Gaze-Anchored Social Net: Decoding Implicit Relations via Joint Modeling

Yuqi Hou[1][0009−0002−3086−8698], Zhuo Chen[1][0000−0003−2988−9387], Han Hu[1][0000−0002−9667−1698], Je Woo Kim[2][0009−0004−5073−7307], Jianbo Jiao[1][0000−0003−0833−5115], and Hyung Jin Chang[1][0000−0001−7495−9677]

[1] University of Birmingham, United Kingdom
{yxh029,zxc417,hxh347}@student.bham.ac.uk, {h.j.chang,j.jiao}@bham.ac.uk
[2] Korea Electronics Technology Institute, Korea
jwkim@keti.re.kr

Abstract. Human gaze does more than point to visual targets; it serves as a subtle indicator of social intent within static images, whereas standard models typically process individuals independently, treating gaze as an i.i.d. quantity or predicting social semantics in isolation. Recent multi-person methods attempt to address this but often treat social relations as rigid, post-hoc classifications decoupled from the gaze estimation process. This oversimplification fails to capture the nuanced nature of social intent, which acts as an underlying driver of gaze behavior rather than a secondary categorical output. We address these limitations by proposing ANCHOR, a target-centric paradigm designed to decode gaze-anchored social intent by modeling the joint distribution of visual attention and latent implicit relations. Our approach surfaces these dependencies as the latent structural scaffolding of gaze behavior. The architecture utilizes a relational attention mechanism to capture fine-grained interpersonal links, leveraging feature-wise modulation for efficient multi-person parsing from a single vision backbone. To stabilize the training of this coupled formulation, we implement an optimization synergy to resolve the inherent conflicts between spatial gaze accuracy and latent social reasoning. This approach ensures robust generalization by seeking stable, flat minima while simultaneously harmonizing competing task gradients. We validate our framework on an extended benchmark featuring dense multi-person annotations and novel social influence rankings. Our results demonstrate state-of-the-art performance and provide the first quantitative evidence that implicit social hierarchies can be robustly disentangled and learned directly from static gaze patterns.

Keywords: social net · static multi-person gaze · optimization synergy.

1 Introduction

Human gaze conveys a wide range of social cues. Prior work shows that where people look reflects their mental states, interactions, and these signals extend

well beyond the role of gaze in basic visual processing [18, 19, 26]. In multi-person scenes, gaze patterns encode rich information about social intent, and interpersonal relationships that extend far beyond individual visual attention.

Despite this potential, current gaze-following methods [12, 28] share a key limitation: they process individuals independently, modeling gaze as $P(X^{(i)}_{\text{gaze}} \mid X_{\text{scene}})$ for each person, implicitly assuming conditional independence across individuals and ignoring inter-person interactions.

Even multi-person models [33, 28] primarily optimize for efficiency while maintaining this independence assumption. Other approaches [8, 9] add auxiliary social tasks such as looking-at-head, shared attention, or looking-at-each-other, but these remain pairwise and rely on coarse, discrete labels within a decoupled framework that models gaze and social cues separately as $P(X_{\text{gaze}} \mid X_{\text{scene}})$ and $P(X_{\text{social}} \mid X_{\text{gaze}}, X_{\text{scene}})$, which cannot express the continuous subtleties of social behavior. Similar limitations appear in broader social group research [15].

We argue that the challenge is not to infer social intent post-hoc as $P(X_{social} | X_{gaze}, X_{scene})$, but rather to model the joint distribution $P(X_{gaze}, X_{social} | X_{scene})$. This formulation reveals a crucial observation: features necessary for accurate gaze prediction are inherently entangled with latent social intent X_{social}. In other words, a model achieving high accuracy on gaze prediction must have already implicitly learned to decode social dynamics to succeed. The real challenge, therefore, is not adding discrete labels, but explicitly surfacing this latent information already encoded in the learned features.

Surfacing this latent social information introduces three fundamental challenges that have not been addressed in prior work: (i) latent decoding of implicit social intent without explicit supervision; (ii) synergistic optimization to mitigate gradient interference between gaze and social tasks; and (iii) quantitative validation in the absence of established benchmarks.

Preliminary experiments indicated that modeling global social networks via GNNs introduces significant noise from distant agents, while explicit social annotations often prove subjective and unreliable. Social perception is naturally organized from a self-centered reference point, where observers prioritize cues relative to a focal individual [20]. Leveraging the relational richness inherent in modern vision backbones [21], we propose a target-centric paradigm. Instead of a noisy global graph, we decode social dynamics through the subjective perspective of the target head, filtering spatial noise while preserving high-fidelity relational context.

To address the Decoding Challenge, we introduce the ANCHOR(Fig 1) framework, designed to surface the implicit relations that drive visual attention. At its core, the architecture employs a relational attention mechanism that models the local interplay between a target individual and their geometric neighbors. By treating the target's features as a query against its surroundings, the model explicitly decodes the joint dependencies that constitute gaze-anchored social net. This relational context is then integrated through feature-wise modulation, allowing for precise, social-aware parsing of multiple individuals from a single frozen encoder.

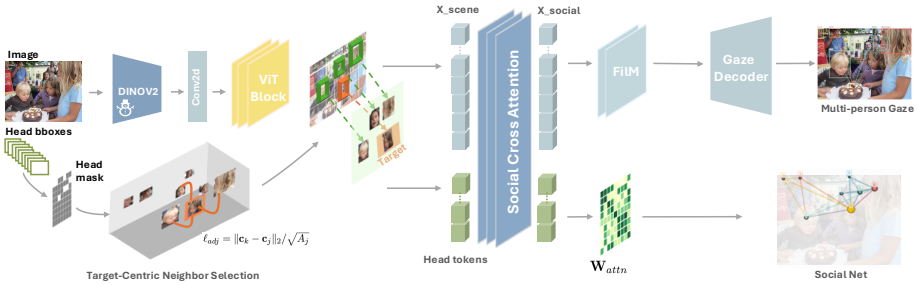


Fig. 1: Overview of the ANCHOR Framework.

Jointly optimizing gaze and social intent triggers gradient conflicts that destabilize training. We resolve these by enforcing flat minima and harmonizing gradients to mitigate interference. A self-supervised consistency objective provides geometric signals to refine social modules without manual labels.

To address the Validation Challenge, we introduce two novel extensions to the GazeFollow test set: (1) multi-person gaze annotations to validate joint-distributional modeling, and (2) Social Influence Ranking (SIR) labels. By capturing collective gaze convergence, SIR transforms ambiguous social perceptions into a measurable ranking task. We evaluate this via the Social Consistency Index (SCI) to validate internal geometric mechanisms and Social Leader Accuracy (SLA) to measure practical leader recognition. Together, these benchmarks provide the first quantitative evidence that implicit social dynamics can be robustly disentangled and learned.

Our work is the first to identify and address the fundamental limitations in current gaze-social modeling paradigms. Our specific contributions are:

- We propose ANCHOR, a target-centric framework that decodes a gaze-anchored social net through relational attention and a self-supervised L_{SCI} loss, uncovering latent social dynamics without manual supervision.
- Our synergistic optimization strategy resolves gradient conflicts in multi-task learning, ensuring stable convergence and robust generalization.
- We introduce the Extended GazeFollow benchmark with multi-person gaze and Social Influence Ranking (SIR) labels, achieving state-of-the-art multi-person gaze prediction performance and providing the first quantitative validation of learned social dynamics.

2 Related Work

From Independent to Multi-person Gaze Following. Early gaze-following methods typically relied on a dual-stream architecture integrating head-orientation cues with spatial saliency maps [24, 25, 31]. Subsequent research has progressively enhanced this foundation by incorporating diverse modalities such as 3D

gaze models [2], scene depth [6, 13, 11], and panoramic context [17], or leveraging high-level semantics including human-object interactions [37, 35] and audio-visual cues [12]. Despite providing robust features, these works primarily target single-subject scenarios. Recent multi-person methods [33, 8, 9] add auxiliary social tasks but remain constrained by coarse, discrete interaction categories.

The Conditional Independence Fallacy. Recent architectures treat multi-person gaze as a set-prediction problem, yet often rely on a conditional independence fallacy. For instance, several frameworks [36, 28, 33] process individual heads in isolation or suffer from noise-prone padding due to rigid architectural constraints. Even methods that incorporate auxiliary social tasks [8, 9] remain limited to coarse, pairwise categories, failing to capture the continuous, joint-distributional nature of social behavior. In contrast, ANCHOR explicitly decodes joint dependencies by modeling $P(X_{gaze}, X_{social}|X_{scene})$ through a flexible, padding-free framework.

Social Interaction and Group Detection. Traditional research identifies explicit social groups via discrete labels, ranging from early F-formations [29, 1] to categorical IDs in recent datasets like JRDB-Social [15]. While effective for high-level classification, these oversimplified, binary labels suffer from subjective annotation bias and fail to capture the continuous subtleties of implicit social intent. ANCHOR diverges by rejecting discrete labels to decode latent social dynamics directly from gaze features. This paradigm is validated through our novel Social Influence Ranking (SIR) metric, which transforms subjective social perceptions into a continuous, quantifiable measure of intent.

Gaze Following Datasets. GazeFollow [24] established the foundation for large-scale gaze detection, with subsequent datasets addressing temporal dynamics (VAT [3]), shared attention (VideoCoAtt [4]), object context (VACATION [5], GOO [34]), and specialized domains [32, 12]. However, these lack the comprehensive multi-person annotations and quantitative ranking labels necessary for joint-distributional modeling. Existing multi-person works often rely on subjective binary labels (e.g. LAH, SA, LAEO), which fail to capture the nuanced spectrum of social influence. To bridge this gap, our Extended GazeFollow benchmark introduces: (1) multi-person gaze annotations and (2) Social Influence Ranking (SIR) labels. This allows us to move beyond discrete classification toward a fine-grained, quantitative understanding of implicit social dynamics.

3 Extended GazeFollow Benchmark

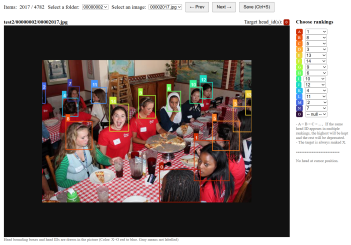
While existing datasets lack the quantitative social measures required for decoding (Fig. 3c) , the original GazeFollow [24] offers a uniquely diverse range of multi-person social scenarios (Fig. 3b). To bridge this gap, we enriched its 3,018 test images with two novel layers of manual annotations.

Multi-Head Gaze Annotations. To evaluate true multi-head performance, we provide comprehensive gaze annotations for *all* valid heads in all 3,018 test images, using refined head boxes from [33]. We deprecated heads with minimal

Table 1: **GazeFollow Benchmark Comparison.**

Method	Multi-person Gaze			Single-person Gaze		
	AUC ↑	Min L2 ↓	Avg L2 ↓	AUC ↑	Min L2 ↓	Avg L2 ↓
Recasens et al.[24]	-	-	-	0.878	0.113	0.19
Chong et al.[3]	-	-	-	0.921	0.077	0.137
Hu et al.[14]	-	-	-	0.923	0.069	0.128
Gupta et al.[10]	-	-	-	0.943	0.056	0.114
Tafasca et al.[33]	0.917	0.169	0.186	0.944	0.057	0.113
Gaze-LLE (B)[27]	0.918	0.190	0.205	0.956	0.045	0.104
Gaze-LLE (L)[27]	0.921	0.183	0.197	0.958	0.041	0.099
ANCHOR (B)	0.928	0.181	0.197	0.955	0.052	0.112
ANCHOR (L)	**0.932**	**0.173**	**0.188**	**0.956**	**0.047**	**0.106**

Fig. 2: **UI of Social Influence Ranking Labeling.**



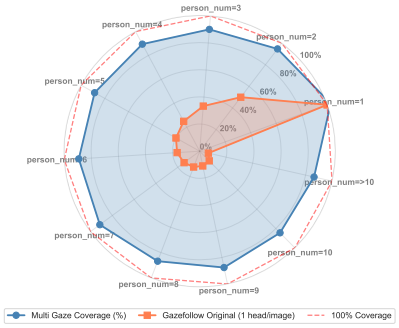
information ($area < 4$ px^2). This resulted in 17,313 total labels, providing dense supervision across complex social scenes (36% of images contain > 3 heads). As shown in Fig. 3a, our benchmark provides over 86% label coverage for images with many heads, correcting a key deficiency in the original sparse dataset.

Social Influence Ranking (SIR) Labels. To create a ground truth for the latent variable X_{social}, we introduce SIR labels. We engaged 4 human annotators to rank all surrounding individuals based on their perceived social influence on the primary target. This ordinal ranking (Top-15) was successfully applied to all 1,428 test images containing sufficient social context ($>= 2$ neighbors).

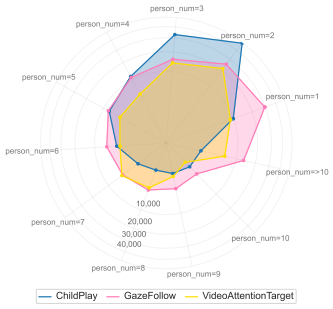
Annotation Procedures and Annotator Agreement. Since GazeFollow typically annotates only the most salient head in each image, we treat this original head as the target subject. Based on this target subject, we first add multi-head gaze annotations: we remove incorrectly detected, overly small, or visually unclear head boxes, and then annotate gaze points for each valid head using the same protocol as the original dataset. We also use the same target subject for Social Influence Ranking (SIR), using the interface shown in Fig. 2. Annotators assign an A–O influence level (ranging from highest to lowest social impact) to every other person in the image using only visual cues such as body orientation, depth-adjusted distance, and simple signs of interaction. Each person receives a unique rank, and no gaze labels or geometric metadata are provided. All four annotators follow the same instructions, and although small differences appear in fine-grained cases, the overall patterns are consistent. We merge their labels into a single consensus ranking, which serves as the final ground truth.

4 Methodology

Problem Definition. Given an RGB image $\mathcal{I}$ and K head bounding boxes $\mathbf{B}$, we aim to predict a person-specific gaze heatmap $\mathbf{H}_k \in [0,1]^{H\times W}$ for every head k. Our premise is that accurate prediction requires modeling the joint distribution $P(\mathcal{H}, \mathbf{Z}|\mathcal{I}, \mathbf{B})$ by explicitly surfacing the latent social intent $\mathbf{Z}$ linking individual behaviors.



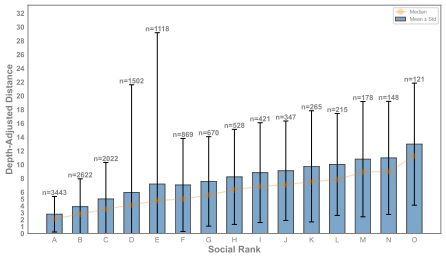
(a) Multi-gaze label coverage.



(b) Datasets coverage comparison.

Dataset	Multi-head Gaze	Social Impact
GazeFollow [24]	✗	✗
VideoGaze [25]	✗	✗
VideoCoAtt [4]	✓	✗
Gaze360 [16]	✓	✗
VideoAttentionTarget [3]	✓	✗
GazeFollow360 [17]	✓	✗
GOO [34]	✗	✗
ChildPlay [32]	✓	✗
Ours (GF test ext.)	✓	✓

(c) Comparison of existing gaze datasets.



(d) Distances across social ranks.

Fig. 3: **Dataset Characteristics:** (a) Gaze annotation coverage across head-number groups. (b) Coverage comparison of existing multi-person gaze datasets. (c) Comparison of dataset annotation properties with prior work. Ours is the extended GazeFollow test split. (d) Average depth-adjusted distances across social ranks ($\pm\sigma$).

4.1 Model Architecture

ANCHOR(Fig.1) integrates four primary components: a frozen backbone, a relational attention mechanism, a synergistic optimization strategy, and an evaluation protocol.

Frozen Encoder. ANCHOR employs a frozen DINOv2 backbone to extract shared scene features from image $\mathcal{I}$. Using the K head bounding boxes in $\mathbf{B}$, we extract head-specific features $\mathbf{c}_k$ via masked average pooling over the corresponding spatial regions. These features are projected to a 256-dimensional space and, along with the scene tokens, processed by three stacked ViT blocks to yield refined global scene tokens $\mathbf{S}$.

Relational Attention Mechanism. Given the extracted features $\mathbf{c}_k$ and global tokens $\mathbf{S}$, ANCHOR decodes the latent social intent $\mathbf{Z}$ by first identifying the local social context $\mathbf{T}$ for a target head k. This context is formed by selecting the top-N neighbors via a depth-adjusted distance $\ell_{adj} = \|\mathbf{p}_k - \mathbf{p}_j\|_2 / \sqrt{A_j + \epsilon}$,

where $\mathbf{p}_i$ denotes the normalized image-plane center of the i_{th} head box and A_j is the bounding box area. The area normalization incorporates monocular scale cues into image-plane proximity: for candidates with similar center distances, a larger head box yields a smaller adjusted distance and is therefore treated as a closer local neighbor. We collect these distances into a geometric pseudo-label vector $\mathbf{D}_{inv} = [\ell_{adj}(k,j)^{-1}]_{j\in\mathcal{N}_k}$, representing the inverse proximity to neighbors. We then employ a cross-attention block, using $\mathbf{c}_k$ as the query and $\mathbf{T}$ as the key-value pairs, to model the interplay between the individual and their surroundings. The output of this block, $\mathbf{X}^k_{social}$, represents social-aware features, while its internal attention weights $\mathbf{W}_{attn}$ serve as the explicit proxy for $\mathbf{Z}$. Finally, a FiLM module [23] modulates the scene tokens $\mathbf{X}^k_{scene}$ using parameters (γ_k, β_k) derived from the social-aware features $\mathbf{X}^k_{social}$, yielding the socially-consistent feature map:

$$\mathbf{X}^k_{scene} = (1+\gamma_k)\mathbf{X}^k_{social} + \beta_k \tag{1}$$

Joint Decoding. ANCHOR yields a dual output for each target head: (1) the person-specific gaze heatmap $\mathbf{H}_k$, and (2) the latent social intent $\mathbf{Z}$. While $\mathbf{H}_k$ is produced by refining and upsampling the modulated features $\mathbf{X}^k_{scene}$ to 64×64 resolution, the social intent $\mathbf{Z}$ is explicitly represented by the attention weights $\mathbf{W}_{attn}$ extracted during the relational mechanism. This joint formulation enables the model to simultaneously recover spatial gaze targets and the underlying social net.

4.2 Loss Design

Our training objective is designed to maximize predictive accuracy for the primary gaze task while simultaneously enforcing the learning of latent social net in a stable manner.

Target Gaze Loss $\mathcal{L}_T$. The primary supervisory signal for the gaze task is the pixel-wise Binary Cross-Entropy (BCE) loss applied to the target head's heatmap prediction. The ground-truth heatmap is generated by placing a 2D Gaussian distribution (with $\sigma = 3$) around the target gaze annotation. We define the target gaze loss $\mathcal{L}_T$ as the mean BCE over all target heads in the batch, $\mathcal{M}_T$:

$$\mathcal{L}_T = \frac{1}{|\mathcal{M}_T|}\sum_{k\in\mathcal{M}_T}\mathrm{BCE}(\mathbf{H}_k, \mathbf{H}^{GT}_k) \tag{2}$$

Self-Supervised Social Consistency Loss $\mathcal{L}_S$. To explicitly train the Social Cross-Attention block to decode the local social prior, we introduce the self-supervised Social Consistency Loss $\mathcal{L}_S$, which maximizes the correlation between the learned mean attention weights $\mathbf{W}_{attn}$ and the geometric pseudo-label $\mathbf{D}_{inv}$ defined in Sec. 4.1:

$$\mathcal{L}_S = -\mathrm{Corr}(\mathbf{W}_{attn}, \mathbf{D}_{inv}) \tag{3}$$

The final objective is $\mathcal{L}_{total} = \mathcal{L}_T + \lambda_s\mathcal{L}_S$. To mitigate potential gradient conflicts between these tasks, we employ the synergistic optimization strategy.

4.3 Synergistic Optimization Strategy

The joint modeling of gaze and social intent introduces a significant optimization challenge: gradient conflicts where $\nabla\mathcal{L}_S$ can degrade the primary gaze task performance. We resolve this via a two-stage optimization stack.

Stage 1: Robustness via SAM. To ensure stable generalization and enforce a flat local minimum for the primary gaze task, we apply Sharpness-Aware Minimization (SAM) [7] exclusively to the gaze objective $\mathcal{L}_T$. The robust gaze gradient $\mathbf{g}_T^{SAM}$ is derived by solving for the optimal neighborhood perturbation ϵ:

$$\mathbf{g}_T^{SAM} = \nabla_\theta \mathcal{L}_T(\theta + \epsilon), \quad \text{where } \epsilon = \arg\max_{\|\epsilon\| \leq \rho} \mathcal{L}_T(\theta + \epsilon) \tag{4}$$

Stage 2: Synergy via Gradient Projection. To mitigate interference, we utilize the Projecting Conflicting Gradients (PCGrad) [38] principle. We treat $\mathbf{g}_T^{SAM}$ as the anchor; if the social gradient $\mathbf{g}_S$ conflicts with it ($\mathbf{g}_T^{SAM} \cdot \mathbf{g}_S < 0$), $\mathbf{g}_S$ is projected orthogonally onto the normal plane of $\mathbf{g}_T^{SAM}$:

$$\mathbf{g}_S^{proj} = \mathbf{g}_S - \frac{\mathbf{g}_T^{SAM} \cdot \mathbf{g}_S}{\|\mathbf{g}_T^{SAM}\|_2^2} \mathbf{g}_T^{SAM} \tag{5}$$

The final gradient update $\mathbf{g}_{final}$ combines the robust gaze anchor and the conflict-free social component, ensuring that relational learning only proceeds in directions that do not degrade gaze accuracy:

$$\mathbf{g}_{final} = \mathbf{g}_T^{SAM} + \lambda_s \cdot \mathbf{g}_S^{proj} \tag{6}$$

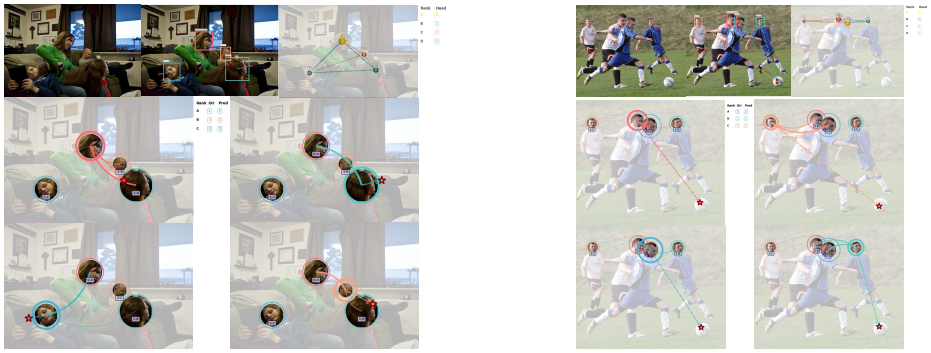
5 Experiments

Our evaluation focus on three aspects: (i) the target-centric architecture's superiority over global-map models, (ii) the synergistic optimization strategy in resolving task-level gradient conflicts, and (iii) the quantitative decoding of latent social intent. Results demonstrate that ANCHOR achieves state-of-the-art static multi-gaze accuracy while successfully recovering underlying social net.

5.1 Evaluation Metrics

To validate the framework's performance, we use standard gaze metrics and introduce novel social evaluation metrics enabled by our test set extensions.

Gaze Prediction Metrics. We evaluate gaze prediction accuracy using three metrics: AUC (Area Under the Curve), Min L2 Distance (minimum Euclidean distance between predicted maximum and all ground-truth points), and Avg L2 Distance (Euclidean distance between predicted maximum and the mean ground-truth point). For all metrics, we report performance in two scopes: the Target Gaze metric, which aligns with the standard GazeFollow benchmark by focusing only on the single designated person per image, and the comprehensive Multi



(a) (b)

Fig. 4: **Visualisation of multi-person gaze prediction and decoded social net.**

Gaze metric ('_multi' suffix) validating the joint-distributional model across all annotated heads in the scene.

Integrated Social-Relational Metrics. To quantify the recovery of both geometric priors and perceived social ranking, we introduce three complementary metrics: Social Consistency Index (SCI), Social Leader Accuracy (SLA), and Social Influence Rank Correlation (ρ_{SIR}). While the observed negative correlation between social impact and distance l_{adj} (Fig. 3d) justifies our $\mathcal{L}_{SCI}$ prior, high variance indicates that social intent is a complex latent variable exceeding mere geometry. This underscores the necessity of our multi-faceted evaluation—using SCI to validate internal geometric mechanisms, SLA to measure practical leader recognition, and ρ_{SIR} to assess the alignment with the full human-annotated hierarchy. Metrics are calculated per target head k and averaged over the dataset.

1. Social Consistency Index (SCI). This index validates the relational mechanism by measuring how attention weights W_{attn} align with geometric proximity priors. For a target k, we compute the Pearson correlation [22] between the attention vector $\mathbf{w}^{(k)}$ and the inverse distance vector $\mathbf{d}_{\text{inv}}^{(k)} = [\ell_{\text{adj}}(k,n)^{-1}]_{n\in\mathcal{N}_k}$:

$$\text{SCI}^{(k)} = \text{Corr}_{\text{Pearson}}\left(\mathbf{w}^{(k)}, \mathbf{d}_{\text{inv}}^{(k)}\right). \tag{7}$$

A high SCI indicates that the model successfully learns the linear proportionality hypothesis ($W_{\text{attn}} \propto \ell_{\text{adj}}^{-1}$) defined in our self-supervised objective.

2. Social Leader Accuracy (SLA). This metric assesses the model's ability to identify the most salient individual (the "leader") in a social scene. We define the prediction as correct if the neighbor n^* receiving the maximum attention weight belongs to the highest human-annotated social ranking $\mathcal{R}_{gt}top1$:

$$\text{SLA} = \frac{1}{|\mathcal{M}|}\sum_{k\in\mathcal{M}} \mathbb{I}\left(R\left(\text{argmax}_{j\in\mathcal{N}k} w_j^{(k)}\right) \in \mathcal{R}_{gt}^{top1}\right). \tag{8}$$

SLA provides a discrete validation of the model's practical social reasoning capability in identifying centers of attention.

3. Social Influence Rank Correlation (ρ_{SIR}). To evaluate whether ANCHOR captures the full human-perceived social hierarchy, we use Spearman's [30] rank correlation. This compares the model-derived ordering of neighbors against the human-annotated ordinal ranks $\mathcal{R}_{\text{gt}}^{(k)}$:

$$\rho_{\text{SIR}}^{(k)} = \text{Corr}_{\text{Spearman}} \left(\mathcal{R}(\mathbf{w}^{(k)}), \mathcal{R}_{\text{gt}}^{(k)} \right). \tag{9}$$

Unlike social leader accuracy, ρ_{SIR} is invariant to scale and measures the monotonic agreement across the entire social graph, confirming the model's depth in decoding complex relations beyond simple geometry.

5.2 Implementation Details

We train on the standard GazeFollow dataset [24] and evaluate on our Extended GazeFollow Benchmark. All models use a frozen DINOv2 backbone and a consistent latent dimension of $C = 256$. Our final model is trained for 15 epochs.

5.3 Qualitative Analysis

Fig. 4 illustrates ANCHOR coupling gaze prediction with target-aware social reasoning. In the first row, multi-gaze heatmaps align with the influence graph, where shared targets and proximity correlate with higher ranks and thicker connections. The model moves beyond simple distance priors, leveraging joint gaze direction and body orientation to assign social influence. The second row shows ANCHOR correctly recovering top influential neighbours (ranking A, B, C) even among equidistant candidates. While minor discrepancies occur in lower ranks where human labels are less sharp, re-targeting rows demonstrate plausible graph reorientation toward new focal subjects. These results confirm a flexible, target-centric social representation that successfully captures the latent structural scaffolding of gaze behavior.

5.4 Quantitative Analysis

Comparison to State-of-art Methods We compare ANCHOR with state-of-the-art methods on two tasks(Tab 1): (1) the full Multi Gaze task, which evaluates true joint-distributional performance, and (2) the standard single-person Target Gaze task. On Multi Gaze (Table 1), ANCHOR (DINOV2-L) achieves an AUC of 0.932, Min L2 of 0.173, Avg L2 of 0.188, surpassing the best i.i.d. model, Gaze-LLE (L) (AUC=0.921, Min L2=0.183, Avg L2=0.197). This demonstrates the empirical cost of the conditional independence assumption. The marginal gap between ANCHOR and Gaze-LLE on Target Gaze is expected. While Gaze-LLE focuses solely on i.i.d. pixel-level accuracy, ANCHOR is designed to decouple

Table 2: **Model ablation study.** Comparison of different model variants on **GazeFollow**. Multi Gaze columns measure joint consistency across people, while Single Gaze columns evaluate individual prediction quality.

Backbone	Ablation Configuration				Multi Gaze			Single Gaze		
	w/o FiLM	w/o Cross Attn	w/o L_{SCI}	w/o TA-RO	AUC ↑	Min L2 ↓	Avg L2 ↓	AUC ↑	Min L2 ↓	Avg L2 ↓
DINOv2(B)	✗	✗	✗	✗	0.916	0.193	0.206	0.955	0.048	0.108
DINOv2(B)	✓	✗	✗	✗	0.916	0.215	0.228	0.949	0.050	0.109
DINOv2(B)	✓	✓	✗	✗	0.916	0.215	0.228	0.949	0.051	0.109
DINOv2(B)	✓	✓	✓	✗	0.920	0.214	0.226	0.951	0.049	0.108
DINOv3 (B)	✓	✓	✓	✓	0.915	0.223	0.235	0.951	0.051	0.110
DINOv3 (L)	✓	✓	✓	✓	0.894	0.246	0.263	0.932	0.096	0.162
ViT(S)	✓	✓	✓	✓	0.852	0.313	0.328	0.887	0.150	0.222
CNN(ResNet50)	✓	✓	✓	✓	0.830	0.329	0.344	0.876	0.156	0.230
DINOv2(B)	✓	✓	✓	✓	0.928	0.181	0.197	0.955	0.052	0.112
DINOv2(L)	✓	✓	✓	✓	**0.932**	**0.173**	**0.188**	**0.956**	**0.047**	**0.106**

high-quality social latent variables from complex scene contexts. This necessitates a trade-off: we accept a minimal sacrifice in individual AUC to achieve jointly-optimized, socially-consistent predictions, effectively trading narrow task-specific precision for meaningful social-relational understanding.

Ablation Study We conduct a series of ablations to validate our two primary claims: the necessity of the ANCHOR architecture and the synergistic optimization stack.

Backbone Selection Analysis: We summarize here only the key outcome of our backbone study. Frozen DINOv2 backbones consistently provide the most stable and accurate performance, while partial or full fine-tuning leads to severe degradation. Detailed results, comparisons with DINOv3, and the full backbone selection analysis are provided in the supplementary material.

Component Ablation Analysis: Table 2 reveals the progressive emergence of joint-distributional capability in our model. The frozen DINOv2 baseline provides strong single-head performance but fails to capture multi-person consistency, indicating that backbone strength alone is insufficient. Simply adding FiLM or the Social Cross-Attention block does not yield meaningful gains and can even degrade results, demonstrating that naive architectural modifications cannot resolve the core challenge. Substantial improvement first appears when introducing the self-supervised $\mathcal{L}_{SCI}$ objective, which encourages the model to internalize geometric-social relations and raises multi-head accuracy accordingly. The decisive transition, however, comes from the full synergistic optimization strategy, which stabilizes optimization and prevents the collapse observed in earlier variants, enabling the model to jointly optimize social reasoning and gaze localization. Scaling the backbone to ViT-L further strengthens performance. Overall, the ablation shows that $\mathcal{L}_{SCI}$ provides the necessary supervisory signal for extracting latent social structure, while synergistic optimization strategy is essential for reliably integrating these signals within the shared representation space, making both components indispensable to the final ANCHOR architecture.

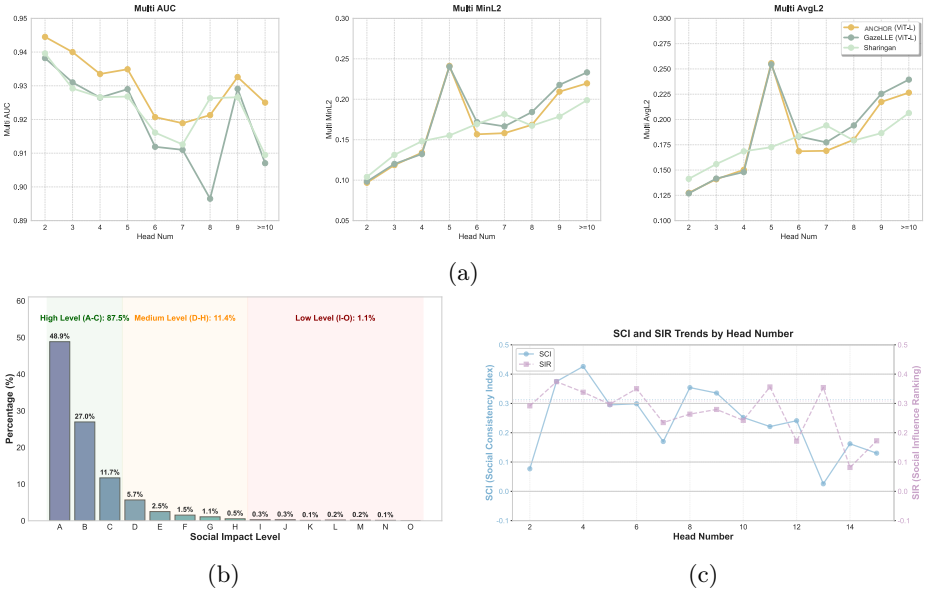


Fig. 5: **Comprehensive Model Performance and Social Analysis:** (a) Gaze following performance comparison (AUC and L_2 distance) across varying head numbers for ViT-B and ViT-L backbones. (b) Detailed analysis of Social Leader Accuracy (SLA) for Top-1 High-Level neighbor prediction across different social tiers. (c) Measurements of Social Consistency Index (SCI) and Social Influence Ranking (ρ_{SIR}) metrics across impact levels.

5.5 Qualitative and Social Analysis

Finally, we analyze our model's performance in complex scenes and its ability to decode social intent.

Robustness in Complex Scenes: Figure 5a demonstrates ANCHOR's robustness. Our model (ViT-L, yellow) consistently outperforms the SOTA i.i.d. model Gaze-LLE (ViT-L, gray) on Multi Gaze AUC across all head counts. The performance gap *widens* in high-density scenes (e.g., 8, 10, and 15 heads), proving our joint-distributional model is superior at handling the complex dependencies that i.i.d. models ignore.

Quantitative Proof of Social Intent: We evaluate the decoded social vector $\mathbf{W}_{\text{attn}}$ using our three social metrics. **SCI** provides mechanism-level validation: a positive mean score (0.282) shows that the SCA block successfully learns the geometric prior encouraged by $\mathcal{L}_{SCI}$. **SLA** (Top-1 High-Level Accuracy) offers the clearest practical evidence of decoded intent: ANCHOR (ViT-L) achieves an **88.2%** accuracy in identifying the highest-tier social neighbor, demonstrating strong leader recognition. Finally, $\rho_{\mathbf{SIR}}$ captures the full social hierarchy; a significant positive correlation (0.280) shows the model recovers not only the leader but the continuous ranking structure perceived by humans.

6 Discussion

Our experiments validate ANCHOR's key claims. First, single-target evaluation hides the challenge of joint gaze reasoning: i.i.d.-optimized baselines such as Gaze-LLE perform well on Single Gaze but drop on Multi Gaze, whereas ANCHOR (ViT-L) reaches 0.932 AUC and stays robust in complex scenes. Second, ablations show FiLM and SCA offer little alone; gains arise only with $\mathcal{L}_{SCI}$, and stable performance requires TA-RO, confirming that exposing X_{social} creates a gradient conflict needing robust, projection-aware optimization. Third, our social metrics validate decoded intent: SCI captures geometric priors, SLA (88.2%) finds the most influential neighbor, and ρ_{SIR} aligns with human rankings. These results provide the first quantitative evidence that models can infer latent social structure from gaze cues. Remaining limitations include dependence on a frozen backbone, tunable optimization weights, and the focus on static images, suggesting future extension to temporal settings.

7 Conclusion

In this work, we address the conditional-independence flaw in gaze-following research by modeling the joint distribution of gaze and latent social intent. We introduce ANCHOR, which integrates four primary components: (1) a frozen backbone for robust feature extraction; (2) a relational attention mechanism to decode social context; (3) a synergistic optimization strategy to resolve gradient conflicts; and (4) a valuation protocol featuring the Extended GazeFollow Benchmark. Our experiments demonstrate that ANCHOR achieves state-of-the-art multi-head performance while proving that the synergistic optimization strategy is essential for stable training. Furthermore, through our SCI, SLA and ρ_{SIR} metrics, we provide the first quantitative evidence that implicit social dynamics can be reliably decoded from static gaze patterns.

References

1. Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., Lanz, O., Sebe, N.: Salsa: A novel dataset for multimodal group behavior analysis. IEEE transactions on pattern analysis and machine intelligence **38**(8), 1707–1720 (2015)
2. Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In: Proceedings of the European conference on computer vision (ECCV). pp. 383–398 (2018)
3. 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)
4. 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)

5. 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 International Conference on Computer Vision. pp. 5724–5733 (2019)
6. Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., Zhai, G.: Dual attention guided gaze target detection in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11390–11399 (2021)
7. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: International Conference on Learning Representations (2021)
8. Gupta, A., Tafasca, S., Chutisilp, N., Odobez, J.M.: A unified model for gaze following and social gaze prediction. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–9. IEEE (2024)
9. Gupta, A., Tafasca, S., Farkhondeh, A., Vuillecard, P., Odobez, J.m.: Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. Advances in Neural Information Processing Systems **37**, 15646–15673 (2025)
10. 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)
11. Horanyi, N., Zheng, L., Chong, E., Leonardis, A., Chang, H.J.: Where are they looking in the 3d space? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2677–2686 (2023)
12. Hou, Y., Zhang, Z., Horanyi, N., Moon, J., Cheng, Y., Chang, H.J.: Multi-modal gaze following in conversational scenarios. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1186–1195 (2024)
13. Hu, Z., Yang, D., Cheng, S., Zhou, L., Wu, S., Liu, J.: We know where they are looking at from the rgb-d camera: Gaze following in 3d. IEEE Transactions on Instrumentation and Measurement **71**, 1–14 (2022)
14. 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)
15. Jahangard, S., Cai, Z., Wen, S., Rezatofighi, H.: Jrdb-social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22087–22097 (2024)
16. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6912–6921 (2019)
17. Li, Y., Shen, W., Gao, Z., Zhu, Y., Zhai, G., Guo, G.: Looking here or there? gaze following in 360-degree images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3742–3751 (October 2021)
18. Mayrand, F., McCrackin, S.D., Ristic, J.: Intentional looks facilitate faster responding in observers. Communications Psychology **2**, 90 (2024). https://doi.org/10.1038/s44271-024-00137-x
19. Mayrand, F., Capozzi, F., Ristic, J.: Gaze communicates both cue direction and agent mental states. Frontiers in Psychology **15**, 1472538 (2024)
20. Myers, D., Abell, J., Sani, F.: EBook: Social Psychology 3e. McGraw Hill (2020)
21. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalid, W., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

22. Pearson, K.: Notes on regression and inheritance in the case of two variables. Proceedings of the Royal Society of London **58**, 240–242 (1895)
23. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
24. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? Advances in neural information processing systems **28** (2015)
25. Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze across views. arXiv preprint arXiv:1612.03094 (2016)
26. Ristic, J., Capozzi, F.: Mechanisms for individual, group, and crowd-based attention to social information. Nature Reviews Psychology (2022). https://doi.org/10.1038/s44159-022-00118-z
27. Ryan, F., Bati, A., Lee, S., Bolya, D., Hoffman, J., Rehg, J.M.: Gaze-lle: Gaze target estimation via large-scale learned encoders. arXiv preprint arXiv:2412.09586 (2024)
28. Ryan, F., Bati, A., Lee, S., Bolya, D., Hoffman, J., Rehg, J.M.: Gaze-lle: Gaze target estimation via large-scale learned encoders. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28874–28884 (2025)
29. Setti, F., Russell, C., Bassetti, C., Cristani, M.: F-formation detection: Individuating free-standing conversational groups in images. PloS one **10**(5), e0123783 (2015)
30. Spearman, C.: The proof and measurement of association between two things. The American Journal of Psychology **15**(1), 72–101 (1904)
31. Sumer, O., Gerjets, P., Trautwein, U., Kasneci, E.: Attention flow: End-to-end joint attention estimation. In: The IEEE Winter Conference on Applications of Computer Vision. pp. 3327–3336 (2020)
32. 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)
33. Tafasca, S., Gupta, A., Odobez, J.M.: Sharingan: A transformer architecture for multi-person gaze following. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2008–2017 (2024)
34. Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R., Guinto, R.: Goo: A dataset for gaze object prediction in retail environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3125–3133 (2021)
35. Tonini, F., Dall'Asen, N., Beyan, C., Ricci, E.: Object-aware gaze target detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21860–21869 (2023)
36. 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)
37. Wei, P., Liu, Y., Shu, T., Zheng, N., Zhu, S.C.: Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6801–6809 (2018)
38. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in neural information processing systems **33**, 5824–5836 (2020)