., Wu, S., Yang, Y., Liu, J.: Gaze target estimation inspired by interactive attention. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8524–8536 (2022)

[22] 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)

[23] 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)

[24] Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Michael, K., Fang, J., imyhxy, Lorna, Wong, C., Yifu, Z., V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, UnglvKitDe, tkianai, yxNONG, Skalski, P., Hogan, A., Strobel, M., Jain, M., Mammana, L., xylieong: ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations (Aug 2022). https://doi.org/10.5281/zenodo.7002879, `https://doi.org/10.5281/zenodo.7002879`

[25] Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., , Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: IEEE International Conference on Computer Vision (ICCV) (October 2019)

[26] 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)

[27] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)

[28] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)

[29] Maenner, M., et al.: Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, united states, 2020. Morbidity and mortality weekly report. CDC surveillance summaries / Centers for Disease Control **72**(2) (2023)

[30] Marín-Jiménez, M.J., Kalogeiton, V., Medina-Suárez, P., , Zisserman, A.: LAEO-Net++: revisiting people Looking At Each Other in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). https://doi.org/10.1109/TPAMI.2020.3048482

[31] Marin-Jimenez, M.J., Kalogeiton, V., Medina-Suarez, P., Zisserman, A.: Laeo-net: Revisiting people looking at each other in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

[32] 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)

[33] Min, K., Roy, S., Tripathi, S., Guha, T., Majumdar, S.: Learning long-term spatial-temporal graphs for active speaker detection. In: European Conference on Computer Vision. pp. 371–387. Springer (2022)

[34] Nonaka, S., Nobuhara, S., Nishino, K.: Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2192–2201 (June 2022)

[35] Otsuka, K., Takemae, Y., Yamato, J., Murase, H.: A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In: Inter. Conf. on Multimodal Interfaces. pp. 191–198 (2005)

[36] Otsuka, K., Kasuga, K., Kohler, M.: Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 191–199 (2018)

[37] Palmero, C., Selva, J., Bagheri, M.A., Escalera, S.: Recurrent CNN for 3d gaze estimation using appearance and shape cues. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. p. 251. BMVA Press (2018)

[38] Park, S., Aksan, E., Zhang, X., Hilliges, O.: Towards end-to-end video-based eye-tracking. In: European Conference on Computer Vision (ECCV) (2020)

[39] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021)

[40] Recasens*, A., Khosla*, A., Vondrick, C., Torralba, A.: Where are they looking? In: Advances in Neural Information Processing Systems (NIPS) (2015), * indicates equal contribution

[41] Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze in video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1435–1443 (2017)

[42] Senju, A., Johnson, M.H.: Atypical eye contact in autism: models, mechanisms and development. Neuroscience & Biobehavioral Reviews **33**(8), 1204–1214 (2009)

[43] Sheikhi, S., Odobez, J.M.: Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions. Pattern Recognition Letters **66**, 81–90 (2015)

[44] Shepherd, S.V.: Following gaze: gaze-following behavior as a window into social cognition. Frontiers in integrative neuroscience **4**, 5 (2010)

[45] Stiefelhagen, R., Yang, J., Waibel, A.: Modeling focus of attention for meeting indexing based on multiple cues. IEEE Trans. Neural Networks **13**(4), 928–938 (2002). https://doi.org/10.1109/TNN.2002.1021893, https://doi.org/10.1109/TNN.2002.1021893

[46] Sumer, O., Gerjets, P., Trautwein, U., Kasneci, E.: Attention flow: End-to-end joint attention estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3327–3336 (2020)

[47] Tafasca, S., Gupta, A., Kojovic, N., Gelsomini, M., Maillart, T., Papandrea, M., Schaer, M., Odobez, J.M.: The ai4autism project: A multimodal and interdisciplinary approach to autism diagnosis and stratification. In: Companion Publication of the 25th International Conference on Multimodal Interaction. pp. 414–425 (2023)

[48] 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)

[49] Tafasca, S., Gupta, A., Odobez, J.M.: Sharingan: A transformer architecture for multi-person gaze following. In: Int. Conference on Computer Vision and Pattern Recognition (CVPR) (2024)

[50] 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)

[51] 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 (ICCV). pp. 21860–21869 (October 2023)

[52] 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)

[53] Vuillecard, P., Farkhondeh, A., Villamizar, M., Odobez, J.M.: Ccdb-hg: Novel annotations and gaze-aware representations for head gesture recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG) (2024)

[54] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box (2022)

A Supplementary Material

The supplementary material is organized as follows:

- B: details on the construction of VSGaze;
- C: breakdown of performance on VSGaze by component dataset;
- D: ablations on temporal window length and novel architecture components;
- E: implementation details;
- F: experiments on incorporating auxiliary information (speaking status);
- G: qualitative analysis of predictions;
- H: details on our DPT based gaze heatmap decoder;
- I: discussion on limitations;
- J: discussion on broader impact.

Unless specified, all experiments and visualizations in the supplementary use LAH and LAEO predictions obtained via the post-processing strategy, and SA predictions from the corresponding decoder. This is because post-processing gaze following outputs for LAH and LAEO ensures that outputs for these three tasks align. We also observe slightly better performance for LAH and LAEO using the post-processing approach compared to using the predictions from their respective decoders (Table 2, Ours-PP vs Ours).

B VSGaze Construction Details

We provide the detailed protocol for obtaining gaze following and social gaze annotations on VSGaze:

Gaze Target Point. For people sharing attention in VideoCoAtt (12), we compute their gaze points as the center of the SA object's bounding box. Similarly, for a person pair LAEO in UCO-LAEO (31), we compute their gaze points as the center of the other person's head bounding box.

LAH. We generate LAH annotations for all datasets. To do so, similarly to Tafasca *et al*. (48), we check whether the gaze point for an annotated person falls inside any other person's head bounding box. For the GazeFollow test set, at least 2 of the annotated gaze points should fall inside another person's head bounding box.

LAEO. We use the provided annotations for UCO-LAEO. For VAT and ChildPlay, we generate LAEO annotations by using the LAH annotations, checking whether the LAH target for a pair of people corresponds to the other person. We cannot obtain LAEO for GazeFollow as most images are annotated for a single person, and for VideoCoAtt because if a person is an SA target, they are not in the set of people sharing attention and their gaze is not annotated.

SA. We use the provided annotations for VideoCoAtt. For VAT and ChildPlay, we generate novel SA annotations from the LAH annotations, checking whether two person share their attention to the same third person. We cannot obtain SA for GazeFollow as most images are annotated for a single person, and for UCO-LAEO as a pair of people annotated with LAEO cannot be sharing attention.

C Breakdown of Results on VSGaze

We provide a breakdown of results on VSGaze by component dataset in Table 4. Note that following the results in the main paper, LAH and LAEO results for Ours-noGF and Ours are obtained from their respective decoders. We can observe that performance trends on individual datasets can differ from the aggregated results on VSGaze. For instance, although we have a small improvement for LAH compared to the baselines across VSGaze, we perform significantly better on ChildPlay. In general, as VideoCoAtt represents the highest number of samples in VSGaze, it also has the highest impact.

Also on ChildPlay, once again we see that better gaze following performance does not translate to better social gaze performance. Although Gupta* has a better distance score compared to Chong$_S$*, it performs significantly worse for all social gaze tasks.

D Ablations

D.1 Temporal Window Length

We compare performance of our model for different temporal window lengths on VSGaze in Table 5. Note that $T = 1$ corresponds to a static model. We observe that incorporating temporal information

Dataset	Method	PP	Dist. ↓	AP_{IO} ↑	$F1_{LAH}$ ↑	$F1_{LAEO}$ ↑	AP_{SA} ↑
VAT (9)	$Chong_S$* (9)	✓	0.132	0.798	0.785	0.486	0.288
	$Chong_T$* (9)	✓	0.137	0.843	0.783	0.479	0.332
	Gupta* (17)	✓	0.138	0.795	0.766	0.518	0.300
	Ours-noGF	✗	-	-	0.766	0.503	0.435
	Ours-noSoc	✓	0.121	**0.847**	0.812	**0.557**	0.440
	Ours	✗	**0.112**	0.845	0.791	0.526	**0.521**
	Ours-PP	✓	**0.112**	0.845	**0.825**	0.548	0.497
ChildPlay (48)	$Chong_S$* (9)	✓	0.123	0.973	0.597	0.470	0.154
	$Chong_T$* (9)	✓	0.137	0.985	0.572	0.416	0.165
	Gupta* (17)	✓	0.119	0.979	0.571	0.428	0.132
	Ours-noGF	✗	-	-	0.609	0.404	0.207
	Ours-noSoc	✓	0.118	**0.994**	0.620	0.412	0.188
	Ours	✗	**0.113**	0.993	**0.682**	0.426	0.179
	Ours-PP	✓	**0.113**	0.993	0.651	**0.436**	**0.216**
VideoCoAtt (12)	$Chong_S$* (9)	✓	0.120	-	0.793	-	0.290
	$Chong_T$* (9)	✓	0.126	-	0.790	-	0.337
	Gupta* (17)	✓	0.115	-	0.815	-	0.347
	Ours-noGF	✗	-	-	0.733	-	0.524
	Ours-noSoc	✓	0.107	-	0.822	-	0.335
	Ours	✗	**0.106**	-	0.804	-	**0.601**
	Ours-PP	✓	**0.106**	-	**0.825**	-	0.345
UCO-LAEO (31)	$Chong_S$* (9)	✓	0.031	-	0.986	0.811	-
	$Chong_T$* (9)	✓	0.064	-	0.941	0.774	-
	Gupta* (17)	✓	0.031	-	0.989	0.859	
	Ours-noGF	✗	-	-	0.989	**0.939**	-
	Ours-noSoc	✓	0.043	-	0.978	0.840	-
	Ours	✗	**0.027**	-	0.990	0.888	-
	Ours-PP	✓	**0.027**	-	**0.994**	0.870	-
VSGaze	$Chong_S$* (9)	✓	0.121	0.918	0.778	0.562	0.288
	$Chong_T$* (9)	✓	0.130	**0.956**	0.764	0.529	0.331
	Gupta* (17)	✓	0.119	0.929	0.784	0.590	0.335
	Ours-noGF	✗	-	-	0.738	0.579	0.515
	Ours-noSoc	✓	0.111	0.945	0.802	0.598	0.339
	Ours	✗	**0.107**	0.940	0.795	0.590	**0.576**
	Ours-PP	✓	**0.107**	0.940	**0.812**	**0.603**	0.352

Table 4: Comparison against gaze following methods on VSGaze and its component datasets: VAT (9), ChildPlay (48), VideoCoAtt (12) and UCO-LAEO (31). All models were trained on VSGaze. PP indicates social gaze predictions from post-processing gaze following outputs (✓) vs predictions from decoders (✗). Best results are in bold, second best results are underlined.

can improve performance, especially in the case of shared attention. For the other metrics performance remains comparable. As a temporal window of 9 does not necessarily give better performance than a temporal window of 5, we use $T = 5$ for our experiments.

We note that these observations are in contrast to those from the static ($Chong_S$) and temporal models ($Chong_T$) of (9). As seen in Table 2, $Chong_T$ often performs worse than $Chong_S$, with especially lower scores for the distance and LAEO metrics. This follows prior observations from the state of the art regarding temporal modelling for gaze following (Section 2), and illustrates the challenge in leveraging temporal information.

While architecture design may be a reason for the lack of greater improvement in performance, the data itself is an important factor. Firstly, despite the larger number of samples in VSGaze compared to standard video based gaze datasets, there is high redundancy between frames so data diversity is not comparable to that of GazeFollow. Secondly, the moments where temporal information is important, such as during gaze shifts, only form a small percentage of total instances. Hence, improvements for predictions in these moments are not reflected in overall metrics. For instance, gaze shifts form less than 10% of total instances in ChildPlay (48). However, in our qualitative analysis (Section G) we can see situations where temporal information helps. In future work we plan to investigate new metrics for evaluating the performance of temporal models.

D.2 Architecture Components

We systematically remove different novel components of our architecture to analyse their impact and provide results in Table 6.

Interaction Module. Removing the Person-to-Scene Interaction encoder $\mathcal{I}^{\mathrm{b}}_{\mathrm{ps}}$ (Ours-noI_{ps}) has the

https://doi.org/10.52202/079017-0500

T	Dist. ↓	AP_{IO} ↑	$F1_{LAH}$ ↑	$F1_{LAEO}$ ↑	AP_{SA} ↑
1	0.108	**0.946**	0.806	0.599	0.555
5	0.107	0.940	**0.812**	**0.603**	**0.576**
9	**0.106**	0.943	0.811	0.590	0.563

Table 5: Ablations for different temporal window lengths T on VSGaze. Best results are in bold, second best results are underlined.

Method	Dist. ↓	AP_{IO} ↑	$F1_{LAH}$ ↑	$F1_{LAEO}$ ↑	AP_{SA} ↑
Ours-noI_{ppt}	0.108	0.937	0.806	**0.612**	0.547
Ours-noI_{sp}	**0.105**	0.936	**0.813**	0.581	0.545
Ours-noI_{ps}	0.112	0.939	0.799	0.605	0.538
Ours-noDPT	0.111	**0.941**	0.810	0.587	0.528
Ours	0.107	0.940	0.812	0.603	**0.576**

Table 6: Ablations on different novel components of our architecture on VSGaze. I_{ppt} refers to the Spatio-Temporal Social Interaction component, I_{sp} refers to the Scene-to-Person encoder, I_{ps} refers to the Person-to-Scene encoder and DPT refers to the gaze heatmap decoder. Best results are in bold, second best results are underlined.

largest impact on performance, especially for distance, LAH and SA. Without this encoder, the frame tokens cannot access the person tokens, so encoding their gaze relevant salient items is much harder. Removing the Scene-to-Person Interaction encoder $\mathcal{I}^{b}_{sp}$ (Ours-noI_{sp}) decreases LAEO and SA performance. Without this encoder, the person tokens cannot access the frame tokens, so they cannot capture the locations of gazed at salient items. As the frame tokens may be able to adapt to this change, gaze following performance is not impacted negatively. Finally, removing the Spatio-Temporal Social Interaction component $\mathcal{I}^{b}_{pp}$, $\mathcal{I}^{b}_{pt}$ (Ours-noI_{ppt}) decreases LAH and SA performance. Without this component, there is no interaction between person tokens, so identification of social dynamics is hindered. Interestingly, we see a boost in LAEO performance. However, a major portion of LAEO positives come from the UCO-LAEO dataset (see Table 1) which consists mainly of two person scenes, so capturing social interactions may be less important.

DPT Decoder. We replace our proposed modified DPT decoder for gaze heatmap prediction (Section H) with a simpler decoder (Ours-noDPT). This decoder projects the frame and person tokens from the last Interaction block, performs a dot product between them, and then upscales the output to the heatmap resolution. Using the simpler decoder results in drops in performance for the distance, LAEO and SA metrics. Unlike the DPT, it lacks multi-scale representations which impacts heatmap prediction and supervision of tokens.

Overall, we observe that removing the components tends to impact shared attention performance. This is similar to the observation in the previous section regarding temporal information. Unlike LAH and LAEO where the target is always another person, SA is more challenging as the shared attention target can be any person or object/point. Hence, this task may be benefitting more from additional information or architecture components.

E Implementation Details

The Gaze Backbone $\mathcal{G}_{\text{stat}}$ is a ResNet18 pre-trained on Gaze360 (25), and processes the head crops at a resolution of 224×224. We use a ViT-base model (1) $\mathcal{V}$ initialized with MultiMAE weights (3) to process the scene at 224×224. The temporal position embedding $\mathbf{x}$ is zero-initialized, while all other layers are randomly initialized. For training, we use the AdamW optimizer (28) with warmup and cosine annealing. The loss coefficients are set as $\lambda_{\text{HM}} = 1000, \lambda_{\text{VEC}} = 3, \lambda_{\text{IO}} = 2$ and $\lambda_{\text{LAH}} = \lambda_{\text{SA}} = 1$. To counter the class imbalance in positive vs negative social gaze labels, we weight positive samples for the social gaze loss by 2.

All models were trained on an internal cluster using a single Nvidia RTX 3090 GPU with 24GB memory. Training time for a single experiment is approximately 8 hours on both GazeFollow and VSGaze. Total compute across all experiments is approximately 140 GPU hours.

F Incorporating Auxiliary Information

Previous studies on analyzing conversations during meetings have shown that people usually look at the other speaking participants (45), and such cues can be exploited for gaze target selection

Dataset	Method	Dist. ↓	AP_{IO} ↑	$F1_{LAH}$ ↑	$F1_{LAEO}$ ↑	AP_{SA} ↑
VSGaze	Ours-spk	**0.107**	0.938	0.810	0.588	**0.590**
	Ours	**0.107**	**0.940**	**0.812**	**0.603**	0.576

Table 7: Performance for incorporating people's speaking information in our model. Best results are in bold.

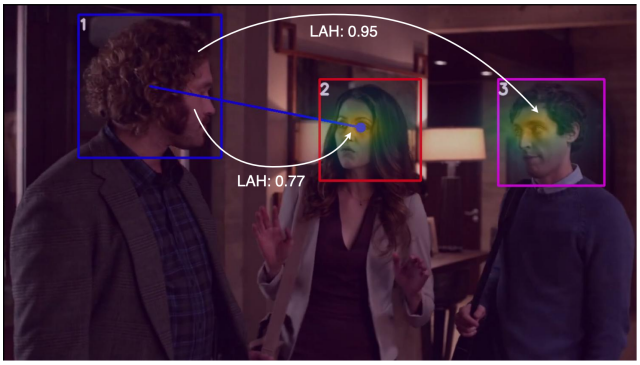


Figure 3: An illustration of the few cases where the predicted gaze point does not match with the predicted LAH label. The uncertainty in the gaze target is reflected in the heatmap, while the uncertainty in the LAH target is reflected in the LAH scores.

(35). Hence, we expect that identifying speaking persons can provide better scene understanding for gaze following, and help recognize attentiveness towards people, especially speakers. The latter is especially important in autism diagnosis, as eye contact is closely monitored by the clinician when they call out to the tested child (2).

F.1 Experiments and Results

To incorporate speaking information in our model, we adapt the Person Module (Section 3). Specifically, we first obtain speaking scores for each person using an active speaker detection model (33). The model is retrained to detect speakers using the video alone and with no audio. The obtained speaking scores $\mathbf{s}_{i,t}$ for each person are linearly projected to the token dimension using $\mathcal{P}_{\text{spk}}$, and added to the person token. Hence, the new person token is obtained as:

$$\mathbf{p}_{i,t} = \mathcal{P}_{\text{gaze}}(\mathbf{g}_{i,t}^{\text{temp}}) + \mathcal{P}_{\text{spk}}(\mathbf{s}_{i,t}) + \mathcal{P}_{\text{box}}(\mathbf{h}_{i,t}^{\text{box}}). \tag{5}$$

We note that this formulation can easily be extended to incorporate other kinds of person-specific auxiliary information such as gestures. We explore this aspect in a follow up work where we extract gestures and other cues using vision-language models (18).

We provide results for incorporating speaking information in Table 7. On VSGaze, once again we see improvements for SA while other metrics remain similar. As the speaker detection model tends to fail in cases with side-view faces or for children in the case of ChildPlay, we may further benefit from a better speaker detection model. In addition, investigating other ways to better capture and incorporate such auxiliary information in our model is a direction for future research.

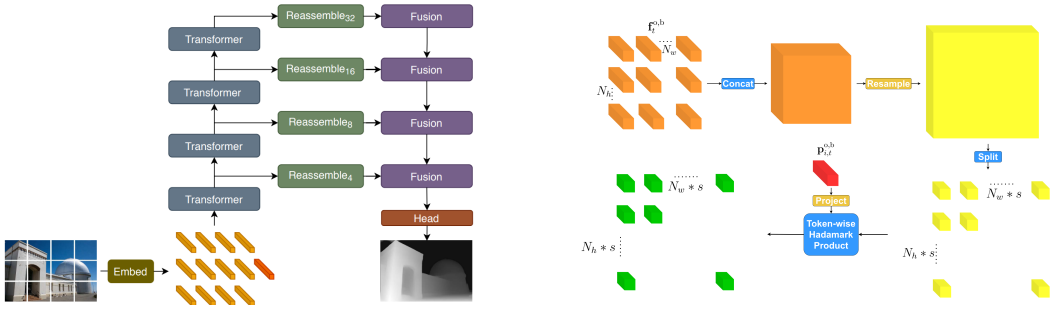
G Qualitative Analysis

G.1 Qualitative Results and Comparisons

We provide qualitative results from our models in Figure 5. We observe that all models perform well, accurately capturing people's gaze targets and social gaze behaviour. We also see that incorporating temporal and speaking information can help improve predictions.

In the first sequence, the static model occasionally makes an error for person 1, and picks person 2's hands as the predicted gaze target. This is because it cannot distinguish blinking from when a person lowers their gaze. On the other hand, the temporal models recognize blinking and maintain the target as person 2's face. In the second sequence, the static model misses shared attention between persons 1,2,3 in the first frame and persons 2,3 in the third frame. This sequence is challenging due to the presence of subtle head motions which the temporal models can better capture. In the third sequence,

https://doi.org/10.52202/079017-0500



(a) (b)

Figure 4: The standard DPT (a, taken from (39)) and our proposed person-conditioned re-assemble stage (b). This transformed DPT is used for predicting gaze heatmaps for each person in the scene.

both the static model and our proposed model make an error when predicting person 3's gaze target in frame 2. However, our model with speaking information recognizes person 1 as the target given their high speaking score.

In addition, we provide qualitative comparisons of our model against other methods in Figure 6. We see that our model performs better overall, accurately inferring people's gaze target and social gaze behaviour despite the complexity of the scenes with multiple salient targets, obscured eyes, varied settings and age groups.

G.2 Alignment Between Gaze Following Outputs and Social Gaze Decoders

We perform analysis to understand the difference in performance between post-processing gaze following outputs or using the predictions from the task specific decoders. For LAH, we find that predictions for the two schemes align 87% of the time. On visualizing outputs, we identify that cases where they don't match are usually where the model is confused between two potential targets. So the arg max of the gaze heatmap for obtaining the gaze point picks one target, while the arg max of the LAH scores picks the other target. This confusion in target selection is illustrated for person 1 in Figure 3. We see that the predicted gaze heatmap highlights both person 2 and 3, and similarly, the LAH scores for looking at both persons 2 and 3 are high.

H Gaze Heatmap Prediction Details

As mentioned in Section 3, we rely on the standard DPT (39) decoder that has been developed for dense prediction, and propose an interesting way to transform it for performing person-conditioned gaze heatmap prediction.

Similar to an FPN (27), the DPT assembles the set of ViT image tokens into image-like feature representations at various resolutions. The feature representations are then progressively fused into the final dense prediction. Specifically, the DPT decoder contains two stages: 1) A *Re-assemble* stage to construct feature maps at specific resolutions at each block, and 2) a *Fusion* stage where the feature maps across consecutive blocks are upscaled and combined.

To include the person specific information, as represented by the person token in our architecture, we modify the re-assemble stage to filter only the information relevant for the given person (Figure 4). More precisely, following the standard DPT, we first project the input frame tokens $\mathbf{f}_t^{\text{o,b}}$ at level b to a lower token dimension using a 1×1 conv layer, followed by spatial upsampling or downsampling using a strided 3×3 transposed conv layer or conv layer respectively;

$$\mathbf{f}_t^{\text{DPT,b}} = \text{Split}(\text{Resample}^b(\text{Concat}(\mathbf{f}_t^{\text{o,b}}))) \tag{6}$$

where Split is the reverse of the spatial concatenation operation, Concat, and Resample is the spatial upsampling/downsampling operation. To condition on a person, we then perform a token-wise hadamard product of the projected frame tokens with the the projected person token (using projection layer $\mathcal{P}^b_{\text{DPT}}$) from the corresponding block b.

$$\mathbf{f}_t^{\text{DPT-c,b}} = \mathbf{f}_t^{\text{DPT,b}} * \mathcal{P}^b_{\text{DPT}}(\mathbf{p}_{i,t}^{\text{o,b}}) \tag{7}$$

where $*$ denotes the hadamard product. The standard Fusion stage then follows to obtain the predicted gaze heatmap. Note that superscripts DPT, DPT-c denote updates to the frame tokens.

https://doi.org/10.52202/079017-0500

I Limitations

As discussed in Supp. G, gaze following outputs and predictions from the social gaze decoders do not always align (13% cases). One possible solution is to post-process gaze following outputs for LAH and LAEO as done in the case of Ours-PP. This ensures that outputs for gaze following, LAH and LAEO align. However, further aligning the outputs with SA is more challenging as post-processing for SA does not account for whether the gaze points fall on the same semantic item. Ensuring that outputs for all social gaze tasks and gaze following align is a challenging and interesting direction for future research. Potential directions include more refined post-processing techniques, or the addition of task consistency losses and regularization. The latter is particularly interesting as it may further increase the benefit of the social gaze losses.

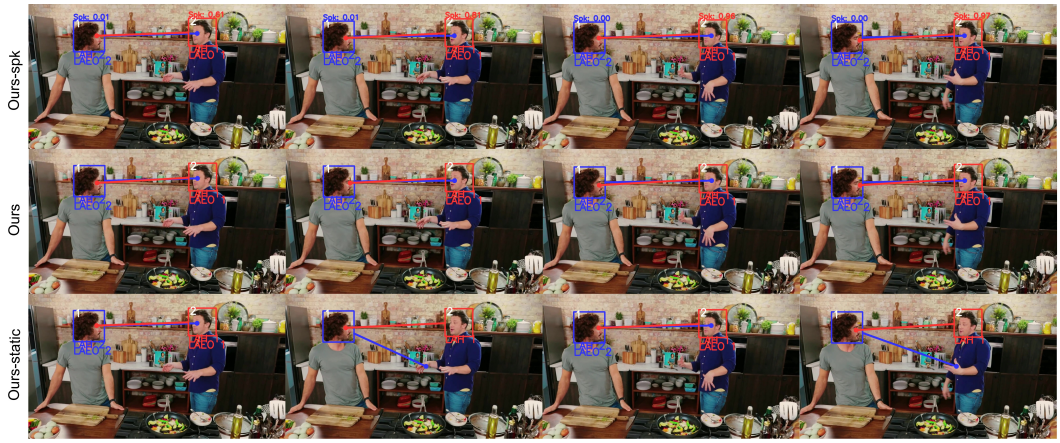
Secondly, as discussed in supplementary D, temporal information does not seem to bring large improvements in performance, with improvements observed mainly for SA. Investigating new architectures, datasets and metrics to improve training and evaluation of temporal information is another important and interesting direction for future research.

J Broader Impact

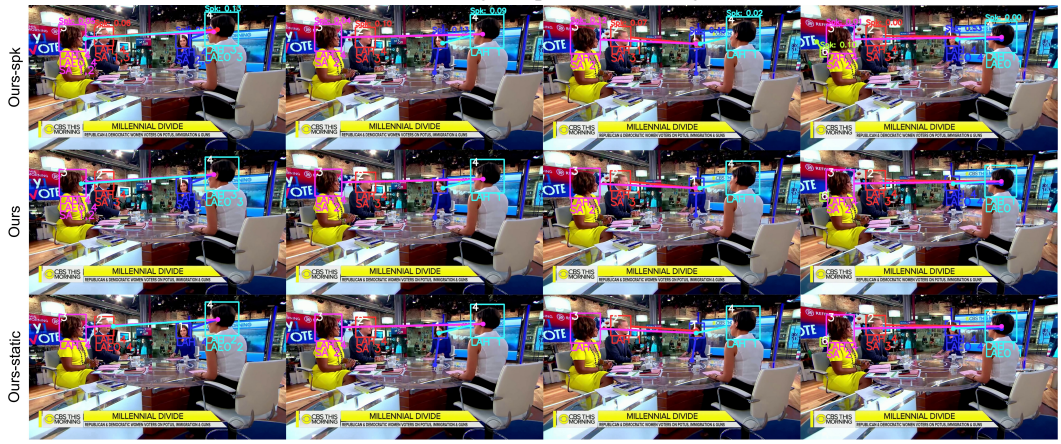
Gaze following and social gaze prediction has strong potential for positive societal impact. Indeed, eye-contact and shared attention are monitored during the ADOS autism diagnostic test for children (2) as a marker for social communication. Given the time consuming nature of clinical tests and large scale prevalence of autism (1 in 36 children (29)), automatic tools for autism screening based on gaze can help flag potential autism cases and reduce the burden on doctors (47).

Nevertheless, as with any medical application, such screening tools must be used carefully to avoid missing cases, for instance through human-in-the-loop systems. It is also important to ensure that gaze algorithms do not invade the privacy of people. They should be deployed with appropriate consent and only for specific applications.

https://doi.org/10.52202/079017-0500



(a) Ours-static fails to recognise person 1 blinking in frames 2,4



(b) Ours-static misses shared attention behaviour in frames 1,3

(c) Ours-spk captures the right target for person 3 in frame 2

Figure 5: Qualitative results of our proposed model (Ours), our model with speaking information (Ours-spk) and our model without temporal information (Ours-static). When the target is predicted to be inside the frame, we display the predicted gaze point and the social gaze tasks with the associated person id(s).

https://doi.org/10.52202/079017-0500

Figure 6: Qualitative comparison of our model against other methods Chong$_S$, Chong$_T$ (9), Gupta (17). Our model performs better overall, outperforming other methods in complex scenes with obscured eyes, multiple salient targets, varied settings and age groups.

Figure 7: Samples from VACATION. Person 3 in both cases is missed as the associated person is already annotated with a social gaze 'state'.

Figure 8: Results from Tonini et al. (2023) demonstrating incorrect head detections. Annotations are in white.

https://doi.org/10.52202/079017-0500

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim three major contributions: a novel architecture described in Section 3, a new dataset described in Section 4.1.1, and new social gaze protocols and metrics described in Section 4.3. Our experimental results are provided in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have provided a discussion on limitations in Supplementary I.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe our architecture in Section 3, and our dataset construction in Section 4.1.1 with further details in Supplementary B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-0500

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the data and code upon cleanup.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and test details including data splits, hyperparameters and type of optimizer are provided in Section 4.2 with additional details in Supplementary E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We are unable to compute error bars due to constraints in terms of access to computational resources and training time.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details on computer resources in Supplementary E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and observed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Potential positive and negative societal impacts of the work have been discussed in Supplementary J.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-0500

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not think our data and models are high risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original papers for leveraged datasets, code and models have been cited. Appropriate licenses will be included when releasing code and data.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No assets have been currently released with the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

https://doi.org/10.52202/079017-0500

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.