

# Beyond Gaze Overlap: Analyzing Joint Visual Attention Dynamics Using Egocentric Data

Kumushini Thennakoon, Yasasi Abeysinghe, Bhanuka Mahanama, Vikas Ashok, Sampath Jayarathna  
Department of Computer Science, Old Dominion University, Norfolk, VA, USA  
kumushini@cs.odu.edu, yasasi@cs.odu.edu, bhanuka@cs.odu.edu, vganjigu@cs.odu.edu, sampath@cs.odu.edu

**Abstract**—Joint visual attention (JVA) provides informative cues on human behavior during social interactions. The ubiquity of egocentric eye-trackers and large-scale datasets on everyday interactions offer research opportunities in identifying JVA in multi-user environments. We propose a novel approach utilizing spatiotemporal tubes centered on attention rendered by individual gaze and detect JVA using deep-learning-based feature mapping. Our results reveal object-focused collaborative tasks to yield higher JVA (44-46%), whereas independent tasks yield lower (4-5%) attention. Beyond JVA, we analyze attention characteristics using ambient-focal attention coefficient  $\mathcal{K}$  to understand the qualitative aspects of shared attention. Our analysis reveals  $\mathcal{K}$  to converge instances where participants interact with shared objects while diverging when independent. While our study presents seminal findings on joint attention with egocentric commodity eye trackers, it indicates the potential utility of our approach in psychology, human-computer interaction, and social robotics, particularly in understanding attention coordination mechanisms in ecologically valid contexts.

**Index Terms**—Joint visual attention, Egocentric data, Eye-tracking

## I. INTRODUCTION

The ability to understand where and how individuals direct their visual attention during social interactions provides details about the underlying cognitive processes, social cues, and coordination strategies that shape mutual understanding [1]. Observing moments of Joint Visual Attention (JVA) during different social interactions between two or more individuals is a well-established method in developmental psychology [2], [3]. JVA is used in language acquisition research with children and in the early detection of autism [4]. Beyond psychology, JVA is widely used in various domains, including human-computer interaction [5], [6], education, neuroscience [7], and social robotics [8]. Traditional methods relied on qualitative approaches (e.g., gaze following [9], pointing) to analyze JVA. However, with advances in eye-tracking technology in conjunction with wearable eye-tracking, it has become easier for researchers to collect accurate gaze data, making the study of JVA more efficient and objective [10], [11].

Wearable eye-tracking technology has emerged as a powerful tool to capture egocentric visual behavior in real-world environments [12]–[15]. These devices provide a mobile and unobtrusive means of collecting multi-modal data from the wearer’s point of view. Although much research on JVA has been conducted in controlled laboratory settings, only a few studies have focused on naturalistic, dynamic environments

with more inherited constraints when studying the visual attention of multiple individuals.

In this work, we present an analysis of JVA using egocentric video data and gaze data collected using wearable eye-tracking glasses. By leveraging the mobility and naturalistic recording capabilities of the devices, we aim to explore how JVA can be detected and interpreted in real-world social interactions. Our research contributes to ongoing efforts to bridge the gap between controlled experimental designs and the complexity of real-world behavior. We used the advance gaze metric ambient-focal attention coefficient  $\mathcal{K}$  [16]–[18] to observe the attention patterns of the participants. Data were obtained from the publicly available Aria everyday activities dataset [19] released by Meta as part of their larger vision for developing Augmented Reality (AR) and Artificial Intelligence (AI) technologies. This dataset was collected using Project Aria glasses [20] during various daily activities such as cooking, watching television, making coffee, and engaging in conversations.

## II. RELATED WORK

JVA has been studied across multiple disciplines with varying approaches and objectives. This section reviews relevant literature that informs our approach to analyzing joint visual attention in everyday activities using egocentric data.

Traditionally, JVA has been a central concept in developmental psychology, where it has been extensively used to understand typical development patterns and identify developmental conditions like autism spectrum disorder [2], [4], [21]. These studies relied mainly on qualitative observational methods [22], [23] to assess when and how individuals coordinate their attention to the same object or event. While qualitative approaches provided valuable insights into the social aspects of attention, they lacked precision in measuring the exact properties of visual attention.

Advancements in eye-tracking technology have enabled researchers to measure JVA with higher precision. Schneider et al. [1] proposed quantifying JVA in collaborative environments based on gaze overlapping. However, this simplified approach tends to reduce the complex nature of joint attention to a binary measurement (overlap or no overlap), overlooking the rich temporal and qualitative characteristics of shared attention. Our work aims to provide details beyond gaze overlap metrics to develop a more sophisticated approach to observing and understanding JVA.

Several studies have explored the identification of JVA using egocentric data from wearable cameras or eye trackers [24]–[26]. However, these studies have focused only on identifying instances of JVA but analyzing the attention behaviors of multiple participants simultaneously. Our research extends previous work by not only identifying moments of joint attention but also analyzing patterns and quantifying the percentage of JVA in everyday activities involving two participants.

Object detection has emerged as a promising approach for identifying JVA in egocentric settings [24], [27], particularly in controlled environments with limited sets of objects. However, this approach faces significant challenges when applied to everyday activities where participants interact with numerous diverse objects that may not be easily recognizable by standard object detection algorithms. Moreover, our primary goal is not to develop new methods for identifying JVA but rather to employ established methods of identification while focusing on deeper analysis of attention patterns.

Researchers have attempted to localize attention by analyzing the angle of direction from multiple video sources [28], [29]. While this approach provides valuable spatial information, it often lacks the precision offered by eye-tracking data. Eye tracking provides more accurate information about the exact focus of attention [16], which is particularly important when analyzing fine-grained attentional behaviors in everyday activities where the objects of interest may be in close proximity to one another.

Our method builds upon the idea of work by Kera et al. [26] where they create a spatiotemporal tube by extracting regions of interest around the gaze position from egocentric video frames and calculating similarity between the defined tubes to identify instances of JVA. However, we extend this approach by incorporating advanced gaze measures to analyze the identified instances of JVA. This additional analytical step allows us to move beyond simple identification to understand the patterns of joint attention in everyday activities captured through egocentric recordings. By addressing these research gaps, our work contributes to a more comprehensive understanding of JVA in naturalistic settings, with implications for both theoretical models of social attention and practical applications in fields such as education,

### III. METHODOLOGY

#### A. Dataset

For this study, we utilized the publicly available Aria Everyday Activities dataset [19], which provides egocentric recordings of participants engaged in common daily activities. This dataset contains time-synchronized video data captured from a first-person perspective as individuals perform routine tasks such as cooking, making coffee, watching television, etc. In addition to video, the dataset provides multi-modal sensor data recorded using Project Aria glasses [20], including per-frame 3D eye gaze directional vectors. The dataset is particularly valuable for our research as it includes dual-participant recordings where two individuals interact in the same environment while both wearing Project Aria glasses.

This synchronized dual-perspective video data along with gaze data allows us to examine JVA as it naturally occurs between pairs of participants during routine interactions. In this study, we used the dual-participant recordings from the dataset to analyze how individuals coordinated their visual attention across various everyday tasks, enabling the investigation of JVA patterns in real-world settings.

The Aria Everyday Activities dataset comprises 143 sequences of daily activities recorded by multiple participants across five geographically diverse indoor locations. However, most of these recordings involve only a single participant or missing recordings from one of the two participants. Since our focus is on dual-participant interactions, we excluded such data, resulting in a subset of 10 recordings. Among these, 5 recording sessions were further excluded due to highly dynamic motion, face-to-face conversations without shared object interactions, or a lack of common visual perspectives. After this filtering process, we retained 5 recordings for analysis. In the selected sessions, five participants, paired in varying combinations, took part in the five distinct activity sessions.

The data for each dual-participant task were processed through the pipeline shown in Figure 1 to facilitate the analysis of JVA.

#### B. Extracting Regions of Interest (ROIs) around the Gaze Position

To analyze JVA between dual participants, we needed to identify their shared visual perspective during the task. We began by decomposing the egocentric video streams into individual frames. Using Project Aria tools [19], we converted the per-frame 3D eye gaze directional vectors into 2D gaze coordinates and projected these coordinates onto the corresponding video frames.

To detect moments of shared gaze, we initially compared the full image frames between participants. However, this approach yielded low similarity scores due to the presence of many unrelated objects in the scene. To better capture the visual content near the gaze point, we adopted a method based on spatiotemporal tubes around points of gaze [26], which allows for extracting and comparing the specific regions attended to by each participant.

From the gaze-annotated frames, we extracted a  $400 \times 400$  pixel region centered on each participant’s gaze point (see Figure 2). This window size of 400 pixels was empirically chosen to cover more than 25% of both the width and height of each frame (resolution:  $1408 \times 1408$ ), effectively capturing the focal area of visual attention while reducing peripheral visual noise. Assembled over time, these gaze-centered ROIs formed what we refer to as spatiotemporal tubes, enabling a more precise analysis of shared attention.

#### C. Calculating Similarity Between Spatiotemporal Tubes

In this study, we assess the visual similarity between two sets of image frames, referred to as spatiotemporal tubes using a deep learning-based feature extraction approach. Each

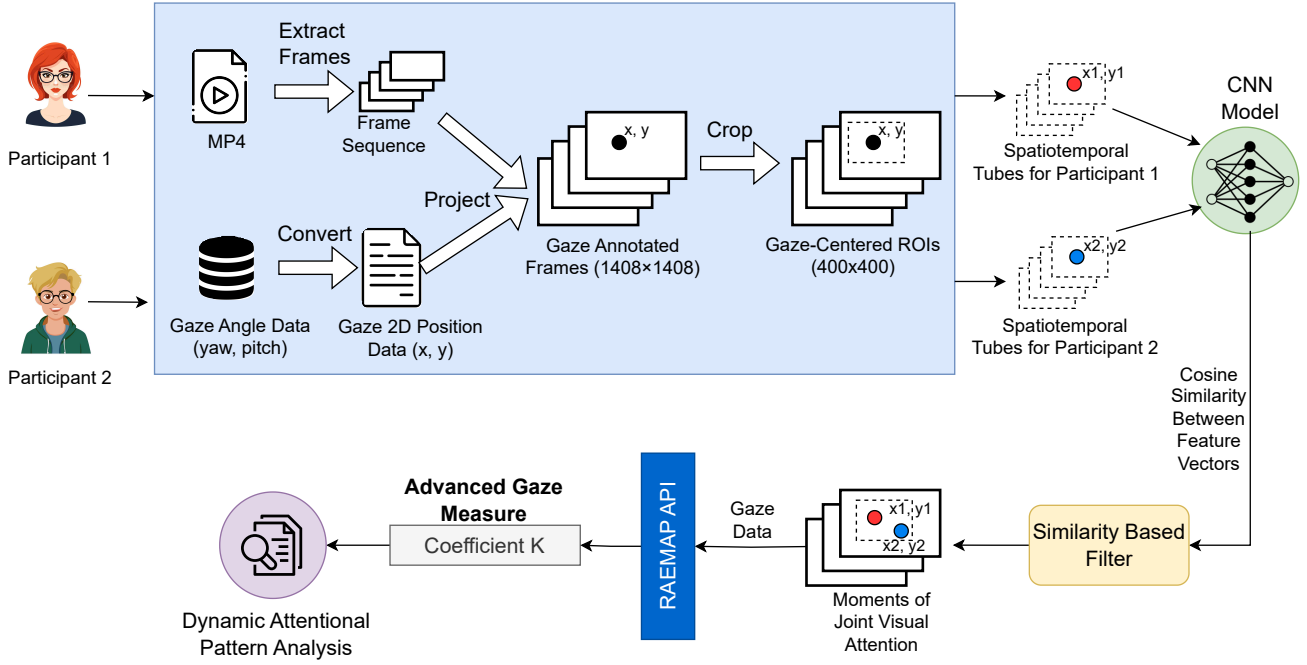


Fig. 1. **Processing pipeline for analyzing JVA between dyads in an egocentric setting.** We use egocentric video and gaze data from each participant in dual-participant activities in the Aria Everyday Activities Dataset [19]. Image frames are extracted from the video stream, and corresponding 2D gaze points are projected onto each frame. For each gaze-annotated frame, a 400×400 pixel region centered at the gaze point is cropped. Time-synchronized pairs of these gaze-centered regions (spatiotemporal tubes) are compared using a deep-learning model. Using a similarity threshold, we filtered the moments of joint visual attention. Gaze data from these moments are then processed through the REAMAP API [30] to compute Coefficient  $\mathcal{K}$ , which is analyzed to assess dynamic attention characteristics.

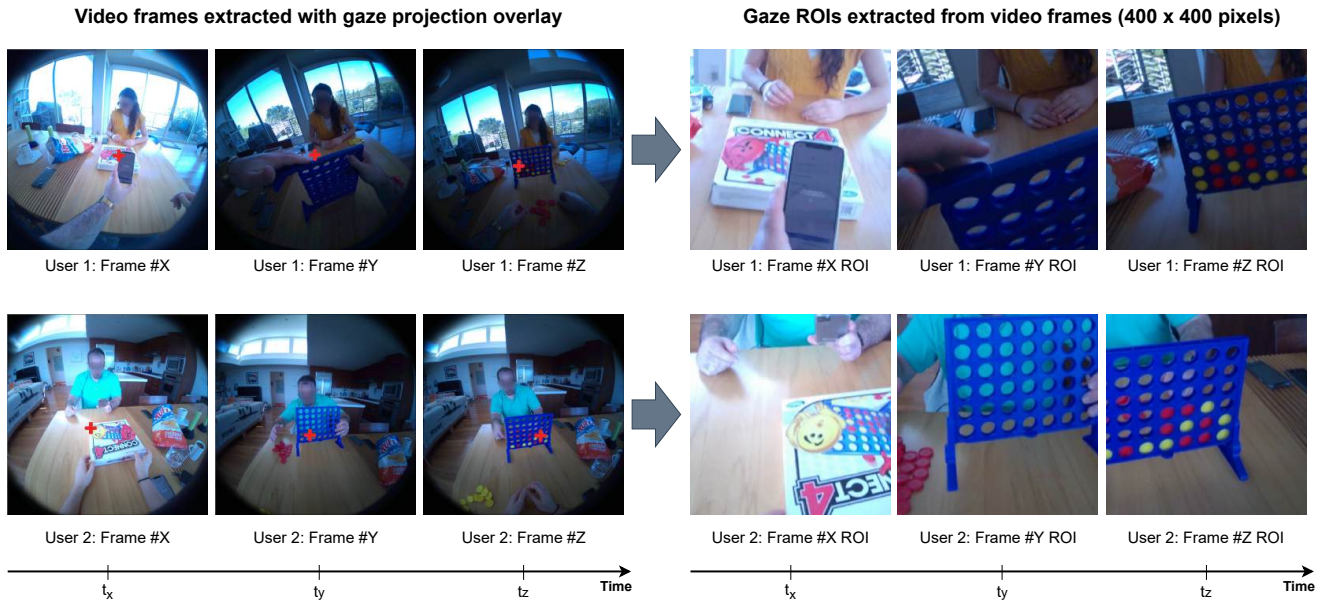


Fig. 2. **Extracting regions of interest (ROIs) around the gaze position from video frames.** The images on the left show the extracted frames from videos for both participants along the time. The images on the right show the 400×400 area around the gaze position extracted from each video frame, making it a spatiotemporal tube [26].

frame is processed with a pre-trained ResNet-50 convolutional neural network (CNN) [31], a model widely adopted for image recognition tasks. Rather than using the network for classification, we remove its final classification layer and use the remaining layers to extract deep feature representations that capture high-level visual patterns from each image.

These feature vectors are then compared using cosine similarity, a metric that quantifies the angular distance between two vectors in high-dimensional space. A higher cosine similarity score indicates greater visual similarity between the frames. By aligning frames based on their temporal positions, we compute similarity scores between image pairs across the two spatiotemporal tubes. This allows us to analyze how closely the visual content is aligned between two users or perspectives over time.

#### D. Moments of Joint Visual Attention and Dynamics of Coefficient $\mathcal{K}$

To identify moments of JVA between participants, we established a similarity threshold of 0.7 for comparing the spatiotemporal tubes. This threshold was empirically determined through careful manual inspection and comparison of various cases of joint attention and independent viewing. Different threshold values were evaluated, and the final value was selected based on their consistency in distinguishing similarities. When the similarity score between the tubes of both participants exceeded this threshold, the corresponding frames were considered as moments of JVA.

Following the identification of JVA moments, we analyzed the dynamic attention characteristics of each participant during these periods using the ambient-focal attention  $\mathcal{K}$ . To compute coefficient  $\mathcal{K}$  for each participant, we used the Real-Time Advanced Eye Movements Analysis Pipeline (RAEMAP) [32], an eye movement processing library. The coefficient  $\mathcal{K}$  is a dynamic indicator that captures the fluctuation between ambient and focal visual search behaviors [17] (see Equation 1).

$$\mathcal{K}_i = \frac{d_i - \mu_d}{\sigma_d} - \frac{a_{i+1} - \mu_a}{\sigma_a} \quad (1)$$

Where  $\mu_d$ ,  $\mu_a$  are the mean fixation duration and saccade amplitude, respectively, and  $\sigma_d$ ,  $\sigma_a$  the standard deviation of the fixation duration and saccade amplitude respectively, which is then computed over all  $n$  fixations. A positive  $\mathcal{K}$  value indicates a more focal scanning pattern, while a negative value reflects a more ambient scanning strategy during joint attention.

## IV. RESULTS

### A. Percentages of JVA

To quantify the prevalence of JVA throughout the interaction sessions, we calculated the percentage of JVA as a ratio between the number of frames filtered as JVA moments (those exceeding our established similarity threshold of 0.7) and the total number of frames in videos of each session. This approach provides a measure of joint attention frequency that

allows meaningful comparisons across varying activities and contexts. Our method of quantifying JVA is nearly related to the technique introduced by Schneider et al. [1], which used the number of overlapping gaze positions during a time window over the total gaze points during the entire activity to quantify the JVA in a collaborative setting. By adapting this approach to egocentric data, our measure captures the proportion of shared attention during interactions. As shown in Table I, our methods captured the highest level of joint attention in A4 and A5, where the participants interacted with a shared object, while the lowest was in A2 and A3, where the participants performed most tasks independently. While our observation proves the intuitive idea of the level of collaboration in each task, it also establishes the potential utility of our approach in joint attention detection.

TABLE I  
PERCENTAGE OF JOINT VISUAL ATTENTION IN DIFFERENT EVERYDAY ACTIVITIES

Activity	Percentage of JVA
<b>A1:</b> Playing a game	23.74%
<b>A2:</b> Making coffee	3.97%
<b>A3:</b> Cooking and eating	5.12%
<b>A4:</b> Watching a video on mobile	46.44%
<b>A5:</b> Watching Television	44.16%

### B. Dynamics of coefficient $\mathcal{K}$

We analyzed the dynamics of the ambient-focal attention coefficient  $\mathcal{K}$  to investigate changes in participants' attention patterns during periods of joint visual attention. Our temporal analysis divided each experiment into four epochs and computed the attentional characteristics. For this purpose, we employed the coefficient  $\mathcal{K}$  as a quantitative measure of visual attention, where negative readings indicate ambient attention, while positive values indicate focal attention. Our examination of  $\mathcal{K}$  values revealed changes in individual attention patterns across four temporal segments.

For a macro-level analysis of the behaviors, we manually annotated the events occurring each time by a panel of three experts, assigning a high-level annotation to the overall event during each temporal window. Then, we compared the behavior of  $\mathcal{K}$  with the events in each temporal window for a qualitative study (see Table II). Among instances where participants interact with the same shared object, we observed the  $\mathcal{K}$  to converge between participants (A1:T3-T4, A4:T4, A5:T1-T4). Moreover, we observed dissimilar  $\mathcal{K}$  readings where interactions were independent (A2:T1-T4) or when attention spread on different objects (A1:T1-T2). Our results suggest the potential of  $\mathcal{K}$  for measuring joint attention dynamics when combined with environmental events.

## V. DISCUSSION

The popularity of smart wearable devices, including AR/VR headsets and head-mounted cameras, has significantly advanced the collection and analysis of egocentric data for social attention research [33]. These technologies enable researchers

TABLE II  
HIGH-LEVEL EVENT ANNOTATIONS FOR TIME EPOCHS IN JOINT ATTENTION STUDY.

Activity	Epoch Annotation			
	Epoch 1	Epoch 2	Epoch 3	Epoch 4
A1: Playing a game	Preparing	Preparing	Playing	Playing
A2: Making coffee	Grabbing items	Grabbing items	Pouring	Drinking
A3: Cooking and eating	Preparing	Preparing	Eating	Eating
A4: Watching a video on mobile	Walking	Watching	Watching	Conversation, Watching
A5: Watching Television	Watching	Watching	Watching	Watching

to capture first-person perspectives complete with eye tracking, head movement, and environmental context during natural interactions. This methodological approach offers advantages for studying JVA

Our methodology extends previous work by adding another step to the identification of joint attention moments by incorporating analysis of attention patterns exhibited during these socially significant interactions. We use the ambient-focal attention with coefficient  $\mathcal{K}$  as an analytical measure to explore the potential of advanced gaze metrics to study JVA which will be a novel contribution to the field.

Our study encountered several technical challenges inherent to egocentric data analysis. The dynamic nature of the visual field in egocentric recordings presents substantial processing difficulties, as participant movements constantly alter the frame of reference. This is particularly evident in the Table I, in Activities 2 and 3. Additionally, the relatively low resolution of Project Aria glasses decreases the critical details, and distinguishing features become obscured, reducing the discriminative power of similarity metrics. Varying lighting conditions across different recording environments also affected our similarity calculations and ultimately affected in detecting frames with JVA.

Looking ahead, we plan to enhance the JVA identification component of our methodology by incorporating emerging techniques like Segment Anything [34] model, an advanced image segmentation algorithm by Meta. By detecting specific objects rather than relying solely on visual similarity, we may overcome some of the limitations imposed by lighting variations and dynamic viewpoints.

The introduction of the ambient-focal analysis into JVA research opens new research questions about the relationship between attention patterns and effective collaboration. Future studies could investigate whether certain patterns of ambient-focal coefficient  $\mathcal{K}$  dynamics correlate with more successful collaborative outcomes or more efficient task completion. We plan to expand this utility study on advanced eye tracking metrics to incorporate other advanced gaze measures like gaze transition entropy [9]. Ultimately, our study lays the groundwork for future research on JVA with detailed interpretations.

## VI. CONCLUSION

This study highlights the value of analyzing joint visual attention (JVA) through egocentric data in natural settings. Our methodology combines spatiotemporal tube analysis with advanced gaze metrics to identify and characterize shared

attention moments. Results show JVA patterns vary by task type, with higher rates during collaborative object-focused activities compared to independent tasks. The convergence of ambient-focal attention coefficients during shared object interaction suggests attentional synchronization that may enhance collaboration. These insights provide a detailed understanding of visual attention coordination in everyday activities. Our approach offers researchers a tool for investigating social attention in ecologically valid contexts, with applications in developmental psychology, human-computer interaction, and social robotics.

## REFERENCES

- [1] S. D'Angelo and B. Schneider, "Shared gaze visualizations in collaborative interactions: Past, present and future," *Interacting with Computers*, vol. 33, no. 2, pp. 115–133, 2021.
- [2] D. A. Baldwin, "Understanding the link between joint attention and language," in *Joint attention*. Psychology Press, 2014, pp. 131–158.
- [3] M. Tomasello and M. J. Farrar, "Joint attention and early language," *Child development*, pp. 1454–1463, 1986.
- [4] T. Charman, "Why is joint attention a pivotal skill in autism?" *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 358, no. 1430, pp. 315–324, 2003.
- [5] B. Mahanama, M. Sunkara, V. Ashok, and S. Jayarathna, "Disetrac: Distributed eye-tracking for online collaboration," in *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, 2023, pp. 427–431.
- [6] Y. Abeysinghe, B. Mahanama, G. Jayawardena, Y. Jayawardana, M. Sunkara, A. T. Duchowski, V. Ashok, and S. Jayarathna, "A-disetrac advanced analytic dashboard for distributed eye tracking," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 15, no. 1, pp. 1–20, 2024.
- [7] J. Duncan, "Converging levels of analysis in the cognitive neuroscience of visual attention," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 353, no. 1373, pp. 1307–1317, 1998.
- [8] B. Scassellati, "Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot," in *International Workshop on Computation for Metaphors, Analogy, and Agents*. Springer, 1998, pp. 176–195.
- [9] Z. Cui, T. Sato, A. Jackson, S. Jayarathna, M. Itoh, and Y. Yamani, "Gaze transition entropy as a measure of attention allocation in a dynamic workspace involving automation," *Scientific Reports*, vol. 14, no. 1, p. 23405, 2024.
- [10] B. Schneider and T. Bryant, "Using mobile dual eye-tracking to capture cycles of collaboration and cooperation in co-located dyads," *Cognition and Instruction*, vol. 42, no. 1, pp. 26–55, 2024.
- [11] S. Becker, S. Mukhametov, P. Pawels, and J. Kuhn, "Using mobile eye tracking to capture joint visual attention in collaborative experimentation," in *Physics Education Research Conference 2021 Proceedings*, 2021, pp. 39–44.
- [12] J. J. Macinnes, S. Iqbal, J. Pearson, and E. N. Johnson, "Wearable eye-tracking for research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices." *bioRxiv*, 2018. [Online]. Available: <https://www.biorxiv.org/content/early/2018/06/28/299925>

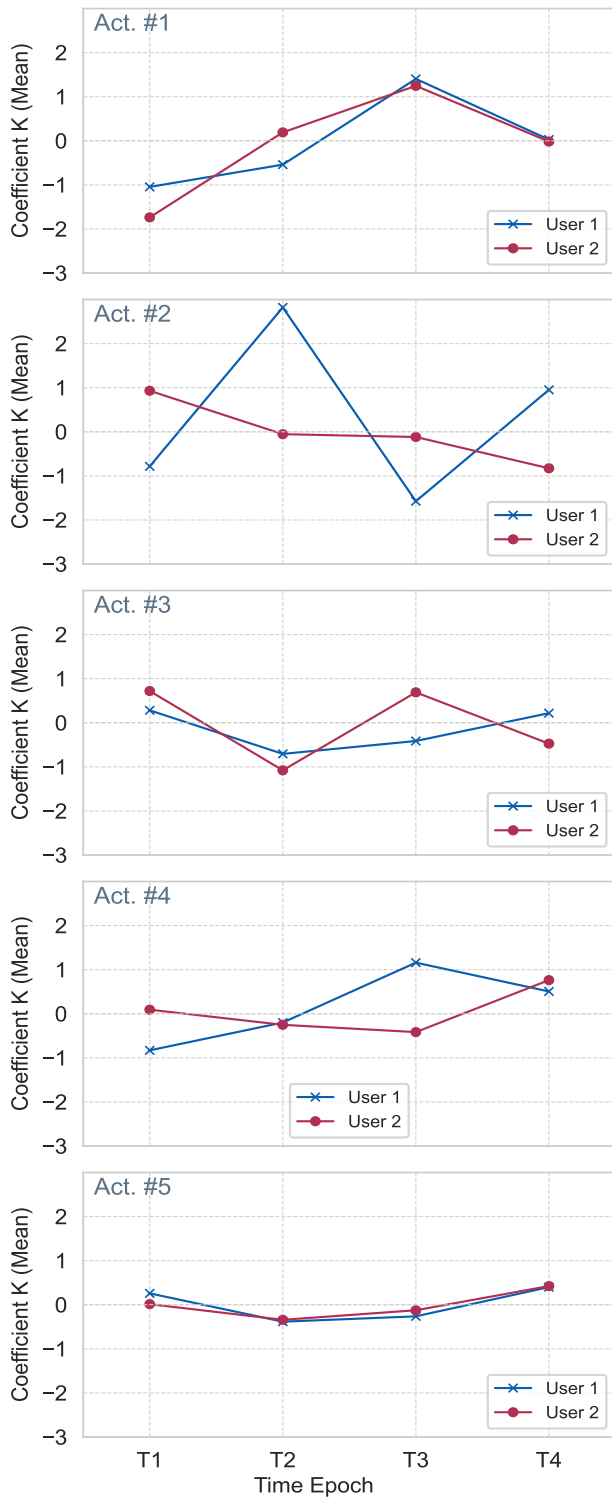


Fig. 3. Dynamics of ambient and focal attention using coefficient  $\mathcal{K}$  between users in each session across four time-epochs.

[13] M. Bock, H. Kuehne, K. Van Laerhoven, and M. Moeller, "Wear: An outdoor sports dataset for wearable and egocentric activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–21, 2024.

[14] Y. Wang, S. Panchadsaram, R. Sherhati, and J. J. Clark, "An egocentric video and eye-tracking dataset for visual search in convenience stores," *Computer Vision and Image Understanding*, vol. 248, p. 104129, 2024.

[15] S. Alletto, D. Abati, G. Serra, and R. Cucchiara, "Exploring architectural details through a wearable egocentric vision device," *Sensors*, vol. 16, no. 2, p. 237, 2016.

[16] B. Mahanama, Y. Jayawardana, S. Rengarajan, G. Jayawardana, L. Chukoskie, J. Snider, and S. Jayarathna, "Eye movement and pupil measures: A review," *frontiers in Computer Science*, vol. 3, p. 733531, 2022.

[17] K. Krejtz, A. Duchowski, I. Krejtz, A. Szarkowska, and A. Kopacz, "Discerning ambient/focal attention with coefficient  $k$ ," *ACM Transactions on Applied Perception (TAP)*, vol. 13, no. 3, pp. 1–20, 2016.

[18] G. Jayawardana, Y. Jayawardana, Y. Abeysinghe, B. Mahanama, S. Jayarathna, and J. Gwizdka, "A real-time approach to capture ambient and focal attention in visual search," in *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, 2025, pp. 1–7.

[19] Z. Lv, N. Charron, P. Moulon, A. Gamino, C. Peng, C. Sweeney, E. Miller, H. Tang, J. Meissner, J. Dong, K. Somasundaram, L. Pesqueira, M. Schwesinger, O. Parkhi, Q. Gu, R. D. Nardi, S. Cheng, S. Saarinen, V. Baiyya, Y. Zou, R. Newcombe, J. J. Engel, X. Pan, and C. Ren, "Aria everyday activities dataset," 2024.

[20] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talatoff, A. Yuan, B. Souti, B. Meredith *et al.*, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.

[21] V. Corkum and C. Moore, "Development of joint visual attention in infants," in *Joint attention*. Psychology Press, 2014, pp. 61–83.

[22] C. Yu and L. B. Smith, "Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination," *PLoS one*, vol. 8, no. 11, p. e79659, 2013.

[23] H. Bradley, B. A. Smith, and R. B. Wilson, "Qualitative and quantitative measures of joint attention development in the first year of life: A scoping review," *Infant and child development*, vol. 32, no. 4, p. e2422, 2023.

[24] R. E. Peters, A. Amatuni, S. E. Schroer, S. Naha, D. Crandall, and C. Yu, "Modeling joint attention from egocentric vision," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, no. 43, 2021.

[25] Y. Huang, M. Cai, and Y. Sato, "An ego-vision system for discovering human joint attention," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 4, pp. 306–316, 2020.

[26] H. Kera, R. Yonetani, K. Higuchi, and Y. Sato, "Discovering objects of joint attention via first-person sensing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 7–15.

[27] D. Tu, W. Shen, W. Sun, X. Min, and G. Zhai, "Joint gaze-location and gaze-object detection," *arXiv preprint arXiv:2308.13857*, 2023.

[28] H. Park, E. Jain, and Y. Sheikh, "3d social saliency from head-mounted cameras," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[29] G. Jayawardana and S. Jayarathna, "Automated filtering of eye movements using dynamic aoi in multiple granularity levels," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 12, no. 1, pp. 49–64, 2021.

[30] G. Jayawardana, A. Michalek, and S. Jayarathna, "Eye tracking area of interest in the context of working memory capacity tasks," in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2019, pp. 208–215.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] G. Jayawardana, "Raemap: Real-time advanced eye movements analysis pipeline," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Adjunct. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: <https://doi.org/10.1145/3379157.3391992>

[33] Y. Abeysinghe, K. Cauchi, V. Ashok, and S. Jayarathna, "Framework for measuring visual attention in gaze-driven vr learning environments using meta quest pro," in *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, 2025, pp. 1–3.

[34] Meta, "Introducing Segment Anything: Working toward the first foundation model for image segmentation — ai.meta.com," <https://ai.meta.com/blog/segment-anything-foundation-model-image-segmentation>, 2023, [Accessed 14-04-2025].