

# Toward Gaze Target Detection of Young Autistic Children

Shijian Deng<sup>1</sup>, Erin E. Kosloski<sup>2</sup>, Siva Sai Nagender Vasireddy<sup>1</sup>, Jia Li<sup>1</sup>, Randi Sierra Sherwood<sup>2</sup>, Feroz Mohamed Hatha<sup>1</sup>, Siddhi Patel<sup>2</sup>, Pamela R. Rollins<sup>2</sup>, Yapeng Tian<sup>1</sup>

<sup>1</sup>Department of Computer Science, The University of Texas at Dallas

<sup>2</sup>School of Behavioral and Brain Sciences, The University of Texas at Dallas

{shijian.deng, erin.kosloski, sivasainagender.vasireddy, jia.li, randi.sherwood, feroz.hatha, siddhi.patel, rollins, yapeng.tian}@utdallas.edu

## Abstract

The automatic detection of gaze targets in autistic children through artificial intelligence can be impactful, especially for those who lack access to a sufficient number of professionals to improve their quality of life. This paper introduces a new, real-world AI application for gaze target detection in autistic children, which predicts a child’s point of gaze from an activity image. This task is foundational for building automated systems that can measure joint attention—a core challenge in Autism Spectrum Disorder (ASD). To facilitate the study of this challenging application, we collected the first-ever Autism Gaze Target (AGT) dataset. We further propose a novel Socially Aware Coarse-to-Fine (SACF) gaze detection framework that explicitly leverages the social context of a scene to overcome the class imbalance common in autism datasets—a consequence of autistic children’s tendency to show reduced gaze to faces. It utilizes a two-pathway architecture with expert models specialized in social and non-social gaze, guided by a context-awareness gate module. The results of our comprehensive experiments demonstrate that our framework achieves new state-of-the-art performance for gaze target detection in this population, significantly outperforming existing methods, especially on the critical minority class of face-directed gaze.

**Code** — <https://github.com/ShijianDeng/AGT>

## Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by challenges in social interaction, communication, and the presence of restricted or repetitive behaviors (Guha 2014). The CDC estimates that about 1 in 31 eight-year-olds has been identified with ASD (Shaw 2025). A core feature often observed in young autistic children is a difference in social attention, particularly in initiating and responding to joint attention (Bruinsma, Koegel, and Koegel 2004; Jones and Klin 2013; Zwaigenbaum et al. 2005). Joint attention refers to the shared focus of two individuals on an object and is a critical milestone in early social development. Its assessment is a cornerstone of early autism diagnosis and intervention, yet it is a process that is labor-intensive, requires highly trained professionals, and is difficult to scale (Luyster et al. 2009).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

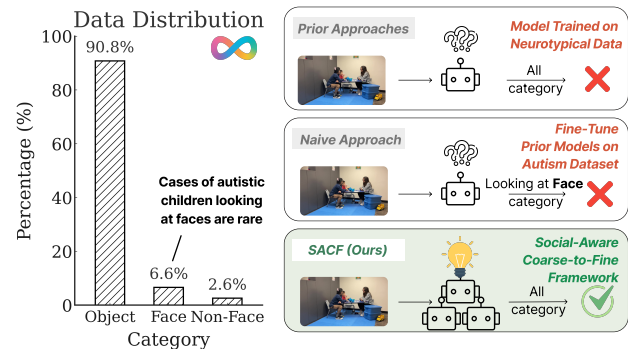


Figure 1: For gaze target detection in young autistic children: (a) Existing off-the-shelf models trained on neurotypical data cannot handle autism scenarios well because autistic individuals can have a different behavior distribution in the same social environment compared to neurotypical individuals, such as avoiding eye contact or not responding to verbal interaction. (b) Models directly fine-tuned on an autism dataset still struggle with the rare case of an autistic child looking at a person’s face. (c) Our Socially Aware Coarse-to-Fine framework uses an MLLM as a router to dynamically utilize a Socially Aware Gaze Expert model and a Socially Agnostic Gaze Expert to address these issues.

Artificial intelligence (AI) holds immense promise to augment the capabilities of clinicians by providing objective, scalable, and accessible tools for behavioral assessment. One of the most fundamental technical challenges in building such a tool is automatic gaze target detection: can an AI model determine where an autistic child is looking within a naturalistic setting? If successful, such a system could quantify instances of joint attention, track progress over time, and provide valuable feedback to caregivers and clinicians.

However, existing research on gaze target detection has focused almost exclusively on neurotypical adults (Recasens et al. 2015; Chong et al. 2020), and more recently, neurotypical children (Tafasca, Gupta, and Odohez 2023). These models are typically trained on datasets where gaze toward faces is common. When applied to autistic children, their performance degrades significantly. This is due to a fundamental difference in behavior: autistic children tend to look at

faces less frequently than their neurotypical peers (Klin et al. 2002; Chawarska, Macari, and Shic 2013).

To address the bias introduced by neurotypical datasets, we collected the first autism-specific dataset for gaze target detection. Nevertheless, autism related gaze behaviors also introduce a severe class imbalance. Gaze toward faces, although clinically significant, is underrepresented in the data. As a result, the trained models tend to predict non-social targets and could miss critical moments of social engagement.

To overcome this challenge, we developed a sophisticated, Socially Aware framework to handle the low-frequency gaze-to-face events in the data. Our socially aware framework leverages the large-scale prior knowledge learned by multi-modal large language models (MLLMs) (Bai et al. 2025) to first coarsely recognize the social context, then process and route the input to fine-grained gaze target detection expert models that can better handle such scenarios. Our approach helps us achieve new state-of-the-art performance on gaze target detection in autistic children, especially in socially-related cases, which is crucial for downstream tasks such as joint attention detection.

Our contributions can be summarized as follows: 1) We tackle a new, challenging, and socially impactful task: **Autism Gaze Target Detection**, which takes an image containing an autistic child and outputs the gaze target of that child; 2) We collect and annotate the first-ever **Autism Gaze Target (AGT)** dataset, a real-world dataset of autistic children in semi-structured play sessions, designed specifically for this task. It contains 16,582 annotated frames and reflects the natural gaze distribution of this population; 3) We propose a **Socially Aware Coarse-to-Fine (SACF)** gaze detection framework. Instead of a single, monolithic model, SACF first assesses the social context of the scene and then deploys specialized expert pathways for social (face-directed) and non-social gaze. This approach directly confronts the class imbalance problem and improves performance on the crucial, yet rare, social gaze events in the autistic population; 4) Our experiments show that SACF significantly outperforms state-of-the-art baselines, demonstrating its potential as a foundational component for AI-driven tools aimed at supporting the autism research community.

## Related Work

### AI for Autism

The application of AI and computer vision to autism research has gained significant traction. Early work focused on analyzing static images or videos to classify behaviors. Researchers have developed models to detect autism-related behaviors (Rajagopalan, Dhall, and Goecke 2013; Pandey et al. 2020; Negin et al. 2021; Wei et al. 2023; Ribeiro, Grellert, and Carvalho 2023; Deng et al. 2024a,b), analyze gesture and gait (Zahan et al. 2023), and classify movements (Singh et al. 2024). Some efforts have shifted towards quantifying attention dynamics. For example, (Jiang and Zhao 2017) uses a DNN-based model with eye-tracking data in free image reviewing. Others have developed an attention-based autism screening model that uses photo-taking and image reviewing for ASD (Chen and Zhao 2019). Building

on this line of research, we focus on gaze target detection in young autistic children in naturalistic social settings. Rather than analyzing gaze in isolation, we aim to understand how a child visually engages with the surrounding environment, including people and objects. Gaze target detection in this context offers a socially meaningful lens into early social attention development and supports the creation of tools that may assist in developmental assessment and intervention.

### Gaze Target Detection

Gaze target detection aims to identify the object or person someone is looking at in an image or video. Early methods focused on narrow areas (Fathi, Hodgins, and Rehg 2012). The advent of natural settings revolutionized the field. The GazeFollow model by Recasens et al. (2015) was a seminal work, introducing a dataset that was not limited to a specific domain. This in-the-wild approach spurred further research. Subsequent research introduced datasets leveraging temporal information from video (Chong et al. 2020). However, these prominent models and datasets are based on neurotypical adults (Fang et al. 2021; Gupta, Tafasca, and Odohez 2022; Bao, Liu, and Yu 2022; Jin et al. 2022). The recently proposed Childplay dataset (Tafasca, Gupta, and Odohez 2023) made strides by focusing on neurotypical children in play scenarios. While valuable, their data distribution still reflects a high frequency of social gaze. No existing work has systematically studied or addressed the unique challenges of gaze detection in autistic children, where the data distribution is inherently and dramatically skewed away from social targets. Although previous research (Tafasca, Gupta, and Odohez 2023) emphasizes the importance of the direction of autism gaze target detection, and Chong et al. (2020) provided preliminary evidence for automating gaze-based joint attention detection for toddlers, our work is the first to systematically tackle this specific, challenging, and clinically vital domain on a large-scale dataset.

### Autism Gaze Target Dataset

To facilitate the new research, we constructed the Autism Gaze Target (AGT) dataset. This dataset was collected and annotated under Institutional Review Board (IRB) approval with informed consent from all participants' legal guardians.

### Data Collection

We extracted nearly 20,000 video frames from 59 ethically sourced video recordings, which are part of the Autism Corpus built by autism experts on our team (Rollins et al. 2021; Rollins and De Froy 2023). Video recordings of autistic children were collected during administration of the Communication and Symbolic Behavior Scales-Developmental Profile (Wetherby and Prizant 2002) by a research clinician, as part of two broader studies.

### Data Annotation

From the recorded videos, we sampled frames at 1 frame per second (fps) and grouped every five consecutive frames into a clip. This approach reduces temporal redundancy while preserving meaningful attentional shifts. For each sampled

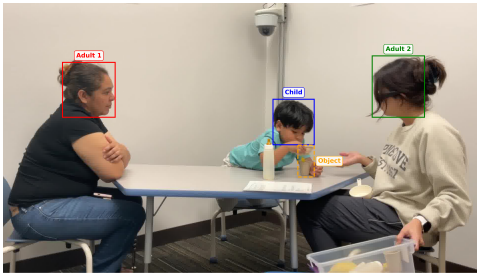


Figure 2: Example from our Autism Gaze Target (AGT) dataset. The child’s head is bounded by a blue box. The clinician’s and parent’s faces are marked with green and red boxes, respectively. The ground-truth target (a toy) is highlighted with a dashed orange box.

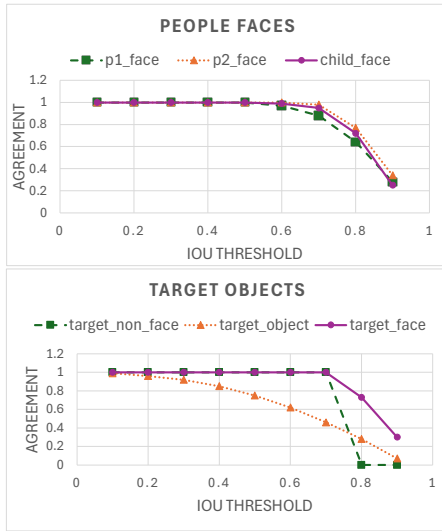


Figure 3: Annotation agreement. Most gaze targets are objects. Agreement remains high and drops slowly as the IoU threshold increases, indicating strong consistency between annotators.

frame, a team of trained annotators used the Computer Vision Annotation Tool (CVAT.ai 2018) to manually label social and non-social gaze targets. The annotation process involved the following steps:

Annotators used bounding boxes to mark the positions of:

1. **Child Face:** A bounding box for the child’s face.
2. **Adult Faces:** Bounding boxes for all adults’ faces in the scene. These were the clinician’s face or the parent’s face.
3. **Ground-Truth Target:** Annotators were instructed to carefully examine the child’s head pose and eye direction to identify the target the child was most likely looking at. Each gaze target was labeled as one of three categories: a target object, a target face, or a target located on a person but not the face, with the corresponding visual region marked. A “Noninclusive” category was used when the gaze is directed outside the frame or at no specific target.

Each frame was annotated by at least one annotator, result-

	Object	Face	Non-face
Object	499	4	12
Face	4	27	1
Non-face	12	1	1

Table 1: Confusion matrix of looking-target categories also shows high agreement between annotators.

Gaze Target Class	Frame Count	Percentage
Face (More-social)	1,088	6.6%
Not Face (Less-social)	15,494	93.4%
<b>Total</b>	<b>16,582</b>	<b>100.0%</b>

Table 2: Gaze target distribution in the Autism Gaze Target (AGT) dataset. The severe imbalance, with ‘Face’ as the minority class, poses a significant challenge for standard models.

ing in a total of 16,582 fully annotated video frames after filtering out frames without a gaze target. One example is illustrated in Figure 2. To ensure annotation quality, we randomly selected 800 frames for double annotation, where two annotators independently labeled the same frames. Figure 3 presents the inter-annotator agreement based on the Intersection over Union (IoU) criterion, computed as:

$$\text{agreement} = \begin{cases} 1 & \text{if IoU} > \text{threshold,} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We also generated a category confusion matrix (Table 1), which yields a Cohen’s Kappa score (Kvaalseth 1989) of 0.757, indicating substantial agreement.

## Data Statistics

The final AGT dataset reflects the expected gaze patterns in the young autistic population. As shown in Table 2, the distribution of gaze targets is highly imbalanced. Gaze is directed towards the adult’s face in only 6.6% of the frames. The majority of gaze is not directed towards face (93.4%), but instead toward other targets, including objects and person-non-face regions. This imbalance means standard models will likely struggle with this data, as the most clinically relevant class (‘Face’) is a minority class. We split the data into training (9874 frames), validation (3344 frames), and test (3364 frames) sets to ensure that our models generalize to unseen participants.

## Method

In this work, we aim to accurately predict a child’s gaze target, with a particular emphasis on correctly identifying the rare but crucial instances of social gaze towards a face.

## Problem Formulation

The task of gaze target detection aims to identify the specific location a person is looking at within a visual scene. In this paper, we formalize this task for the unique context of young

autistic children in naturalistic settings. Given an input image  $I \in \mathbb{R}^{H \times W \times 3}$  and the bounding box of the child’s head  $B_{head}$ , the primary objective is to predict the 2D coordinates of the gaze target,  $(x_t, y_t)$ , within the image frame.

Beyond spatial localization, a critical aspect of this problem involves the semantic interpretation of the target. Let  $C = \{\text{Face}, \text{Not Face}\}$  represent the set of possible target categories. In addition to predicting the gaze coordinates  $\mathbf{p}_t = (x_t, y_t)$ , the model must also assign a semantic label  $c_t \in C$  to the identified target. Therefore, the problem is to build a comprehensive framework  $\mathcal{F}$  that maps the input image  $I$  and head bounding box  $B_{head}$  to a tuple containing both the predicted location and its class:

$$(\mathbf{p}_t, c_t) = \mathcal{F}(I, B_{head}) \quad (2)$$

Successfully learning this function is essential for downstream clinical applications. The primary challenge is to optimize for this joint objective in a domain characterized by severe data imbalance, where the prior probability of the ‘Face’ class is significantly lower than that of the ‘Not Face’ class, i.e.,  $P(c_t = \text{Face}) \ll P(c_t = \text{Not Face})$ .

### Preliminary

To address the localization challenge defined above, modern methods typically train a model to produce a probabilistic gaze heatmap, from which the final coordinates are derived. Our framework builds upon such a state-of-the-art model with prior knowledge augmented by MLLMs.

**Gaze Target Detection Model.** For fine-grained gaze target localization, we use **GazeLLE** (Ryan et al. 2025), a state-of-the-art architecture designed for efficiency and accuracy. It revolutionizes previous methods by leveraging a single, frozen, pre-trained image encoder,  $\Phi_{enc}$  (e.g., DINOv2 (Oquab et al. 2023)), to extract a feature representation,  $F_{scene} \in \mathbb{R}^{N \times D}$ , from the entire scene image  $I$ .

Instead of a secondary network for the head, GazeLLE injects the subject’s location directly into the feature space using a person-specific positional prompt,  $E_{pos}$ , derived from the head bounding box  $B_{head}$ . This combined representation is then processed by a lightweight transformer decoder,  $Dec_{gaze}$ , to produce the final gaze heatmap,  $M_{gaze} \in [0, 1]^{H' \times W'}$ . The process can be summarized as:

$$M_{gaze} = Dec_{gaze}(\Phi_{enc}(I) + E_{pos}(B_{head}); \theta_{dec}) \quad (3)$$

where only the decoder parameters,  $\theta_{dec}$ , are trained. The model is optimized using a ground-truth heatmap,  $\hat{M}_{gaze}$ , which is generated by applying a Gaussian blur centered on the ground-truth target coordinates. The training objective is to minimize the pixel-wise Binary Cross-Entropy (BCE) loss between the predicted and ground-truth heatmaps:

$$\begin{aligned} \mathcal{L}_{gaze} = & -\frac{1}{H'W'} \sum_{i,j} \left( \hat{M}_{gaze}^{(i,j)} \log(M_{gaze}^{(i,j)}) \right. \\ & \left. + (1 - \hat{M}_{gaze}^{(i,j)}) \log(1 - M_{gaze}^{(i,j)}) \right) \end{aligned} \quad (4)$$

This architecture’s effectiveness and streamlined design make it an ideal base model for our problem. We also use

the recent **Sharingan** model (Tafasca, Gupta, and Odobez 2024) as another base model.

**Multimodal Large Language Model** To tackle the semantic interpretation aspect of the problem, we leverage **Qwen2.5-VL-7B-Instruct** (Bai et al. 2025), a small MLLM. We employ it as a Social Context Awareness (SCA) module to provide a high-level, coarse understanding of the scene. Given the image  $I$  and a textual prompt  $T$ , the MLLM estimates a social context score  $s$  corresponding to the probability that the child is looking at a face:

$$s = P_{SCA}(V_{answer} = \text{looking at face} | I, T; \theta_{MLLM}) \quad (5)$$

This score allows our framework to dynamically adapt its strategy based on the semantic potential of the scene.

### Socially Aware Coarse-to-Fine Framework

The central challenge in gaze target detection for autistic children stems from the severe class imbalance in the data. Standard models fail on this task because their optimization objectives are inherently compromised by this distribution, where the probability of the clinically critical ‘Face’ class is dwarfed by the ‘Not Face’ class. Such models inevitably learn to prioritize the majority class, achieving low average error by effectively ignoring the rare but vital social gaze.

To counteract this bias, our core idea is that a “divide and conquer” strategy that decomposes the problem into coarse-grained semantic understanding and fine-grained specialized localization would outperform a single, monolithic model. This motivation guides our proposed **Socially Aware Coarse-to-Fine (SACF)** framework,  $\mathcal{F}_{SACF}$ . As illustrated in Figure 4, the framework first assesses the scene’s social context before routing it to an expert model fine-tuned for that specific context.

**Social Context Awareness (SCA) Module.** To enable a “divide and conquer” approach, the framework must first answer: *What is the potential looking target in this scene?* A powerful MLLM, with its vast pre-trained knowledge of visual scenes, can serve as an effective estimator for this high-level task. We therefore implement the SCA module using a fine-tuned Qwen2.5-VL model. It is trained on a binary proxy task derived from our data,  $\mathcal{D}_{SCA} = \{I_i\}_{i=1}^N$ , to estimate the social context score  $s$ , which represents the likelihood that the child is looking at an adult’s face. The framework uses the social context score  $s$  from the SCA module to gate the two pathways via a simple threshold,  $\tau$ . This determines the coarse class prediction  $c_{coarse}$ :

$$c_{coarse} = \begin{cases} 1 & \text{if } s \geq \tau \\ 0 & \text{if } s < \tau \end{cases} \quad (6)$$

Here, 1 denotes face and 0 denotes not face. This provides the coarse assessment of the scene’s social context needed to guide the subsequent fine-grained analysis.

**Two-Pathway Gated Experts.** Given a coarse classification from the SCA, the next challenge is to perform accurate localization. A single fine-grained model would still be influenced by the data imbalance. We therefore hypothesize that training two separate experts,  $Ex_{aware}$  and  $Ex_{agnostic}$ ,

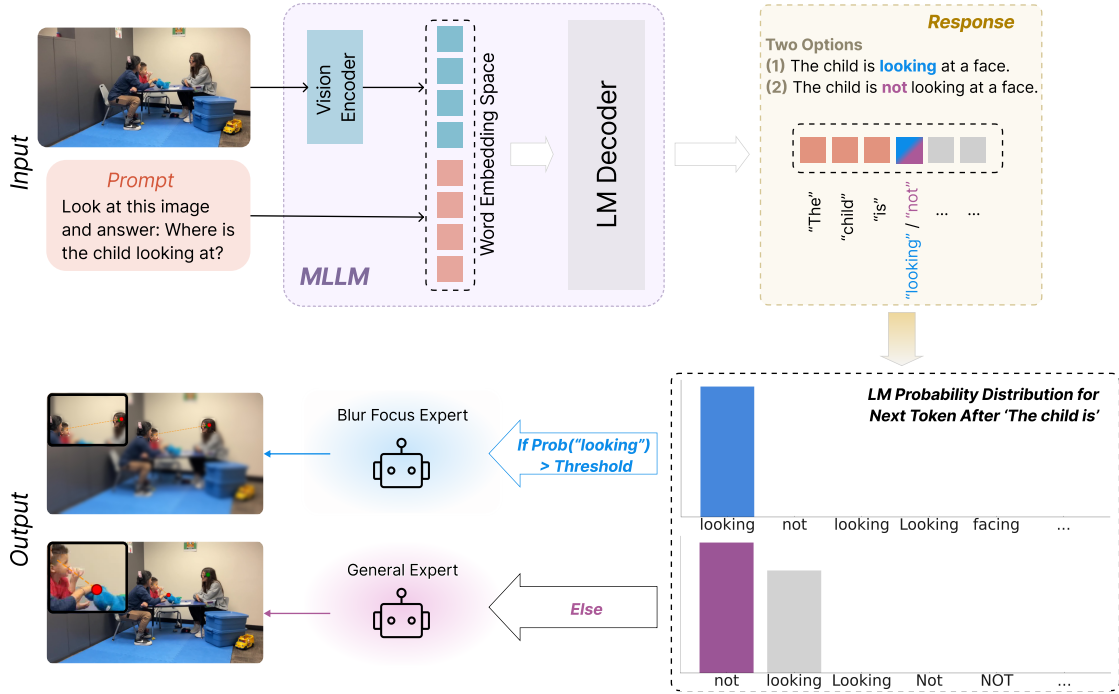


Figure 4: The architecture of our proposed Socially Aware Coarse-to-Fine (SACF) gaze detection framework. An input image is first processed by the Social Context Awareness (SCA) module to generate a social context score  $s$ . This score is used by a gate to dynamically assign the input image to two specialized expert pathways: a Socially Aware Gaze Expert ( $Ex_{aware}$ , Blur Focus) and a Socially Agnostic Gaze Expert ( $Ex_{agnostic}$ , General). The final prediction is derived from the routed expert outputs, allowing the framework to adapt its specialization based on the scene’s social context.

with distinct and specialized objectives will be more effective. Each expert model, based on the GazeLLE architecture, maps an image and head box to a gaze heatmap,  $Map = Ex(I, B_{head}; \theta)$ .

**Socially Aware Gaze Expert ( $Ex_{aware}$ ).** The rationale for this expert is to create a specialist for the case when the social context is clear. We train it on an augmented dataset,  $\tilde{D}$ , where we apply a strong Gaussian blur,  $G_{blur}$ , to irrelevant regions of an image  $I$  when the target is certain. This encourages the model to focus on plausible targets (e.g., faces) and avoid extreme failure cases. The expert is trained by minimizing the gaze loss  $\mathcal{L}_{gaze}$  on this modified data:

$$\min_{\theta_{face}} \sum_{\tilde{I}_i \in \tilde{D}} \mathcal{L}_{gaze}(Ex_{aware}(\tilde{I}_i, B_{head}), \hat{M}_{gaze}) \quad (7)$$

**Socially Agnostic Gaze Expert ( $Ex_{agnostic}$ ).** Conversely, for the frequent non-social and unsure scenarios, the framework needs an expert that is maximally performant on those cases. The rationale is to let this model fully optimize for the most common situations without being constrained by the specialized requirements of the ‘Face’ class. Thus, its parameters,  $\theta_{agnostic}$ , are optimized by minimizing the same

gaze loss  $\mathcal{L}_{gaze}$  on the original, unmodified training set  $\mathcal{D}$ :

$$\min_{\theta_{agnostic}} \sum_{I_i \in \mathcal{D}} \mathcal{L}_{gaze}(Ex_{agnostic}(I_i, B_{head}), \hat{M}_{gaze}) \quad (8)$$

**Gating and Inference.** The final step is to integrate the coarse assessment with the fine-grained expert predictions during inference. Our rationale is that a decisive, dynamic routing mechanism is the most effective way to leverage the experts’ specialized knowledge to predict gaze target coordinates:

$$\mathbf{p}_t^* = \begin{cases} \arg \max_{\mathbf{p}_t} (Ex_{aware}(\tilde{I}, B_{head})) & \text{if } c_{coarse} = 1 \\ \arg \max_{\mathbf{p}_t} (Ex_{agnostic}(I, B_{head})) & \text{if } c_{coarse} = 0 \end{cases} \quad (9)$$

While the coarse assessment guides the choice of expert, the final semantic classification,  $c_{fine}$ , is refined based on these precise output coordinates. Let  $\mathcal{B}_{faces} = \{B_{face}^{(1)}, \dots, B_{face}^{(k)}\}$  be the set of all adult face bounding boxes in the image. The final class is determined by checking if the predicted point  $\mathbf{p}_t$  falls within the union of these bounding box regions:

$$c_{fine} = \begin{cases} \text{Face} & \text{if } \mathbf{p}_t \in \bigcup_{B \in \mathcal{B}_{faces}} \text{Area}(B) \\ \text{Not face} & \text{otherwise} \end{cases} \quad (10)$$

Model	$L^2 \downarrow$	$L^2_{obj} \downarrow$	$L^2_{face} \downarrow$	$L^2_{pnf} \downarrow$	Precision	Recall	F1
Qwen2.5-VL-7B-Instruct-AGT	0.0475	0.0413	0.1237	0.0837	0.7050	0.6260	0.6630
Sharingan	0.0615	0.0615	0.0647	0.0561	0.4377	0.8010	0.5660
Sharingan-AGT	0.0486	0.0451	0.0949	0.0595	0.7744	0.7330	0.7531
Sharingan-SACF (Ours)	<b>0.0480</b>	0.0453	0.0817	0.0616	0.7647	0.7573	<b>0.7610</b>
Sharingan-SACF (Upper Bound)	0.0478	0.0471	0.0378	0.0923	0.9581	1.0000	0.9786
GazeLLE	0.0670	0.0630	0.1092	0.1041	0.3868	0.6553	0.4865
GazeLLE-AGT	0.0460	0.0405	0.1130	0.0804	0.6984	0.6408	0.6684
GazeLLE-SACF (Ours)	<b>0.0453</b>	0.0405	0.1019	0.0804	0.7041	0.6699	<b>0.6866</b>
GazeLLE-SACF (Upper Bound)	0.0396	0.0391	0.0307	0.0752	0.9856	0.9951	0.9903

Table 3: Performance comparison of different gaze target detection methods. Metrics include various  $L^2$ , and standard classification scores.

This two-step classification process—a coarse routing followed by a fine-grained spatial check—allows SACF to dynamically apply the correct specialist while making a final, evidence-based semantic judgment, effectively resolving the performance trade-offs imposed by the imbalanced data.

## Experiments

### Experimental Setup

**Dataset.** We use our Autism Gaze Target (AGT) dataset with the splits described previously.

**Models.** We compare our SACF against other models:

- **Sharingan (Tafasca, Gupta, and Odobez 2024):** A recent model trained on the Childplay dataset with neurotypical children.
- **GazeLLE (Ryan et al. 2025):** Current SOTA trained on the Childplay.
- **Sharingan-AGT:** Sharigan trained on our AGT dataset.
- **GazeLLE-AGT:** GazeLLE trained on our AGT dataset.
- **Qwen2.5-VL-7B-Instruct-AGT:** The MLLM fine-tuned on our AGT training set.
- **SACF (Ours):** Our Socially Aware Coarse-to-Fine gaze detection framework. We select two recent gaze target detection SOTA models: Sharigan and GazeLLE as the base models of our framework.
- **SACF w/ SCA upper-bound:** A version of our model that uses a perfect gate model (i.e., an oracle), showing the framework’s potential with an ideal social context awareness module.

**Evaluation Metrics.** Following established practice in gaze target detection (Recasens et al. 2015; Chong et al. 2020; Fang et al. 2021; Tafasca, Gupta, and Odobez 2023; Ryan et al. 2025), we report seven complementary scores that jointly capture (i) *where* the model points and (ii) *whether* it flags the scene as social (*Face*) or not.

- $L^2$  - mean Euclidean distance (normalised) between the predicted and ground-truth gaze point over the test set.
- $L^2_{obj} / L^2_{face} / L^2_{pnf}$  - the same error computed only on frames whose ground-truth target is, respectively, an

Class	Precision	Recall	F1 Score	Support
Face	69.23%	65.53%	0.673	206
Not-face	97.76%	98.10%	0.979	3,158

Table 4: SAC’s performance for the Face vs. Not-face task.

*Object*, an *Adult Face*, or a *Person-Non-Face* region (e.g. torso or hands). Using three separate classes makes minority-class performance visible instead of being averaged away.

- **Precision, Recall, F1** - macro-averaged over the binary decision *Face* vs. *Not-face*. Macro averaging prevents the 93.4 % majority class from inflating the score.

**Implementation Details.** All models were implemented in PyTorch. We used a Qwen2.5VL-7B as the gate model and GazeLLE for our expert models. Training and testing were performed using NVIDIA A6000 GPUs.

### Results and Analysis

**Overall gains across the board.** Table 3 shows that our SACF framework achieves new state-of-the-art results. Using the *Sharingan* backbone, SACF reduces the  $L^2$  from 0.0486 to 0.0480, while pushing the F1 score from 0.7531 to 0.7610. A similar trend is observed with the *GazeLLE* backbone, where overall  $L^2$  drops from 0.0460 to 0.0453 and F1 climbs from 0.6684 to 0.6866.

**Models trained on neurotypical data fail on our autism dataset.** We found that models trained on neurotypical datasets, like the original Sharigan and GazeLLE, perform poorly on our AGT test set (Table 3) by overestimating the social engagement of autistic children. Their performance improves significantly after being trained on our AGT dataset (Figure 5), highlighting the social distribution shift and the need for autism-specific data.

**Excelling on the rare but crucial *Face* class.** Because only 6.6% of gaze events are directed at a face (Table 2), improvements on this minority class are easily masked by majority performance. SACF overturns that bias: on *GazeLLE*,  $L^2_{face}$  falls from 0.1130 to 0.1019; ( $\downarrow 9.8\%$ ), and on *Sharingan*

Scenario	Sample Count	$Ex_{aware}$ 's $L^2$	$Ex_{agnostic}$ 's $L^2$	Winner
<i>Correct Predictions</i>				
Pred=Face, GT=Face	135	0.0303	0.0898	$Ex_{aware}$ (3x better)
Pred=Not-face, GT=Not-face	3,098	0.0398	0.0409	$Ex_{aware}$ (slightly better)
<i>Incorrect Predictions</i>				
Pred=Face, GT=Not-face	60	0.2818	0.0817	$Ex_{agnostic}$ (3.4x better)
Pred=Not-face, GT=Face	71	0.2090	0.1569	$Ex_{agnostic}$ (1.3x better)

Table 5: Analysis of expert models’ performance for different prediction scenarios. The table is grouped by correct and incorrect predictions. As we can see, since misclassifications can hurt overall performance, it is essential to obtain a good SCA.



Figure 5: A child is playing with an object. However, the baseline Sharingan model trained on the neurotypical Child-play dataset (left) is biased and predicts the child is looking at the clinician’s face. When trained on our AGT dataset (right), it correctly predicts the child is looking at the object. The green dot denotes the ground-truth gaze target, while the red dot denotes the predicted target.

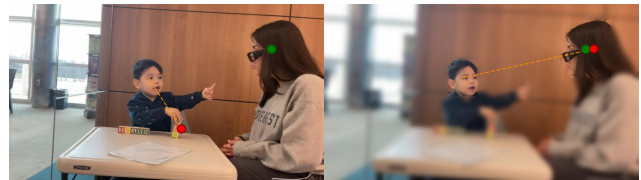


Figure 6: A child is actively interacting with a clinician. Our SACF framework (right) correctly localizes gaze to the face, thanks to the prior knowledge about social context leveraged from the MLLM. The simply trained GazeLLE-AGT (left) fails due to a lack of this social context awareness. Green dot denotes the ground-truth, while red denotes the prediction.

from 0.0949 to 0.0817;(↓13.9%). This directly translates into more reliable detection of socially relevant moments in clinical practice (Figure 6).

**An expert-level view of why SACF works.** Table 5 reveals that the Socially Aware Gaze Expert is  $3\times$  more accurate than the Socially Agnostic Gaze Expert when a face is indeed the target. Conversely, when the scene is mis-routed, the Socially Agnostic Gaze Expert is up to  $3.4\times$  better. The Context-Aware gate therefore acts as a high-stakes switch: when it chooses correctly (96% of the time), the framework benefits from the best of both specialists; when it errs, performance degrades gracefully rather than catastrophically.

**Upper-bound analysis highlights room for growth.** With the performance of our current small MLLM SCA (Table 4), we are already able to improve the final performance of the framework by routing input to different experts based on social context. Replacing the learned gate with an oracle (*SACF Upper Bound*) yields an F1 of 0.9903 and an almost perfect  $L^2_{face}$  of 0.0307. This gap pinpoints the gate as a promising lever for future gains: any advancement in social-context recognition will directly unlock further accuracy. Considering the current rapid development of MLLMs, our framework also stands to benefit significantly.

**Clinical impact.** The 10-14% drop in Face-target localization error means that, on average, the predicted gaze point lands 4-6 pixels closer to the true facial region at a  $224\times 224$  resolution. In downstream joint-attention related scoring pipelines, this improves F1 by around 2% (Table 3), bringing automated assessments one step closer to human-expert reliability.

## Conclusion

In this paper, we introduced gaze target detection in autistic children, a critical task for developing AI-powered tools to support the autism community. We addressed the scarcity of appropriate data by collecting and annotating the AGT dataset, which consists of 16,582 frames of young autistic children in naturalistic play sessions. The dataset’s inherent class imbalance, a direct result of the behavioral characteristics of autism, renders existing gaze detection methods inadequate. To overcome this challenge, we proposed the Socially Aware Coarse-to-Fine (SACF) gaze detection framework. By first assessing the social context of a scene with Social Context Awareness (SCA) module and then using a gated two-pathway system of experts, our model achieves new state-of-the-art performance. It significantly improves the detection of rare but clinically vital gaze-to-face events without compromising its accuracy on more frequent non-social targets. Our work represents a substantial step towards building automated systems for objective, scalable measurement of gaze targets for young autistic children.

## Acknowledgments

This work was supported by a UT Dallas Seed Program for Interdisciplinary Research (SPIRe) Award and grants from the Texas Higher Education Coordinating Board (THECB) Autism Grants Program (Rollins, PI; NOGA #20476 & #27509). The opinions and conclusions expressed are those of the authors and do not necessarily represent the opinions or policies of the funding sources.

## References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.
- Bao, J.; Liu, B.; and Yu, J. 2022. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14126–14135.
- Bruinsma, Y.; Koegel, R. L.; and Koegel, L. K. 2004. Joint attention and children with autism: A review of the literature. *Mental retardation and developmental disabilities research reviews*, 10(3): 169–175.
- Chawarska, K.; Macari, S.; and Shic, F. 2013. Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. *Biological psychiatry*, 74(3): 195–203.
- Chen, S.; and Zhao, Q. 2019. Attention-based autism spectrum disorder screening with privileged modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1181–1190.
- Chong, E.; Wang, Y.; Ruiz, N.; and Rehg, J. M. 2020. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5396–5406.
- CVAT.ai. 2018. Computer Vision Annotation Tool (CVAT). <https://github.com/cvat-ai/cvat>. Open-source, web-based image and video annotation tool (MIT license).
- Deng, A.; Yang, T.; Chen, C.; Chen, Q.; Neely, L.; and Oyama, S. 2024a. Language-assisted deep learning for autistic behaviors recognition. *Smart Health*, 32: 100444.
- Deng, S.; Kosloski, E. E.; Patel, S.; Barnett, Z. A.; Nan, Y.; Kaplan, A.; Aarukapalli, S.; Doan, W. T.; Wang, M.; Singh, H.; et al. 2024b. Hear me, see me, understand me: Audio-visual autism behavior recognition. *IEEE Transactions on Multimedia*.
- Fang, Y.; Tang, J.; Shen, W.; Shen, W.; Gu, X.; Song, L.; and Zhai, G. 2021. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11390–11399.
- Fathi, A.; Hodgins, J. K.; and Rehg, J. M. 2012. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1226–1233. IEEE.
- Guha, M. 2014. Diagnostic and statistical manual of mental disorders: DSM-5. *Reference Reviews*, 28(3): 36–37.
- Gupta, A.; Tafasca, S.; and Odobez, J.-M. 2022. 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*, 5041–5050.
- Jiang, M.; and Zhao, Q. 2017. Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the IEEE international conference on computer vision*, 3267–3276.
- Jin, T.; Yu, Q.; Zhu, S.; Lin, Z.; Ren, J.; Zhou, Y.; and Song, W. 2022. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113: 104924.
- Jones, W.; and Klin, A. 2013. Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504(7480): 427–431.
- Klin, A.; Jones, W.; Schultz, R.; Volkmar, F.; and Cohen, D. 2002. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry*, 59(9): 809–816.
- Kvaalseth, T. O. 1989. Note on Cohen’s kappa. *Psychological reports*, 65(1): 223–226.
- Luyster, R.; Gotham, K.; Guthrie, W.; Coffing, M.; Petrak, R.; Pierce, K.; Bishop, S.; Esler, A.; Hus, V.; Oti, R.; et al. 2009. The Autism Diagnostic Observation Schedule—Toddler Module: A new module of a standardized diagnostic measure for autism spectrum disorders. *Journal of autism and developmental disorders*, 39(9): 1305–1320.
- Negin, F.; Ozyer, B.; Agahian, S.; Kacdioglu, S.; and Ozyer, G. T. 2021. Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders. *Neurocomputing*, 446: 145–155.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Pandey, P.; Prathosh, A.; Kohli, M.; and Pritchard, J. 2020. Guided weak supervision for action recognition with scarce data to assess skills of children with autism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 463–470.
- Rajagopalan, S.; Dhall, A.; and Goecke, R. 2013. Self-stimulatory behaviours in the wild for autism diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 755–761.
- Recasens, A.; Khosla, A.; Vondrick, C.; and Torralba, A. 2015. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*. \* indicates equal contribution.
- Ribeiro, G. O.; Grellert, M.; and Carvalho, J. T. 2023. Stimming behavior dataset-unifying stereotype behavior dataset in the wild. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, 225–230. IEEE.
- Rollins, P. R.; De Froy, A.; Campbell, M.; and Hoffman, R. T. 2021. Mutual gaze: An active ingredient for social development in toddlers with ASD: A randomized control trial. *Journal of Autism and Developmental Disorders*, 51(6): 1921–1938.
- Rollins, P. R.; and De Froy, A. M. 2023. Reexamining Pathways Early Autism Intervention in children before and after the third birthday: A randomized control trial. *Journal of Autism and Developmental Disorders*, 53(3): 1189–1201.
- Ryan, F.; Bati, A.; Lee, S.; Bolya, D.; Hoffman, J.; and Rehg, J. M. 2025. Gaze-llc: Gaze target estimation via large-scale

learned encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28874–28884.

Shaw, K. A. 2025. Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years—Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022. *MMWR. Surveillance Summaries*, 74.

Singh, P.; Pathak, A.; Ganai, U. J.; Bhushan, B.; and Subramanian, V. K. 2024. Video based computational coding of movement anomalies in asd children. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 400–409.

Tafasca, S.; Gupta, A.; and Odobez, J.-M. 2023. Child-play: A new benchmark for understanding children’s gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20935–20946.

Tafasca, S.; Gupta, A.; and Odobez, J.-M. 2024. Sharingan: A transformer architecture for multi-person gaze following. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2008–2017.

Wei, P.; Ahmedt-Aristizabal, D.; Gammulle, H.; Denman, S.; and Armin, M. A. 2023. Vision-based activity recognition in children with autism-related behaviors. *Heliyon*, 9(6).

Wetherby, A. M.; and Prizant, B. M. 2002. *Communication and symbolic behavior scales: Developmental profile*. Paul H Brookes Publishing Co.

Zahan, S.; Gilani, Z.; Hassan, G. M.; and Mian, A. 2023. Human gesture and gait analysis for autism detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3328–3337.

Zwaigenbaum, L.; Bryson, S.; Rogers, T.; Roberts, W.; Brian, J.; and Szatmari, P. 2005. Behavioral manifestations of autism in the first year of life. *International journal of developmental neuroscience*, 23(2-3): 143–152.