Sharingan: A Transformer Architecture for Multi-Person Gaze Following

Samy Tafasca  Anshul Gupta  Jean-Marc Odobez
Idiap Research Institute, Switzerland
École Polytechnique Fédérale de Lausanne, Switzerland
{stafasca, agupta, odobez}@idiap.ch

Abstract

Gaze is a powerful form of non-verbal communication that humans develop from an early age. As such, modeling this behavior is an important task that can benefit a broad set of application domains ranging from robotics to sociology. In particular, the gaze following task in computer vision is defined as the prediction of the 2D pixel coordinates where a person in the image is looking. Previous attempts in this area have primarily centered on CNN-based architectures, but they have been constrained by the need to process one person at a time, which proves to be highly inefficient. In this paper, we introduce a novel and effective multi-person transformer-based architecture for gaze prediction. While there exist prior works using transformers for multi-person gaze prediction [38, 39], they use a fixed set of learnable embeddings to decode both the person and its gaze target, which requires a matching step afterward to link the predictions with the annotations. Thus, it is difficult to quantitatively evaluate these methods reliably with the available benchmarks, or integrate them into a larger human behavior understanding system. Instead, we are the first to propose a multi-person transformer-based architecture that maintains the original task formulation and ensures control over the people fed as input. Our main contribution lies in encoding the person-specific information into a single controlled token to be processed alongside image tokens and using its output for prediction based on a novel multiscale decoding mechanism. Our new architecture achieves state-of-the-art results on the GazeFollow, VideoAttentionTarget, and ChildPlay datasets and outperforms comparable multi-person architectures with a notable margin. Our code, checkpoints, and data extractions will be made publicly available soon.

1. Introduction

Gaze is an important form of communication and was extensively studied across different domains and applications such as consumer behavior understanding [4, 19, 36], soci-
In this work, we aim to tackle these challenges by addressing the multi-person gaze-following task while maintaining the original problem formulation. To this end, we propose Sharingan, a novel, effective, and efficient transformer-based architecture to predict the gaze target of multiple people simultaneously. A key component of this architecture is to represent the person’s gaze information by a single gaze token produced by a gaze backbone and processed alongside the image tokens. This is in stark contrast to previous methods that represent intermediate person-specific gaze information as a visual attention map [7] or gaze cone [11, 12, 21, 35]. We show in our ablations that this is not only unnecessary but can also hinder performance in the context of a transformer. Furthermore, we introduce Conditional DPT: a more sophisticated lightweight multi-scale gaze decoding mechanism that helps improve performance by providing a finer-grained understanding of the scene for gaze target selection. This also has the benefit of producing heatmaps that better capture uncertainty when it is difficult to decide where a person is looking (cf. qualitative results in the supplementary material).

Through extensive ablations and evaluations, we find that Sharingan achieves good performance on all public benchmarks, and even transfers well to other gaze-related tasks such as shared attention and mutual gaze.

2. Related Work

In this section, we present several relevant research topics.

Gaze Following. The task of gaze following was first introduced in the seminal work of Recasens et al. [29]. The idea is to predict the pixel-wise 2D location in the image corresponding to where a target person is looking. The main advantage of this formulation is the lack of constraints which allows methods trained this way to generalize to arbitrary settings (i.e., scene properties, camera parameters, image conditions, etc.). It was later extended by Chong et al. [7] to also include the prediction of whether the given person is looking inside the image frame or somewhere outside.

Traditional methods for gaze following [7, 11, 12, 16, 17, 21, 29] typically rely on convolutional networks and follow a 2-tower architecture. The first branch processes the scene image to highlight salient regions, while the second branch processes the head crop of the target person to infer a general gaze direction. A fusion mechanism then combines information from both parts to produce the final prediction.

The gaze following task is often framed as the prediction of a gaze heatmap where pixels with high intensity represent spatial areas with higher prediction confidence. We devote a section later to discuss the alternative formulation of regressing the 2D location directly (cf. Section 5).

Multi-Person Gaze Following. A major downside of the traditional formulation of gaze following is the need for multiple forward passes when predicting the gaze of different people in the same image. This problem motivated the need for architectures that can natively handle the prediction of gaze for multiple people with a single forward
pass. Jin et al. [16] first proposed a simple convolution-based architecture to handle the multi-person setting where a scene backbone computes a fixed person-agnostic feature representation. This is then fused repetitively with head features computed from the different people using another head backbone before decoding each into its corresponding gaze heatmap. Aside from the architectural differences and limited performance, one of the main drawbacks of this method is that the computation for each person is done independently from the others, which ignores potential interactions between people. Recently, Tu et al. [39] and Tonini et al. [38] proposed transformer-based architectures to perform multi-person gaze target prediction. Their methods only take the image as input and simultaneously predict both the head box and gaze target (among others) for every person in the scene. Their work is inspired by the DETR architecture [5], where the task is formulated as a set prediction problem. Instead of reinventing the wheel, our method focuses differently from the others, which ignores potential interactions between people.

3. Sharingan Architecture

Our Sharingan architecture is illustrated in Figure 3. The main idea is to use a transformer that lets scene tokens and person-specific gaze tokens interact within an attention framework to jointly predict the 2D gaze heatmap of each individual. Thus, the inputs are the image and the head crops that we assume are available. We introduce below the different components of this architecture.

3.1. Image tokens

We follow a standard ViT architecture to produce image tokens. The input scene image $I \in \mathbb{R}^{H \times W \times C}$ goes through a patch projection $P_{\text{img}}$ to produce image tokens that we equip with positional information $x^\text{img} \in \mathbb{R}^{N \times D}$, where $N$ is the number of patches, and $D$ is the token dimension.

3.2. Gaze tokens

The main purpose of a gaze token is to map the gaze information of a person into a token embedded in the same space as the image tokens, which can interact with scene tokens to select the relevant content for prediction. For simplicity, we first introduce this process for a single person.

**Single Person Case.** Let $h_{\text{crop}} \in \mathbb{R}^{H \times W \times C}$ denote the head crop of a person and $h_{\text{bbox}} = (x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}}) \in [0, 1]^4$ her head bounding box. The mapping works as follows. The head crop $h_{\text{crop}}$ is fed to a gaze backbone $\mathcal{G}$ to produce a gaze embedding $g_{\text{emb}} \in \mathbb{R}^{d_{\text{emb}}}$. This embedding is used in two ways. First, it goes through a gaze regressor (i.e. MLP) $\mathcal{O}_g$ to predict a 2D gaze vector $g_v = \mathcal{O}_g(g_{\text{emb}})$. This output is supervised using an angular gaze loss.

Secondly, the gaze embedding is projected to the token dimension using a learnable linear projection $P_{\text{gaze}}$, resulting in the gaze token $x^\text{emb} = P_{\text{gaze}}(g_{\text{emb}}) \in \mathbb{R}^D$. As we want to incorporate information about the person’s location (and size), we also project the head bounding box $h_{\text{bbox}}$ into a bounding box embedding $x_{\text{bbox}}$ using a learnable linear projection $P_{\text{bbox}}$: $x_{\text{bbox}} = P_{\text{bbox}}(h_{\text{bbox}}) \in \mathbb{R}^D$. Finally, we add this embedding to the gaze token to obtain the final location-aware gaze token:

$$x' = x^\text{emb} + x_{\text{bbox}} \in \mathbb{R}^D \quad (1)$$

**Multi-person case.** When $N_p$ persons are detected, the architecture will produce a set of $N_p$ gaze tokens, following the same process described above for each person. Thus, if $h_{\text{bbox}}^i$ and $h_{\text{crop}}^i$ denote the bounding-box and head crop of person $i$, the above process will generate a gaze token $x^g_i$ for this person. To simplify notation, we will also denote by $x^t$ the set of gaze tokens of all people in the scene, with $x^t = x^t_1 \oplus \ldots \oplus x^t_{N_p}$, where $\oplus$ is the concatenation operator.

**Modality Encoding.** Given the different nature of gaze tokens compared to image tokens, we need to encode modality-specific information to distinguish between them. Rather than using an explicit scheme, in practice we expect this modality information to be captured by the bias terms of the different projection operators $P_{\text{gaze}}$ and $P_{\text{img}}$.

3.3. Transformer Encoder

The transformer encoder is a standard ViT [8]. It takes as input the concatenation of the scene tokens $x^\text{img}$, the gaze token(s) $x^g$, according to $x = x^\text{img} \oplus x^g \in \mathbb{R}^{N_t \times D}$, where $N_t = N + N_p$. The set of input tokens goes through a series of $L$ transformer blocks to obtain an output sequence of similar shape, denoted by $x^\text{out} = x(L) \in \mathbb{R}^{N_t \times D}$.

3.4. Gaze Decoder

The goal of the gaze decoder $D_{\text{gaze}}$ is to predict a set of gaze heatmaps. Our Conditional DPT (cf. Figure 4) takes four intermediate representations of the image tokens $x^\text{img}_{l(i)}$ and gaze tokens $x^g_{l(i)}$ and combines them progressively at different simulated resolutions, where lower resolutions have more channels and correspond to deeper layers of the encoder. This can be viewed as the isotropic equivalent of a Feature Pyramid Network [22].

Our design is inspired by DPT [28], which can only handle decoding image tokens alone. In our case, we need this decoding to be conditioned on each person. To this end, after each block of layers, the image tokens $x^\text{img}_{l(k)}, k \in \{4, 8, 16, 32\}$ are reassembled into an image-like representation at resolution $\left(\frac{H}{k}, \frac{W}{k}\right)$ and dimension $d_k$. The gaze to-
tokens $x^g_{(L)}$ at that layer are also projected to the same dimension. Next, we duplicate the image feature maps $N_p$ times, and apply an element-wise dot-product between each gaze token and a copy of the image feature map. Finally, these person-specific image features are stacked, and we merge the batch and person dimensions to produce a final output of dimension $(B \times N_p, d_k, H, W)$. This tensor is passed to a fusion module where it is processed by a small residual convnet and added to the output from the previous fusion block. The result goes through another residual convnet, an upsampling stage to double the resolution, and a projection, leading to a tensor of dimension $(B \times N_p, d_{out}, \frac{2H}{H}, \frac{2W}{W})$. At the end of this process, we get a tensor of dimension $(B \times N_p, d_{out}, \frac{H}{2}, \frac{W}{2})$, which goes through a convolutional head that predicts the heatmaps by bringing the channel dimension down to 1 and resizing the spatial dimension to that of the gaze heatmap. Finally, we separate the batch and person dimensions such that the final shape of the output is $(B, N_p, 1, H_{hm}, W_{hm})$. The rationale behind this design is to gather information from different layers and resolutions, which is important for dense prediction tasks. In this case, it is particularly useful for gaze tokens where information from the early layers might retain more scene-independent gaze cues due to their proximity to the gaze encoder.

3.5. In-Out prediction

The In-Out classifier head $O_{MLP}$ consists of an MLP with 7 layers. It is fed the concatenation of input and output gaze tokens to predict a binary in-vs-out label for each person.

$$o = O_{MLP}([x^g_{(L)}, x^g])$$

3.6. Loss and implementation details

We train our model using a combination of three losses:

Heatmap Loss ($L_{reg}$). The heatmap loss is the pixel-wise MSE between the GT heatmap and the predicted heatmap:

$$L_{hm} = \sum_{x,y} W_{hm} \cdot H_{hm} \cdot (A^g_{x,y} - A^p_{x,y})^2.$$

Angular Loss ($L_{ang}$). The angular loss drives the prediction of a normalized gaze direction vector. It maximizes the cosine of the angle between the predicted and ground truth gaze vectors according to: $L_{ang} = 1 - <g^g_{sv}, g^p_{sv}>$ where $<a, b>$ denotes the inner product between $a$ and $b$.

In-Out Loss ($L_{io}$). The in-out loss is the standard binary cross-entropy for in-vs-out prediction.

Global loss. The final loss is a given by:

$$L = \lambda_{reg}L_{reg} + \lambda_{ang}L_{ang} + \lambda_{io}L_{io}$$
4. Experiments

4.1. Datasets

GazeFollow. GazeFollow [30] is an image-based dataset annotated mostly with head bounding boxes, 2D gaze points. Overall, it has around 130K annotated instances in 122K images. The test set comprises 4782 gaze instances, each of which is labeled by multiple annotators (~10).

VideoAttentionTarget. VideoAttentionTarget [7] is a video-based dataset consisting of 1331 clips from 50 TV shows. It is also annotated with the head bounding boxes, 2D gaze points, and in vs out labels. Overall, it contains 164K instances in 71K frames.

ChildPlay. ChildPlay [35] is a video dataset consisting of 401 clips from 95 YouTube videos of children engaged in play activities. On top of the standard annotations, the authors extend the in vs out label to include other gaze classes (e.g. gaze shift), which we don’t use in this paper. Overall, it contains 257K instances in 120K frames.

4.2. Metrics

We use four metrics to evaluate our gaze following models. The first three are AUC, Distance, and AP which are all standard in the literature [7]. Particularly, the AUC adaptation introduced by [7] for datasets with a single point annotation is not very informative, so we decided to replace it with PLAH, which we define below.

PLAH. Recently introduced in [35], this metric computes the Precision of looking at people’s heads to incorporate semantic information in the evaluation of gaze models. Unlike [35], we consider a prediction positive if the predicted and annotated gaze points fall within the same head box.

4.3. Experimental Protocol

Context People. The training of Sharingan relies on processing multiple people at the same time (cf. section 5), but available benchmarks often annotate 1 person per image. To circumvent this problem, we apply an off-the-shelf head detector trained on the CrowdHuman dataset [31]. We discard detections with a confidence score lower than 0.5 and detections with an IOU score higher than 0.5 with the annotated target person. During training, due to batch constraints, we set $N_{p}^{tr}$ and keep it fixed. For each image, we use the person with the GT annotation and randomly sample $N_{p}^{tr} - 1$ (detected) heads when available, otherwise we use padding for the box and head crop. Incidentally, the loss is computed and propagated solely from the annotated person. At evaluation, for each image $i$ we set $N_{p}^{tr}$ to the number of all people in it and process them in a single forward pass using a batch size of 1. Note that $N_{p}^{tr}$ is a property of the training process, not the architecture, and thus doesn’t restrict the $N_{p}^{tr}$ that can be used during inference. Unless stated otherwise, we use $N_{p}^{tr} = 2$ in all experiments.

Implementation Details. Sharingan processes the input scene image and head crop at a resolution of $224 \times 224$, while the output heatmap is $64 \times 64$. The gaze backbone $G$ is a ResNet-18 [13] pretrained on Gaze360 [18], and the transformer encoder is a ViT-base model [9] initialized with weights from a multimodal MAE [2].

Training. The models are trained for 20 epochs on GazeFollow. For VideoAttentionTarget and ChildPlay, following standard practices, we take the trained GazeFollow model, freeze everything except the gaze decoder and In-Out classifier, and fine-tune them separately for 2 epochs each. We use the AdamW optimizer [23] with a learning rate of $3e-5$, and a cosine annealing schedule. The fine-tuning uses a learning rate of $1e-6$ for the gaze decoder and $3e-4$ for the In-Out classifier. We also make use of Stochastic Weight Averaging [15] to stabilize training on GazeFollow. The loss coefficients are $\lambda_{reg} = 1000$ and $\lambda_{ang} = 3$.

Validation. Since GazeFollow [30] and VideoAttentionTarget [7] do not propose any validation splits, we use the train-val splits proposed by [35]. The best model on the validation set is selected based on the distance metric.

\[^{1}\text{https://github.com/deepakcrk/yolov5-crowdhuman}\]
4.4. Comparison with the State-of-the-art

We summarize our quantitative results on the GazeFollow and VideoAttentionTarget datasets in Table 1, and on ChildPlay in Table 2, compared to previous works\(^2\). Our model sets a new state-of-the-art on all 3 datasets on most metrics and outperforms the only comparable multi-person method by 0.013 on the Avg. Dist. metric. Moreover, the only method that comes close to our results is [12], which we slightly outperform on GazeFollow in both Avg. Dist. and AUC. However, unlike Sharingan, this method uses 2 other modalities (i.e. depth, pose), has a very complex and costly training protocol (i.e. modality-specific backbones are pre-trained separately for the task), and is resource-intensive during inference (i.e. single-person). Moreover, it generalizes poorly as evidenced by a cross-dataset evaluation\(^3\) (i.e. Dist. 0.113 vs 0.134 on VideoAttentionTarget, and 0.109 vs 0.142 on ChildPlay). Incidentally, the image version of this method is also significantly worse (i.e. 0.134 vs 0.113). In terms of multi-person comparison, we beat [16] by 0.013 and 0.02 in Avg. Dist. and Min. Dist. respectively.

We also see similar results on video datasets (VideoAttentionTarget and ChildPlay), and it is interesting to note that the model trained on GazeFollow is achieving remarkable cross-dataset performance without fine-tuning. Surprisingly, it is already improving on its multi-person competitor by a large margin (i.e. Dist. 0.113 vs 0.134), which is a testament to the generalization ability of our model. We also note that the model from [35] is on par with ours on the video datasets. We believe the reason is these datasets have high-quality images compared to GazeFollow, so the depth-based geometric prior in [35] proves useful, but that method remains multimodal and single-person. We provide more details, discussions, and experiments in the supplementary material to further assess robustness and generalization.

4.5. Ablation Experiments

Person Encoding. An important aspect of Sharingan is the way people are encoded into the architecture. Previous methods often represent the head location as a binary mask, and gaze information as a visual attention map [7] or gaze cone [11, 12, 35]. We modify Sharingan to experiment with 3 variants, and opt for single-person training to alleviate the computational cost from some of these formulations: 1. Instead of projecting the box coordinates, we use a head location mask that we tokenize and add the resulting head position embeddings to the image tokens (i.e. Head Mask Embed). 2. Using the same head mask embedding, we tokenize the head crop and append the head tokens to the image ones (i.e. Head Crop Tokens). 3. Instead of processing the head directly, we use the gaze backbone to regress a gaze vector which is used to build a gaze cone that we tokenize (i.e. Gaze Cone Tokens). Please note that formulations 2 and 3 double the number of tokens, and we decode the output image tokens into a gaze heatmap using a normal DPT [28]. As we can see from Table 3 (top), the location-aware gaze token is both efficient and performs the best.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC ↑</th>
<th>Avg. D. ↓</th>
<th>Min. D. ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head Mask Embed</td>
<td>0.940</td>
<td>0.117</td>
<td>0.060</td>
</tr>
<tr>
<td>Head Crop Tokens</td>
<td>0.933</td>
<td>0.138</td>
<td>0.076</td>
</tr>
<tr>
<td>Gaze Cone Tokens</td>
<td>0.934</td>
<td>0.133</td>
<td>0.073</td>
</tr>
<tr>
<td>Gaze Token</td>
<td>0.944</td>
<td>0.113</td>
<td>0.057</td>
</tr>
<tr>
<td>Token to Heatmap</td>
<td>0.647</td>
<td>0.302</td>
<td>0.234</td>
</tr>
<tr>
<td>Dot-Product</td>
<td>0.923</td>
<td>0.120</td>
<td>0.062</td>
</tr>
<tr>
<td>Up &amp; Dot-Product</td>
<td>0.934</td>
<td>0.116</td>
<td>0.059</td>
</tr>
<tr>
<td>Conditional DPT</td>
<td>0.944</td>
<td>0.113</td>
<td>0.057</td>
</tr>
</tbody>
</table>

Table 3. Ablation results for person encoding (top) and gaze decoding (bottom).

Gaze Decoder. We also perform ablations to assess the suitability of our Conditional DPT gaze decoder by comparing it to other baselines as shown in Table 3 (bottom). The first baseline (i.e. Token to Heatmap) regresses a gaze heatmap directly from the output person token using an MLP. This is also the decoding approach undertaken by [39] and [38]. The second baseline (cf. Dot-Product) projects the tokens, then performs a dot-product between each person token and each image token before resizing the output to \(64 \times 64\) to get the final heatmap. The third baseline (cf. Up & Dot-Product) upscales the image representation first, then performs the dot-product. We note that our Conditional DPT outperforms all the other methods, justifying the need for a more sophisticated decoding mechanism.

Angular Loss. Our experiments show that the angular loss doesn’t affect the final performance, but we decided to keep it anyway to always have a reliable gaze direction, even when the person is looking outside the frame.

5. Discussion

Model Efficiency. Unlike most previous methods, Sharingan’s ability to predict the gaze of multiple people at the same time makes it very efficient for real-world applications. Figure 5 (left) shows a fairly large improvement in inference time compared to a lightweight baseline when \(N_p\) increases. It’s worth noting that while Sharingan does most of its processing once within the transformer, the small gaze encoder and decoder themselves are executed for each person independently by combining the batch and person dimensions. This explains why the curve is not constant.

\(^2\)We omit [38, 39] from the table because their evaluation protocol is different, which makes them incomparable to the rest of the methods.

\(^3\)Results are taken from [35].
Table 1. Results of our Sharingan architecture on the GazeFollow and VideoAttentionTarget datasets. The best scores for multi-person models are given in bold, while the best scores in general are underlined. The † symbol means that the model was trained on GazeFollow and evaluated without fine-tuning. The modality column uses the codes I (image), T (time), D (depth), E (eyes), P (pose), and O (objects).

<table>
<thead>
<tr>
<th>Method</th>
<th>Multi</th>
<th>Modality</th>
<th>GazeFollow</th>
<th>VideoAttentionTarget</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>AUC ↑</td>
<td>Avg. Dist. ↓</td>
</tr>
<tr>
<td>Recasens [29]</td>
<td>×</td>
<td>I</td>
<td>0.878</td>
<td>0.190</td>
</tr>
<tr>
<td>Chong [6]</td>
<td>×</td>
<td>I+T</td>
<td>0.896</td>
<td>0.187</td>
</tr>
<tr>
<td>Lian [21]</td>
<td>×</td>
<td>I</td>
<td>0.906</td>
<td>0.145</td>
</tr>
<tr>
<td>Chong [7]</td>
<td>×</td>
<td>I+T</td>
<td>0.921</td>
<td>0.137</td>
</tr>
<tr>
<td>Fang [11]</td>
<td>×</td>
<td>I+D+E</td>
<td>0.922</td>
<td>0.124</td>
</tr>
<tr>
<td>Fang [11]</td>
<td>×</td>
<td>I+D</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Jin [17]</td>
<td>×</td>
<td>I+D</td>
<td>0.920</td>
<td>0.118</td>
</tr>
<tr>
<td>Jin [17]</td>
<td>×</td>
<td>I</td>
<td>0.909</td>
<td>0.137</td>
</tr>
<tr>
<td>Tonini [37]</td>
<td>×</td>
<td>I+D</td>
<td>0.927</td>
<td>0.141</td>
</tr>
<tr>
<td>Gupta [12]</td>
<td>×</td>
<td>I+D+P</td>
<td>0.943</td>
<td>0.114</td>
</tr>
<tr>
<td>Gupta [12]</td>
<td>×</td>
<td>I</td>
<td>0.933</td>
<td>0.134</td>
</tr>
<tr>
<td>Bao [3]</td>
<td>×</td>
<td>I+D+P</td>
<td>0.928</td>
<td>0.122</td>
</tr>
<tr>
<td>Hu [14]</td>
<td>×</td>
<td>I+D+O</td>
<td>0.923</td>
<td>0.128</td>
</tr>
<tr>
<td>Tafasca [35]</td>
<td>×</td>
<td>I+D</td>
<td>0.936</td>
<td>0.125</td>
</tr>
<tr>
<td>Jin [16]</td>
<td>✓</td>
<td>I</td>
<td>0.919</td>
<td>0.126</td>
</tr>
</tbody>
</table>

Table 2. Results on the ChildPlay dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Multi</th>
<th>Dist. ↓</th>
<th>PLAH ↑</th>
<th>AP ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gupta [12]</td>
<td>×</td>
<td>0.113</td>
<td>—</td>
<td>0.983</td>
</tr>
<tr>
<td>Tafasca [35]</td>
<td>×</td>
<td>0.107</td>
<td>0.590</td>
<td>0.986</td>
</tr>
<tr>
<td>Sharingan†</td>
<td>✓</td>
<td>0.109</td>
<td>0.600</td>
<td>—</td>
</tr>
<tr>
<td>Sharingan</td>
<td>✓</td>
<td>0.106</td>
<td>0.600</td>
<td>0.990</td>
</tr>
</tbody>
</table>

Number of People. One consideration of particular importance in the context of Sharingan is the influence of the number of people on training and evaluation. Table 4 shows that increasing $N_{tr}^{p}$ during training does not influence performance. We believe that since the gaze tokens are interchangeable and play a symmetric role (i.e., no order encoding), as long as $N_{tr}^{p} > 1$, the model will be forced to learn how to accommodate an arbitrary number of people in the scene during inference by sharing the learned image representations among them. However, the same cannot be said for the single-person setting ($N_{tr}^{p} = 1$), which is fundamentally different from multi-person ($N_{tr}^{p} > 1$). In this case, the model inevitably learns to tailor its image representations to a single person. To illustrate this behavior, we evaluate a single-person ($N_{tr}^{p} = 1$) and a multi-person ($N_{tr}^{p} = 2$) models twice: first using one person ($N_{i}^{p} = 1$) and a second time by processing all people in the scene ($N_{i}^{p} = all$). Figure 5 (right) shows that the multi-person model is perfectly able to do both single-person and multi-person prediction. However, the single-person model experiences a significant degradation when attempting multi-person prediction.

Heatmap vs 2D Point. Gaze following has always been framed as a heatmap prediction task. To the best of our
and person processing, and selectively fuse their information. One idea worth exploring is to disentangle scene information. It is difficult to understand how the model is combining information. In this section, we explore this formulation further to gain a deeper understanding of the task. To this end, we trained a Sharingan model by replacing the Conditional DPT decoder with a simple MLP to directly regress \((x, y)\) coordinates from the output person token. Since the 2D Point model can only predict a single value, it can’t represent the entire distribution over the 2D space of the image. Instead, we believe that it converges to an expectation of this posterior probability. When this distribution is multimodal (i.e., more than one probable gaze target), the expectation becomes unlikely under that posterior distribution. We illustrate this behavior in Figure 6 by comparing the person-specific attention map from the last layer of the encoder of the 2D Point model to the predicted gaze heatmap from the Heatmap variant. It is clear that both models capture the different modes quite well, but the former outputs \((x, y)\) coordinates resembling a weighted average of these modes, which ends up distant from all of them.

This phenomenon leads to the results given in Table 5, showing that the 2D point model achieves better average distance and PLAH, but lags behind the Heatmap model in terms of minimum distance and RLAH. This is not surprising since the average distance of the gazeFollow test set is a distance to an average of modes (i.e., multiple annotated points), which loosely resembles the objective that the 2D Point model is optimizing for. Consequently, we believe that the average distance should never be the only metric for evaluating gaze following methods, a claim also supported by [3]. On the other hand, positive LAH predictions for the 2D Point model mean that it is very confident, which explains the better PLAH value. However, this also means that the number of false negatives will be high, hence the big gap in RLAH because the predicted point will often be slightly off-target when the model is not confident (e.g., marginally outside the head area when looking at a person).

**Limitation.** Sharingan processes both image and gaze tokens simultaneously in the transformer. This means that the same weights operate on both types of tokens, which makes it difficult to understand how the model is combining information. One idea worth exploring is to disentangle scene and person processing, and selectively fuse their information. This may help improve the stability of predictions on videos, and allow people to interact together in a more meaningful way.

### 6. Conclusion

In this paper, we introduced Sharingan, a novel and efficient transformer-based architecture for gaze target prediction that is designed to support an arbitrary number of people out of the box. Sharingan stands out for its efficiency both in training and inference, delivering SOTA results on public benchmarks. Furthermore, it demonstrates strong generalization when tested on other datasets and naturalistic scenes. We also validated architectural decisions through extensive ablation experiments and discussed key aspects related to the model and the task.

Beyond gaze following, Sharingan’s intuitive design makes it suitable for other research areas in human behavior understanding. Specifically, the architecture can be extended to perform a multi-faceted analysis of social scenes by integrating different modalities (e.g., depth, semantics), and producing more outputs (e.g., gestures, interactions). We intend to explore this direction further in future work.

**Acknowledgement.** This research has been supported by the AI4Autism project (Digital Phenotyping of Autism Spectrum Disorders in children, grant agreement no. CRSII5 202235 / 1) of the the Sinergia interdisciplinary program of the SNSF.

### Table 4. Influence of the number of people during training on GazeFollow. The evaluation uses all available people.

<table>
<thead>
<tr>
<th>(N_{tr}^p)</th>
<th>AUC</th>
<th>Avg. D. (\downarrow)</th>
<th>Min. D. (\downarrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.944</td>
<td>0.113</td>
<td>0.057</td>
</tr>
<tr>
<td>3</td>
<td>0.943</td>
<td>0.114</td>
<td>0.058</td>
</tr>
<tr>
<td>4</td>
<td>0.942</td>
<td>0.113</td>
<td>0.057</td>
</tr>
</tbody>
</table>

### Table 5. Comparison between the heatmap and the 2D point training objectives on GazeFollow. RLAH is the recall of LAH.

<table>
<thead>
<tr>
<th></th>
<th>Avg. D.</th>
<th>Min. D.</th>
<th>PLAH</th>
<th>RLAH</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D Point</td>
<td>0.106</td>
<td>0.066</td>
<td>0.683</td>
<td>0.368</td>
</tr>
<tr>
<td>Heatmap</td>
<td>0.113</td>
<td>0.057</td>
<td>0.667</td>
<td>0.571</td>
</tr>
</tbody>
</table>

![Figure 6. Comparison of Heatmap (i.e. left) and 2D Point regression model (i.e. right) models. The heatmap on the right is obtained by computing the attention weights (i.e. last encoder layer) of the person’s gaze token with the image tokens.](image-url)
References


