GazeCoT: Unleashing Social Intelligence in Multimodal LLMs With Gaze-Informed Chain-of-Thought Reasoning

Zhoutong Ye
Department of Computer Science and Technology, BNRist
Tsinghua University
Beijing, China
yezt24@mails.tsinghua.edu.cn

Xutong Wang
Department of Computer Science and Technology
Tsinghua University
Beijing, China
wangxuto23@mails.tsinghua.edu.cn

Chengwen Zhang
Department of Computer Science and Technology
Tsinghua University
Beijing, China
zcw25@mails.tsinghua.edu.cn

Ruiwen Zhang
Academy of Arts & Design
Tsinghua University
Beijing, China
zrw22@mails.tsinghua.edu.cn

Mingze Sun
Department of Computer Science and Technology
Tsinghua University
Beijing, China
sunmz24@mails.tsinghua.edu.cn

Qinwei Li
Department of Computer Science and Technology
Tsinghua University
Beijing, China
liqw24@mails.tsinghua.edu.cn

Chun Yu*†
Department of Computer Science and Technology, BNRist, College of AI
Tsinghua University
Beijing, China
chunyu@tsinghua.edu.cn

Yuanchun Shi†
Department of Computer Science and Technology, BNRist
Tsinghua University
Beijing, China
Qinghai University
Xining, China
shiyc@tsinghua.edu.cn

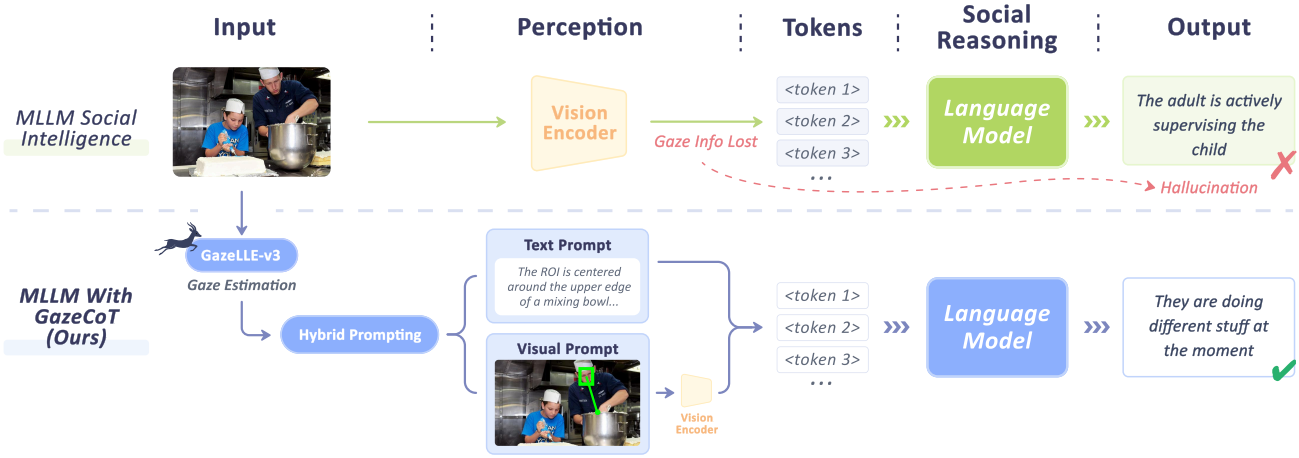


Figure 1: GazeCoT enables multimodal LLMs (MLLMs) to understand gaze as a social cue, which they are unable to do on their own, by leveraging a gaze estimation model and a hybrid visual and text prompting strategy. This unleashes the language model's text-based social intelligence in multimodal scenarios by addressing a critical limitation in multimodal social perception. The example image is from the GazeFollow dataset [103].

*Corresponding author.
†Also with Key Laboratory of Pervasive Computing, Ministry of Education

CHI '26, Barcelona, Spain

Abstract

Social intelligence is vital for effective human-AI interaction. While LLMs demonstrate strong text-based social intelligence, the vision

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790922

modality remains challenging due to the presence of non-verbal social cues. For example, gaze is the primary conveyor of social attention, yet it cannot be accurately perceived and understood by multimodal LLMs (MLLMs). Therefore, we propose GazeCoT, a pipeline using gaze estimation models to provide MLLMs with the attention of people in images or videos. The gaze information is provided as visual and text prompts compiled into a structured context to support MLLM social reasoning. Benchmark evaluation confirms that GazeCoT enhances MLLMs' social intelligence by improving gaze perception. A user study in a challenging application involving parent-child interactions demonstrates that GazeCoT improves perceived explainability and trustworthiness by aligning MLLM social perception and social reasoning with human norms. We hope that GazeCoT, a versatile plug-and-play pipeline, can enable socially aware, MLLM-based HCI applications.

CCS Concepts

• **Human-centered computing** → **Interaction paradigms**; • **Computing methodologies** → **Computer vision**.

Keywords

Multimodal Large Language Models, Human-AI Interaction, Gaze, Artificial Social Intelligence

ACM Reference Format:

Zhoutong Ye, Xutong Wang, Chengwen Zhang, Ruiwen Zhang, Mingze Sun, Qinwei Li, Chun Yu, and Yuanchun Shi. 2026. GazeCoT: Unleashing Social Intelligence in Multimodal LLMs With Gaze-Informed Chain-of-Thought Reasoning. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3772318.3790922

1 Introduction

Artificial social intelligence (ASI) — the AI's ability to perceive, interpret, and act on social cues — is central to human-AI collaboration [4, 37, 57]. It has great implications for human-robot interaction (HRI) [42, 83, 94, 108, 156], smart cities and spaces [91, 156], AI assistants [68, 104, 137] and accessibility [8], among other HCI applications. Large language models (LLMs) opened up a new path to building generalizable ASI. LLMs learn vast world knowledge, including knowledge on human social intelligence, through self-supervised pre-training. As a result, LLMs aided by chain-of-thought (CoT) reasoning [21, 89] have achieved considerable progress in ASI, demonstrating near human-level performance on some text QA benchmarks like SocialIQA [89, 107]. However, ASI goes beyond the text modality. **Social intelligence is inherently multimodal, involving the perception of numerous non-verbal social cues. Among these, human gaze is one of the most important** [12, 34, 37, 40, 65]. It is the primary conveyor of social attention [6, 92, 105, 140], and provides vital clues about a person's intent [96, 120], both of which are crucial for HCI applications [50, 84, 129]. But despite rapid progress in multimodal LLMs (MLLMs), these models are still largely unable to understand gaze cues [161]. This weakness can be attributed to MLLMs' failure to (1) understand fine-grained visual features of a person's head and eyes, and (2) perform spatial reasoning to find the fixation point, both of which are well documented shortfalls of MLLMs [147, 151, 161]. Without accurate gaze perception, which in our case means *accurately representing human gaze as MLLM tokens*, the models' potential in text-based social reasoning would be wasted in multimodal scenarios.

Given the importance of gaze in conveying human attention and intent, there has been work to integrate first-person, egocentric gaze information into MLLMs to improve interactive AI assistants in smart glasses and VR/AR headsets by using a wearable gaze tracker to provide the LLM- or MLLM-based agent with the user's gaze information [9, 15, 67, 68, 104, 126, 137, 144]. **However, integrating the gaze of persons other than the viewer into MLLM reasoning, useful for applications operating in third-person views like human-robot interaction (HRI), smart spaces and joint attention analysis, remains unexplored.** A key reason behind this gap is that, unlike in first-person applications, it is nearly impossible to obtain ground truth gaze data in third-person views in the wild. Despite this severe limitation, recent breakthroughs [106] in third-person gaze target estimation using large-scale pre-trained vision encoders [93, 118] enable a workaround: instead of relying on gaze trackers, we can use gaze estimation models to provide MLLMs with gaze information in third-person views.

To this end, we propose GazeCoT (**Gaze**-Informed **C**hain-**o**f-**T**hought), a plug-and-play pipeline integrating third-person gaze into the CoT reasoning context of MLLMs through a pre-trained gaze estimation model and hybrid visual and text prompting. Specifically, we develop a state-of-the-art gaze estimation model, GazeLLE-v3, that combines the latest advances in gaze estimation [106] and the DINOv3 [118] general-purpose vision backbone. We then combine two tried-and-tested prompting strategies to provide GazeLLE-v3's gaze estimation to MLLMs: (1) direct visual prompting [14, 113, 116, 141] by overlaying the image with a face bounding box, a gaze line and a fixation point, and (2) text prompting about the details on and around the fixation point. The visual prompt conveys GazeLLE-v3's estimations in an inherently grounded manner (i.e. pixel-level integration with the image) [70, 141] with minimal loss of information [157], while the text prompt fills in details and reduces hallucination [158]. Therefore, the two strategies complement each other [62]. Finally, we integrate these strategies into an efficient MLLM pipeline to create GazeCoT, focusing on reducing inference delay and hallucination. In summary, GazeCoT enables MLLMs to leverage GazeLLE-v3's gaze estimations, significantly improving MLLMs' ability to perceive gaze and use it for social reasoning. The plug-and-play nature of GazeCoT enables HCI researchers and practitioners to easily endow their systems with gaze-based social awareness.

We conduct extensive experiments on benchmarks and with users to evaluate the performance of GazeCoT and its impact on human-AI interaction. GazeCoT achieves a 25-percentage-point (pp) improvement in gaze target recognition accuracy over the baseline. To further evaluate GazeCoT's social intelligence in complex situations, we curate the *Gaze-grounded Social Intelligence* (GSI) benchmark that tests social perception, theory-of-mind (ToM) reasoning, and social interaction. GazeCoT achieves a 10pp gain on GSI, indicating improvements in complex social intelligence. Finally, we conduct user studies in a downstream application scenario, parent-child joint media engagement analysis. The results show

that GazeCoT improves output quality, explainability and trustworthiness, through its accurate gaze-informed comprehension of social scenes. The improved explainability and trustworthiness can be attributed to the role of gaze as a common ground between humans and MLLMs in social perception and social reasoning. We hope GazeCoT and our findings can help advance socially intelligent human-AI interaction. In summary, we make the following contributions:

- We propose GazeCoT, a versatile plug-and-play pipeline leveraging gaze estimation models to improve the social intelligence of MLLMs and facilitate smooth and productive human-AI interaction. **We expand the Gaze+MLLM paradigm into third-person scenarios using gaze estimation models.**
- We demonstrate the effectiveness of GazeCoT through evaluation on 2 benchmarks. A user study on a challenging downstream task, parent-child joint media engagement (JME) analysis, further proves that GazeCoT improves social perception and social reasoning in MLLMs.
- Our user study also shows that GazeCoT makes MLLMs more explainable and trustworthy by aligning MLLMs to human norms for social perception and social reasoning. This opens up new possibilities for better human-AI interaction.
- We publicly release our code, the GSI benchmark, and the model weights of the state-of-the-art GazeLLE-v3 gaze estimator, enabling easy off-the-shelf use of GazeCoT in downstream HCI applications. They are available in this GitHub repo.

2 Related Work

2.1 Artificial Social Intelligence and Social Cue Understanding

Artificial social intelligence (ASI) refers to AI's ability in social perception, theory-of-mind (ToM) reasoning, and natural social interaction [37]. It requires AI to interpret complex social interactions from various behavioral signals [37]. This, in turn, requires a precise understanding of the rich spectrum of non-verbal social cues that humans exhibit, such as gaze [12, 34, 40, 65], gesture [87], posture [139], and joint attention [26]. Traditional ASI research has made progress in perceiving and modeling specific human mental states like belief, intention, and joint attention [26, 35, 36, 115]. More recently, LLMs with powerful text-based reasoning and commonsense knowledge have also shown success in navigating social situations described in text [3, 25, 107, 135]. However, LLMs' success in text-centric tasks does not readily translate to the vision domain, as current multimodal LLMs (MLLMs) largely fail in fine-grained visual perception [45, 160], including that of non-verbal social cues like gaze [161].

This limitation is particularly significant for advancing ASI within HCI. Current research on social cue-aware interactive systems has diverged into two main streams. The first approach focuses on building interactive applications by converting social cues — such as gaze [68, 104], joint attention [20, 63], and gestures [52, 56] — into text for processing by LLMs [95]. However, this social cues-to-text pipeline inherently causes information loss, restricting the model's contextual awareness. A second, nascent stream explores the direct integration of social cues with multimodal LLMs (MLLMs) to overcome the limitation with text [137, 144]. While progress has been made in areas like gestures and joint attention [115, 154], the integration of third-person gaze (i.e. the gaze of individuals other than the viewer) remains unexplored. GazeCoT addresses this gap with a novel method that uses hybrid visual and text prompting to enable MLLMs to understand gaze, thereby boosting its utility for building socially intelligent interactive applications.

2.2 Multimodal Large Language Models and Their Limitations

Multimodal large language models (MLLMs) are a class of large language models (LLMs) modified to take non-text input [149]. For the vision modality, this means adding a vision encoder extracting features from images, and a vision adapter projecting the features into tokens understandable by the LLM [13, 31, 74, 77, 149, 166]. However, this setup has its limitations. Many MLLMs preprocess the image through *tiling*, where the image is broken down into smaller square tiles (e.g. 384×384 pixels) on arbitrary lines [24, 46, 72]. Each tile is then encoded separately. This approach not only degrades the semantics of objects broken up by tile boundaries [54, 147], but also assumes that all tiles are equally important and should be represented by the same number of tokens [24, 46]. Moreover, the models are trained on large-scale coarse-grained caption data crawled from the Internet [13]. The scarcity of detail-rich captions [55], as well as captions grounded in specific regions of images [45, 165], leads to models that are poor at capturing fine-grained details [45, 160]. Finally, MLLMs are still trained in a text-centric paradigm and generally have poor spatial reasoning capabilities [145, 147]. These factors exacerbate hallucination [78, 147, 158], a major hurdle for MLLM-powered HCI applications, especially high-stakes ones (e.g. medicine) [17, 60], and ones where it is unrealistic for users to consistently identify and correct hallucinations (e.g. accessibility for blind and low-vision users, real-time interaction, etc.) [18, 44, 123].

Likewise, all three factors, tiling-induced semantic degradation, coarse-grained training data, and lack of spatial awareness, are also relevant to MLLMs' failure to understand gaze [161]. Gaze often spreads across several tiles, while gaze estimation requires fine-grained image understanding and spatial reasoning [22, 27, 103, 106]. GazeCoT aims to mitigate these weaknesses by injecting gaze estimations into MLLM contexts, enabling MLLMs to perceive and reason in a more human-like, socially intelligent manner. In addition to text prompts, we also use visual prompts (i.e. marking the image with bounding boxes [14, 113], circles [116], arrows [14], scribble [141], etc.) to convey gaze information.

2.3 Gaze Target Estimation

Gaze estimation models are key to enabling MLLMs to overcome the aforementioned weaknesses in understanding human gaze. Formally, gaze target estimation is a computer vision (CV) task that involves *dense prediction* [106]. The input is an image and the bounding box of a person's head, while the output is a 2D *heatmap* representing the likelihood of each image patch (e.g. 8×8 pixels) containing the gaze target. Gaze target estimation requires both global scene information and head features. As a result, many models use a dual branch architecture with separate scene and head encoders [22, 27, 103]. Others use additional branches for depth

estimation [5, 38], body pose [47], and eye location [38] to capture more features. The state-of-the-art model for this task, GazeLLE [106], represents a return to simplicity. It takes a single RGB image as input and delegates all feature extraction to a pre-trained, general-purpose vision backbone (DINOv2 [93]), which generates scene tokens. It then decodes these tokens with a lightweight ViT-based gaze decoder. A head position prompt is used to guide the decoding process. The idea is that a general-purpose backbone like DINOv2 can extract all types of features (e.g. scene, depth, head, eye, pose, etc.) all on its own.

The single branch (RGB image input only) architecture of GazeLLE made it easier than previous multi-branch models to be integrated into GazeCoT. However, despite impressive performance metrics, GazeLLE has high failure rates when the task relies less on *scene features* and more on *head features* (see Figure 3 for cases). We hypothesize that the cause is the limited quality of dense features provided by DINOv2 [118]. Therefore, we train GazeLLE-v3, an improved model based on the DINOv3 backbone [118], as the foundation of GazeCoT.

2.4 Gaze-informed MLLMs for HCI

Gaze is a natural expression of a person's attention and intent. Therefore, first-person gaze has been thoroughly investigated as a way to facilitate natural target selection [59, 117], object manipulation [79, 82], and other types of input [51, 101] in AR/VR. For multi-user settings, cone-of-vision has been used to model joint attention and improve mutual awareness in VR-based collaboration [10, 11]. The emergence of LLMs and MLLMs led to similar work on human-AI interaction in XR devices and smart glasses, creating a flourishing field of Gaze+MLLM research. One such line of work focuses on **improving information querying** with gaze. For example, Voila-A [144], G-Voila [137], WalkieTalkie [68] and GazePointAR [67] explores using gaze as an indicator of user attention, reducing ambiguity in queries. Another line of work, including GazeLLM [104] and EmBARDiment [9], uses gaze to **shape MLLM context to match the user's context**. Here, gaze-informed visual and text prompts are used to improve AI productivity agents running on smart glasses and AR headsets. In addition, GazeNoter [126] and Sensible Agent [66] leverage gaze as a means of **natural input to AI agents** in AR/VR devices, while Augmented Object Intelligence [29] envisions MLLM-powered interaction with physical objects in AR, which can be enhanced by incorporating gaze information. Finally, AiGet [15] and SocialEyes [110] use egocentric gaze collected from AR headsets for **user modeling**, inferring mental states and social dynamics from gaze measurements.

Despite promising results in first-person views, the integration of third-person gaze information (i.e. the gaze of individuals other than the viewer) into MLLMs remains underexplored. Understanding gaze in third-person views is crucial for HCI applications that include, among others, human-robot interaction (HRI) [42, 83, 94, 108], smart cities and spaces [91], and joint attention analysis and mediation [131]. Gaze-informed MLLMs have the potential to revolutionize all these applications [32, 42, 85, 119, 153]. The challenge here is the lack of gaze tracking devices in third-person views, making it nearly impossible to collect ground truth third-person gaze data in the wild. However, recent advances in gaze estimation, discussed in Section 2.3, provide a workaround: gaze estimation models can be integrated into MLLM reasoning to mitigate the weaknesses discussed in Section 2.2. This enables us to expand the Gaze+MLLM paradigm in AR/VR research to third-person scenarios, improving MLLM social intelligence in more diverse use cases. To the best of our knowledge, GazeCoT is among the first works to systematically explore third-person gaze-informed MLLMs, offering a versatile plug-and-play solution.

3 GazeCoT

3.1 Overview

Our pipeline, GazeCoT, is designed to improve the social intelligence of MLLMs by injecting gaze information, an important cue in human behavior and social interactions, into the chain-of-thought (CoT) reasoning context. We derive three design goals from the discussions on the nature of MLLMs (Section 2.2) and gaze prompting (Section 2.4):

- **G1**: Injecting accurate and grounded gaze information into MLLM context with minimal loss of information.
- **G2**: Minimizing MLLM hallucination.
- **G3**: Improving efficiency, both in terms of token cost and inference delay.

An overview of GazeCoT is provided in Figure 2. The two types of gaze tools provide gaze prompts to the Task Agent, which uses them to complete the task, delivering gaze-informed and socially intelligent results. Our pipeline relies on hybrid visual and text prompting to achieve optimal results. We both directly overlay the gaze line and fixation point on the image as visual prompts (**G1**), and use MLLM to generate detailed text description of the region of interest (ROI) around the fixation point (**G1, G2**). We design a highly structured format for the Task Agent's CoT reasoning context to reduce hallucination related to long contexts (**G2**). The pipeline is parallelized to reduce latency (**G3**).

Our work can be divided into three main parts: improving gaze target estimation, designing appropriate gaze tools for hybrid visual and text prompting, and implementing an efficient MLLM agent pipeline in the form of GazeCoT. For gaze estimation, we improve upon the GazeLLE [106] model by switching to a more powerful pre-trained backbone, DINOv3 [118] (see Section 3.2 for details). We discuss the gaze tools used in hybrid gaze prompting in Section 3.3. Finally, the full GazeCoT pipeline and relevant design considerations are described in Section 3.4.

3.2 GazeLLE-v3: State-of-the-art Gaze Estimator for GazeCoT

GazeCoT relies on gaze estimation algorithms to circumvent the lack of ground truth provided by gaze trackers in third-person use cases. As discussed in Section 2.3, the original GazeLLE model [106], which uses DINOv2 as the feature extractor, fails in cases where *head and eye orientation* are the most important clues. We present some of the failure modes in Figure 3. We hypothesize that the reason for these failures is DINOv2's **inability to extract fine-grained head and eye features**. Therefore, we switch to DINOv3 [118], a stronger backbone trained on large-scale (1689M) image data, to improve the accuracy of gaze target estimation (**G1**).

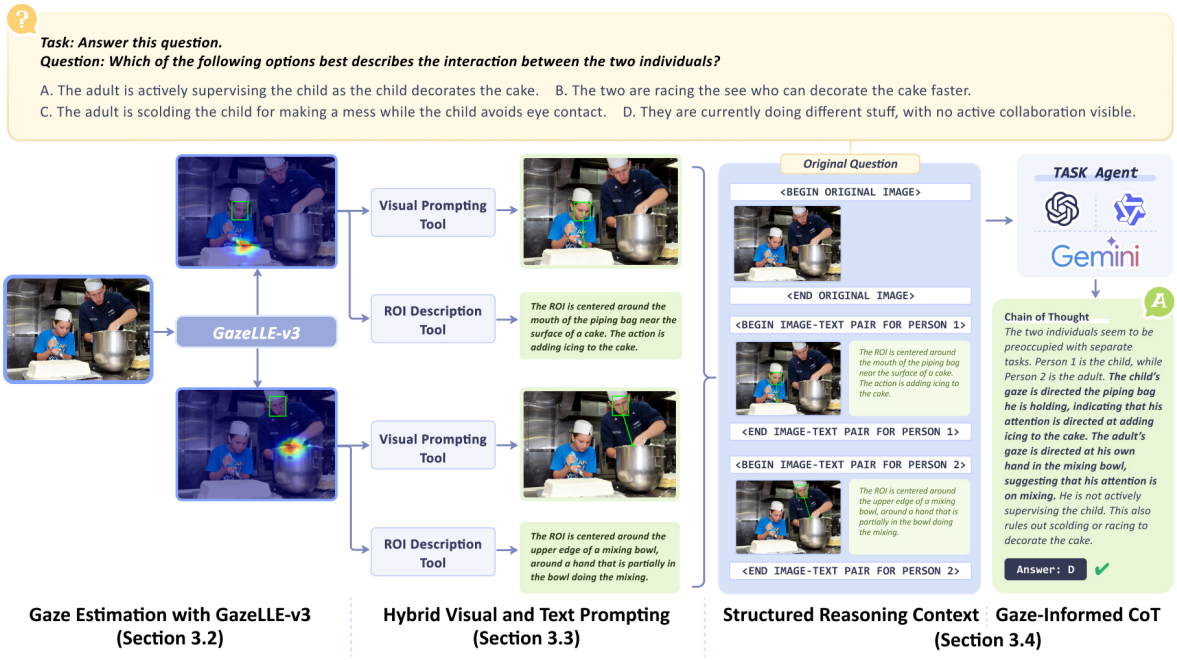


Figure 2: An overview of the GazeCoT pipeline in a social perception task. GazeLLE-v3 generates the gaze estimation of each person in the form of gaze heatmap. The visual prompting tool uses the heatmap to generate an image with visual prompts (facial bounding box, gaze line and fixation point). The region-of-interest (ROI) description tool generates a text description of the area with the highest heatmap value. We compile the visual and text prompts into a structured reasoning context to reduce hallucination. Finally, the Task Agent performs CoT reasoning to generate a gaze-informed output. This example is from the GazeFollow dataset [103].

Another advantage of DINOv3 is the use of Gram anchoring in its training process, which significantly improves the quality of dense feature maps used in gaze target estimation. A visualized comparison of DINOv2 and DINOv3 features is in Section 7.2.

Specifically, we train two variants of GazeLLE-v3, GazeLLE-v3-L and GazeLLE-v3-H, on the GazeFollow dataset [103]. The two models are based on the DINOv3-ViT-L and DINOv3-ViT-H+ backbones, respectively. We increased the input image size from 448×448 to 512×512 to fit the larger 16×16 patch size of DINOv3 (DINOv2 has 14×14 patches). To take advantage of DINOv3's higher quality features, we train our models for longer. Apart from the backbone, input image size and training time, the rest of the training procedure (e.g. output heatmap size, hyperparameters, training data, etc.) is identical to those used by Ryan et al. [106]. More details on model training are available in Appendix A.

Table 1 shows that GazeLLE-v3-L and GazeLLE-v3-H outperform the original GazeLLE models on all three commonly used metrics for gaze estimation. AUC (area under the curve) measures the similarity between the heatmap and the distribution of ground truth human annotations. Average L2 measures the average Euclidean distance between the predicted gaze target and all ground truth annotations. Minimum L2 measures the distance to the nearest ground truth annotation. **The L2 metrics are more important for our use case**. More discussion on the metrics is in Section 7.2. This result validates our choice of DINOv3 as the backbone. As shown in Figure 3, GazeLLE-v3-H is capable of handling cases that require fine-grained understanding of head and eye features.

Model	Backbone	GazeFollow		
		AUC↑	Avg. L2↓	Min L2↓
Human [106]	*N/A*	*0.924*	*0.096*	*0.040*
GazeLLE-B [106]	DINOv2-ViT-B	0.956	0.104	0.045
GazeLLE-L [106]	DINOv2-ViT-L	0.958	0.099	0.041
GazeLLE-v3-L	DINOv3-ViT-L	0.959	0.097	0.040
GazeLLE-v3-H	DINOv3-ViT-H+	**0.960**	**0.093**	**0.038**

Table 1: Results on the official test split of GazeFollow.

3.3 Hybrid Visual and Text Prompting with Gaze Tools

While GazeLLE-v3 models are great at predicting gaze target heatmaps, there remains a gap between that and improvements in MLLMs' social intelligence. The key here is to **inject the gaze information into the MLLM's context in a manner conducive to socially intelligent CoT reasoning**. This requires us to (1) convert the heatmap into visual and text prompts that can be understood by

[1]Used under Pexels' license (https://www.pexels.com/photo/women-talking-to-each-other-3894383/)

Figure 3: Images where the original GazeLLE-L model [106] performs poorly. A bounding box indicates the person whose gaze is being estimated, and a green dot indicates the predicted fixation point (the point with the largest heatmap value). A common theme of these failures is GazeLLE-L disregarding head and eye orientation in its estimations. In contrast, our GazeLLE-v3-H model is able to make correct (the two images on the left) or almost correct (rightmost image) gaze estimations. Image source (left to right): Zhang et al.'s benchmark [161], Pexels[1], the GazeFollow dataset [103].

MLLMs, and (2) reduce hallucination, especially in the vision modality. Figure 4 provides an overview of the tools.

Visual Prompting With Gaze Lines. The first gaze tool is straightforward, **converting the heatmap into an MLLM-friendly visual prompt (G1)**. As discussed in Section 2.2, effective visual prompting practices often leverage bounding boxes [113], circles [116], lines [14] and arrows [14]. Therefore, we use the same geometric shapes to illustrate the gaze information for MLLMs **(G1)**. First, we draw a bounding box around the head of the person whose gaze is being visualized. This helps the model to ground the gaze to a specific person. We then draw a straight line from the center of the head bounding box to the fixation point, which is defined as the point with the highest heatmap value. Finally, we add a dot at the fixation point to further emphasize its importance. This type of direct visual prompting is inherently grounded [70, 141] (i.e. pixel-level integration with the image) and therefore leads to minimal information loss in that regard [157] (**G1**). We adopt a one-prompted-image-per-person strategy here to avoid having too many visual prompts in a single image, where the gaze lines and fixation points overlap with each other, cluttering the scene and making it difficult for the MLLM to make sense of the gaze information **(G1, G2)**.

Text Prompting With ROI Descriptions. One problem with the first tool is that it compresses the entire region-of-interest (ROI) into a single fixation point, which may lead to loss of information. We mitigate this limitation with another tool, the ROI description tool. It is designed to **reduce hallucination (G2)** and improve MLLM's capability to **understand fine-grained details (G1)** around the fixation point. This is achieved through generating a detailed text description of a zoomed-in view of the ROI surrounding the fixation point. This approach combines zoom, an effective strategy to improve MLLM performance most prominently used in OpenAI's ChatGPT o3, and gaze information. We first extract the ROI using the heatmap (see Appendix C for details on the heatmap-to-ROI conversion). We then crop out the ROI and use a small circle to illustrate the location of the predicted fixation point. The crop is then sent to an MLLM, which is prompted to describe the gaze target and surrounding objects and persons in detail. Since the task of describing the ROI is relatively simple, we can use a smaller and more efficient MLLM (e.g. GPT-4.1 Mini, Qwen2.5 VL 7B, etc.) for this step (**G3**). Crucially, we keep only the text description and discard the zoomed-in crop of the ROI because having too many ROI crops in scenes with multiple individuals would lead to increased hallucination for the Task Agent (**G2**) [75, 114, 143, 158].

By combining visual prompts and text prompts, we obtain a hybrid prompting strategy that accurately portrays the global spatial semantics of the gaze information, as well as the details surrounding the fixation point. This fills the gap between an accurate gaze estimation and concrete improvements in MLLMs' ability to understand human gaze, enabling the model to use that information to act in a socially intelligent manner.

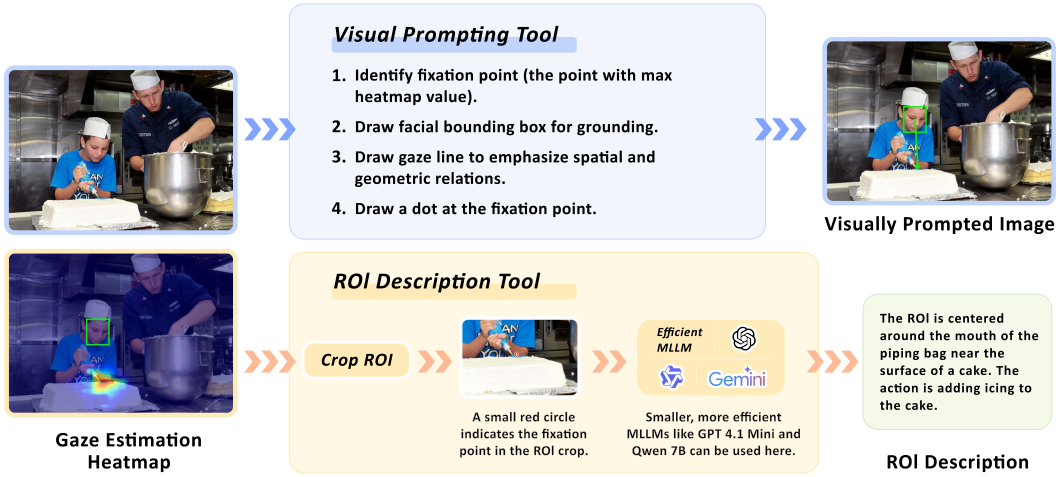


Figure 4: Overview of the two gaze tools used to convert heatmaps into MLLM-readable visual and text prompts. The visual prompting tool preserves most of the spatial information of the gaze estimation by GazeLLE-v3 and converts it into intuitive, MLLM- (and human-) friendly visual prompts. However, this tool collapses the entire region-of-interest (ROI), the red area in the visualization, into a single fixation point. Therefore, we generate a detailed text description of the ROI with the ROI description tool to cover the span of the entire high-probability area to complement the visual prompt.

3.4 GazeCoT: An Efficient Pipeline for Gaze Prompting

Introducing gaze information would inevitably increase delay and computational cost. Since face recognition and GazeLLE-v3 take only fractions of a second per inference, our focus in pipeline design is on the MLLM side, where each inference or API call takes at least several seconds. The objective for GazeCoT here is to (1) parallelize MLLM inference to reduce delay (**G3**), and (2) properly structure the Task Agent's CoT reasoning context to reduce hallucination (**G2**). We leverage the fact that describing different ROIs is a set of mutually independent tasks, parallelizing the ROI description tool through either parallel API calls or local batch inference. For Task Agent context management, we use a highly structured format with clear labels for the original images, the visually-prompted images, and text descriptions of ROIs. This format, illustrated in Figure 2, reduces hallucination caused by long, unstructured context, and improves the overall performance of GazeCoT.

Combining these elements leads to the GazeCoT pipeline illustrated in Figure 2. First, we detect the faces present in the image. We then obtain the gaze estimations of each individual through GazeLLE-v3. After that, we use the visual and text prompting tools described in Section 3.3 to generate hybrid visual and text gaze prompts for the Task Agent, and structure these prompts in a clearly labeled manner to reduce hallucination. Finally, the Task Agent conducts CoT reasoning and generates a gaze-informed output, completing the task.

4 Research Questions

We conducted extensive experiments with GazeCoT, both on benchmarks and with users, to answer the following research questions (RQs):

- **RQ1**: Can mainstream MLLMs understand gaze prompts and use them to infer gaze targets in images?
- **RQ2**: Does GazeCoT improve the social intelligence of MLLMs?
- **RQ3**: How does each component of GazeCoT contribute to overall performance?
- **RQ4**: Does GazeCoT affect other key aspects of human-AI interaction (explainability, trustworthiness, etc.)? If so, what is the underlying mechanism?

We answer **RQ1** with experiments on a controlled gaze target recognition benchmark for MLLMs in Section 5.1. We address **RQ2** in Section 5.2 with our novel Gaze-grounded Social Intelligence (GSI) benchmark, which evaluates social intelligence in complex scenarios. To address **RQ3**, we perform ablation studies on GazeLLE-v3, the ROI description tool, and the structured prompt of the Task Agent on the 2 benchmarks to validate GazeCoT's design (Section 5.3). Finally, we conduct user experiments in a real-world application scenario, namely *parent-child joint media engagement (JME) analysis*, to provide insight on **RQ2** and **RQ4** (Section 6).

5 Benchmark Experiments

5.1 Gaze Target Recognition

To address **RQ1**, we evaluate GazeCoT and a CoT prompting baseline on the Gaze Referential Inference benchmark proposed by Zhang et al. [161]. This visual question answering (VQA) benchmark is specifically designed to test MLLMs' ability to infer gaze target. It includes 907 image-question pairs, all of which are collected in a simple, controlled setting. The questions are solely about which object the person in the image is looking at. This allows us to single out the *gaze estimation* capability of GazeCoT and the baseline, and see if MLLMs can leverage gaze prompts based on GazeLLE-v3, providing a direct answer to **RQ1**. We provide a sample question from this benchmark in Figure 5.

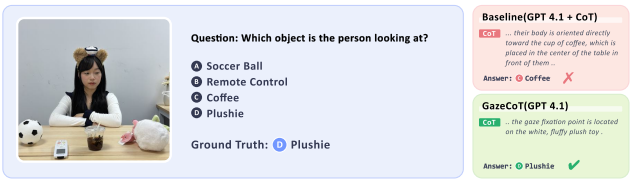


Figure 5: A sample question from the gaze target recognition benchmark proposed by Zhang et al. [161]. This sample demonstrates a common failure mode identified by Zhang et al., namely that MLLMs rely almost exclusively on face and body orientation, and largely disregard eye direction.

Evaluation Procedure. For this benchmark, we adopt the standard procedure for MLLM VQA. All questions are evaluated with zero-shot CoT prompting, with the GazeCoT prompt providing an additional explanation of the visual prompts. We set the temperature (τ) to 0 except for GPT-5, which only accepted $\tau = 1.0$. To account for the observed bias of MLLMs in multiple choice questions (e.g. a tendency to choose the second option regardless of content [71]), we evaluate each question 3 times, randomly shuffling the choices each time. We take the standard LLM-as-a-judge approach [43, 53, 81, 151] for grading the VQA output, where a LLM is prompted to compare the output and the ground truth and provide a judgment of either *right* or *wrong*. Since grading the answer of multiple choice questions is very straightforward, with the only variability being the character set used in the output, we use GPT-4.1 Nano as an efficient judge. We manually inspected 100 evaluations by the GPT-4.1 Nano judge and found no errors as expected. To demonstrate the generalizability of GazeCoT, we run the pipeline with multiple MLLMs, including both open source (Qwen2.5 VL) and proprietary (GPT-4.1, GPT-5) models. The detailed setup for these pipelines is presented in Figure 6.

GazeCoT Version	Gaze Estimation Model	Task Agent Model	Zoom Description Model
Proprietary Models			
GPT 4.1	GazeLLE-v3-H	GPT 4.1	GPT 4.1 Mini
GPT 4.1 Mini	GazeLLE-v3-H	GPT 4.1 Mini	GPT 4.1 Mini
GPT 5	GazeLLE-v3-H	GPT 5 *(Reasoning Effort: Medium)*	GPT 5 Mini *(Reasoning Effort: Low)*
Open Source Models			
Qwen 2.5 VL	GazeLLE-v3-H	Qwen 2.5 VL 72B	Qwen 2.5 VL 7B

Figure 6: The different versions of GazeCoT evaluated in benchmark experiments, implemented with different MLLM families. The Task Agent uses more powerful MLLMs, while the zoom description models are smaller and more efficient. We use the *low* and *medium* reasoning effort settings for GPT-5 to reduce cost and latency, as setting reasoning effort to *high* causes GPT-5 to reason for several minutes and generate thousands of reasoning tokens.

Results and Analysis. We present the results of the gaze target benchmark experiment in Figure 7. Baseline MLLMs performed barely above chance, which is consistent with the original findings of Zhang et al. [161]. This suggests that existing MLLMs are

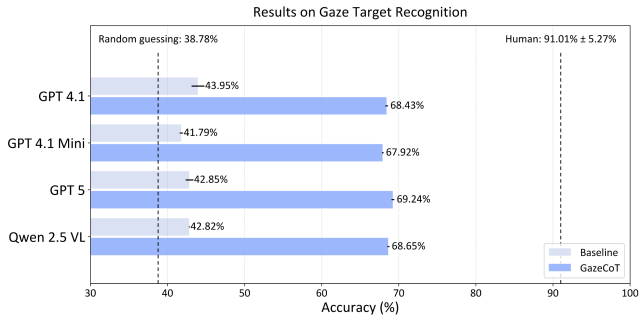


Figure 7: Results on the gaze target recognition benchmark. Standalone MLLMs (baseline) performed only marginally better than random guessing. GazeCoT significantly enhanced performance for all 4 MLLMs, improving accuracy by around 25 percentage points. Despite this leap in accuracy, GazeCoT still performed below human level due to the limits of the GazeLLE-v3-H gaze estimation model. Human performance is cited from [161]. More discussion on improving gaze estimation can be found in Section 7.2.

largely unable to infer gaze targets and, by extension, understand gaze as a social cue. This underscores the necessity of explicitly injecting gaze information into MLLMs' reasoning context. In contrast, GazeCoT consistently improved the performance of all MLLM families evaluated, achieving nearly 70% accuracy in all four setups, demonstrating the effectiveness and generalizability of our approach. However, it should be noted that, while gaze prompts are capable of significantly improving MLLMs' ability to infer gaze targets, there remains a large human-MLLM gap in this specific benchmark, suggesting room for improvement. The bottleneck here is GazeLLE-v3 (see Appendix B for bad cases), and we expect that progress in both general-purpose vision backbones and gaze estimation algorithms will narrow the gap. In conclusion, mainstream MLLMs can leverage gaze predictions to improve gaze perception. This addressesd **RQ1** and lays a solid foundation for improvements in higher-level social intelligence.

5.2 Gaze-grounded Social Intelligence

The results in Section 5.1 show that GazeCoT substantially improves MLLMs' ability to infer gaze targets. In this experiment, our objective is to explore **RQ2**. That is, whether this improvement translates to improved social intelligence, especially in more abstract and high-level aspects like ToM and social interaction. Therefore, we curate the *Gaze-grounded Social Intelligence* (GSI) benchmark to evaluate MLLMs' social intelligence in complex scenarios.

Benchmark Curation Process. In GSI, we use images from three gaze target estimation datasets, GazeFollow [103], GOO-Real [125] and ChildPlay [122], and craft social-intelligence-related questions based on the ground truth gaze annotations. To complement the images from gaze datasets, we also include images with ground truth captions from the Internet, further diversifying the scenarios in GSI. In total, GSI contains 320 social intelligence questions from diverse scenarios, testing the MLLMs' performance in social perception, theory-of-mind (ToM) reasoning, and social interaction. 20.9%

Figure 8: Three sample questions from GSI, representing three different levels of social intelligence. The first question is solely about social perception. The second question requires the model to predict human action, touching on ToM reasoning. The third question requires the model to interact with the person and provide appropriate in-context shopping advice, a form of social interaction.

of GSI questions use GazeFollow images, 30.9% are from GOO-Real, 7.2% are from ChildPlay, and the remaining 40.9% are based on images we collected online. All questions were quality checked by at least two other researchers to ensure that they are related to social intelligence, have unambiguous answers, and are reasonably straightforward to humans (i.e. solvable with common sense). For images collected online, we strictly adhered to copyright laws and licensing. Three sample questions representing different levels of social intelligence are provided in Figure 8.

Evaluation Procedure, Results and Analysis. In this experiment, we follow the same evaluation protocol described in Section 5.1 and evaluate the same MLLM families. We report the results in Figure 9. GazeCoT performed significantly better than baseline MLLMs with CoT reasoning on GSI, demonstrating that improved capability to understand human gaze translates to improved social perception, ToM reasoning, and the ability to engage in social interaction among MLLMs. Improvement is observed on all 4 versions of GazeCoT, each based on different MLLMs, showcasing the generalizability of our approach. The results also show that, even in complex scenes with social cues other than gaze, the latest and most advanced MLLMs like GPT-5 can still benefit substantially from the explicit injection of gaze information. On the other hand, even smaller models like GPT-4.1 Mini performed better with GazeCoT than larger and more advanced models like GPT-5 under the baseline condition, demonstrating the importance of gaze in social intelligence. In conclusion, the experiment on the GSI benchmark directly addresses **RQ2** and proves that GazeCoT can improve the social intelligence of MLLMs through integrating gaze estimation into the CoT reasoning process.

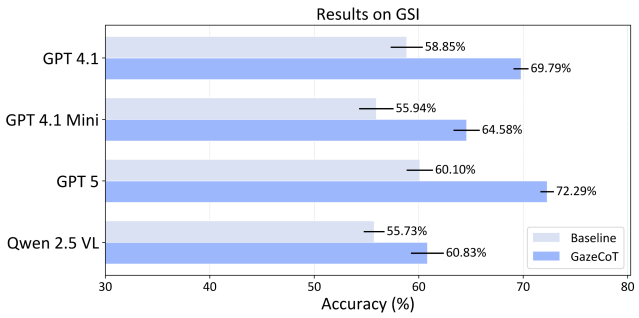


Figure 9: Results on the GSI benchmark. GazeCoT significantly improved performance for all 4 MLLMs.

Another observation is that, unlike in the simpler gaze target recognition benchmark, the performance of GazeCoT on GSI is more dependent on the base model (60.83% for Qwen 2.5 VL vs 72.29% for GPT-5). More advanced models like GPT-5 performed significantly better than weaker models like Qwen 2.5 VL and GPT-4.1 Mini, showing that perception, social knowledge, and reasoning capability all play a role in GSI questions. This validates that the benchmark curation process went as intended and led to a benchmark that evaluates higher-level social intelligence.

5.3 Ablation Study

To dissect the performance contributions of GazeCoT components and provide insight into **RQ3**, we perform ablation studies on GazeLLE-v3, the ROI description tool, and the Task Agent's structured CoT reasoning context. The ablation study does not include the visual prompting tool and the Task Agent because these are essential to the functioning of the pipeline, and removing them would break down GazeCoT entirely. All ablation studies are done on the GPT-4.1 and Qwen 2.5 VL versions of GazeCoT using the procedure described in Section 5.1. We chose these models

Setting. We demonstrate the need to train GazeLLE-v3 by comparing the performance of GazeCoT using gaze estimations by GazeLLE-v3-H and the original GazeLLE-L model released by Ryan et al. [106]. The ablation study on GazeLLE-v3 is conducted on the gaze target recognition benchmark, since the controlled setting allows us to single out the gaze estimation capability for comparison. We report the results in Table 2. The next ablation study involves the ROI description tool, which provides text prompts to complement the main visual prompt. We compare the performance of GazeCoT with and without the ROI description tool. Finally, we compare the performance of GazeCoT with and without the structured Task Agent context. Without structured context, the visually prompted images and text descriptions would be given to the Task Agent without in-context labels. Here, we keep the system prompt unchanged and provide explanations of the images and text descriptions in general. The ablation study on pipeline design is conducted on both the controlled gaze target recognition benchmark and GSI. We report the results in Figure 10.

Results and Analysis. These results demonstrate that GazeLLE-v3-H performs substantially better, more than doubling the gain

Benchmark	Pipeline	Gaze Estimator	Backbone	Accuracy (%)	Improvement (pp)
Gaze Target	Baseline	N/A	N/A	43.95 ± 0.83	N/A
	GazeCoT	GazeLLE-L	DINOv2-ViT-L	54.36 ± 0.39	10.41
	GazeCoT	GazeLLE-v3-H	DINOv3-ViT-H+	68.43 ± 0.17	24.48

Table 2: Performance of GazeCoT (GPT-4.1 version) with different gaze estimators on the gaze target recognition benchmark. GazeLLE-v3-H more than doubles the original GazeLLE-L model's improvement over the baseline. This validates our qualitative observations in Section 3.2 and our decision to train stronger gaze estimation models.

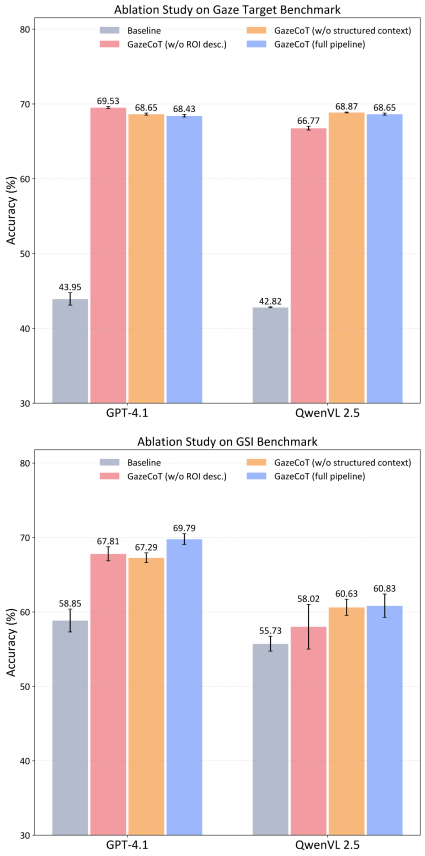


Figure 10: Results of the ablation study. We tested the proprietary GPT-4.1 and open source Qwen2.5 VL 72B on the gaze target recognition benchmark (top) and GSI (bottom).

over the baseline compared to GazeLLE-L in the gaze target recognition benchmark. This validates our choice of training stronger gaze estimation models. This is consistent with the observations in Section 3.2 (Figure 3).

For pipeline components, we first analyze the results on gaze target recognition. For GPT-4.1, the ROI description tool caused a slight drop in accuracy. This is expected, as this benchmark is collected in a very simple and highly controlled environment. As a result, the ROI contains little additional detail, and the image overlaid with visual prompts alone is enough for the MLLM to accurately identify which object is being looked at. However, ROI description still improved performance on gaze target recognition for the weaker open-source model, Qwen 2.5 VL, suggesting that the tool is more important for less capable models. Similarly, since the scene contains only one person, the structured CoT context is almost identical to an unstructured one, leading to similar performance with or without the structured context for both GPT and Qwen. In contrast, results on GSI clearly demonstrate that both ROI description and structured reasoning context contribute to better high-level social intelligence. The ROI description tool led to substantial improvements in GSI performance for both GPT and Qwen, suggesting that the tool is beneficial for social reasoning in complex scenes. The structured CoT reasoning context also improves performance on GSI, especially for GPT-4.1 in complex scenes, suggesting that gaze prompting requires careful context management to ensure optimal performance.

In conclusion, all components of the GazeCoT pipeline, including the GazeLLE-v3 gaze estimator, the ROI description tool, and the structured CoT reasoning context of the Task Agent, contribute to overall performance, especially in complex scenarios. This largely validates the design of GazeCoT and answers **RQ3**.

6 User Study

6.1 Scenario

We adapt GazeCoT to *parent-child Joint Media Engagement (JME) [150] analysis* to demonstrate its potential to benefit downstream applications by improving social perception, a key aspect of artificial social intelligence (ASI) (**RQ2**), and by improving the perceived explainability and trustworthiness of MLLM output (**RQ4**). In parent-child JME analysis, researchers need to analyze video recordings of parent-child interactions to extract information. A widely-used method is taking field notes [2, 33, 111, 148], where researchers observe and record in detail parents' and children's verbal and non-verbal communication. These notes can then be used for research, providing expert advice to parents, or given directly to parents as a form of recap [2, 98]. Automating this process requires advanced social perception and reasoning capabilities, and is out of reach of existing MLLMs [115]. Therefore, we chose automated field note generation and analysis due to its value to both researchers (as a tool for automatic user study coding) and parents (as a source of high quality and always available feedback), as well as its high demand on social intelligence.

Following the experiment format used in GazeLLM [104], we use GazeCoT and baseline MLLMs to generate field notes of 11 play sequences (Sections 6.2 and 6.3), and ask experts to rate the field notes. The metrics cover both output quality (accuracy, comprehensiveness, and usefulness) and subjective perceptions (explainability

and trustworthiness) to address **RQ2** and **RQ4**, respectively. We describe the procedure in Section 6.4 and discuss the result in Section 6.5.

6.2 Play Sequences

In this experiment, we use play sequences recorded in a previous parent-child-AI collaboration study with 3- to 6-year-olds. In that study, the parent and child dyad engage in English as a foreign language (EFL) learning through playing with toys while seated at a table. The AI assistant, running on a laptop or a smartphone, provides support on game progression and English expressions. See Appendix E for more details on how the play sequences were collected. Our use of these play sequences in this study has the full permission of the authors of the original study, is strictly in accordance with the terms agreed to by participating parents, and has been approved by the university ethics committee.

Figure 11: The two types of camera angles involved, front view and side view. Both provide clear view of the parent, the child, and the tabletop. The laptop in the image on the right hosts the AI assistant. Faces are blurred to protect privacy.

The main rationale behind choosing these play sequences is that the scenes are dynamic and complex, with the parent and child manipulating dozens of toys, communicating with each other, and interacting with the AI assistant. This creates a challenging scenario full of verbal and non-verbal cues difficult for MLLMs alone to accurately understand and describe. Another reason is the camera angle, which provides a clear view of the faces and the play area (Figure 11). Finally, using the view for automated field note generation and analysis would directly benefit the downstream application explored by the original study, enabling features like play sequence recap and tailored assistance for the parent based on past field notes. In summary, this scenario allows for the evaluation of fine-grained social perception (**RQ2**) and subjective perceptions of GazeCoT output (**RQ4**) in a real and challenging environment. In total, we selected 11 play sequences from that study, all of which are between 20 and 60 seconds in duration.

6.3 Generating Field Notes With GazeCoT

We slightly modify the GazeCoT pipeline described in Section 3 to meet the unique demands of generating field notes. Specifically, we modify the visual prompting format proposed in Section 3.3 by drawing the facial bounding boxes, gaze lines, and fixation points of both the parent and the child together in a single visually prompted image. Since the play sequence contains only two individuals, having too many visual prompts cluttering the scene and overlapping with each other is not a problem, unlike in scenes with more people. To adapt GazeCoT to video input, we first cut the video into 20-second segments, and extract frames at a constant 1 frame per second (FPS). We empirically chose 20s × 1FPS for two reasons: (1) 1 FPS provides enough fine-grained details on the dynamics of the play sequence, and (2) 20 images per MLLM query is a heavy but not impossible task for GPT-4.1. Furthermore, we modify the structured reasoning context of the Task Agent to include timestamps for each frame to ensure that the field notes are accurately timestamped. In addition to the video, we also use GPT-4o Transcribe to provide the MLLM with the text transcript of the speech in each clip. Finally, we use GPT-4.1 to merge and summarize the field notes of all 20-second video segments, resulting in the final field note for the entire play sequence. We show the modified visual prompts and structured reasoning context, as well as a partial example of the GazeCoT-generated field notes, in Figure 12. The full example can be found in Appendix F.

We use this modified GazeCoT pipeline to generate 1 field note per play sequence. For the baseline condition, we used the same MLLM (GPT-4.1), substituting the visually prompted frames with unedited ones. We also generate 1 field note per play sequence with the baseline MLLM. This results in 11 field notes for each group (GazeCoT and Baseline). We ensured that all field notes have the same formatting to facilitate fair comparison.

6.4 Participants and Procedure

We recruited 10 graduate-level experts (6F, 4M, Age 23.5±2.5) in early childhood education. We refer to them as P1 to P10. All are Ph.D. or graduate students conducting active research in the field. On average, they have participated in 2 early childhood education research projects as the primary experimenter responsible for hosting the experiment and observing the parent-child or instructor-child interaction. We chose Ph.D. and master's students instead of more senior researchers because the students have fresh hands-on experience taking field notes. All participants were compensated for their time at a rate roughly equivalent to 15 USD per hour.

Procedure. For each participant, we first go through a warm-up phase lasting around 15 minutes, where the participant watches a clip of a play sequence and two field notes of that clip. This helps to familiarize the participant with the settings of the experiment and enables them to make sense of the rating criteria. After the warm-up phase, the participant is provided with 11 clips, each with 2 field notes (one generated by GazeCoT, the other by Baseline GPT-4.1), respectively. We do not reveal which system generated each field note to the participant at this stage. After watching the clip and comparing the field notes, the participant is asked to rate the field notes on *accuracy*, *comprehensiveness*, *usefulness*, *explainability*, *trustworthiness*, and *overall preference* on a 7-point Likert scale. The exact definitions of the 6 metrics provided to the participant are shown in Appendix G. As the participant rates each field note, they can choose to watch the video clip again when needed to confirm details or refresh their memory. Finally, after the participant finishes rating the field notes of every clip, we conduct a semi-structured interview lasting 10 to 20 minutes to discuss the experience. During the interview, we would reveal to the participant which system generated each note. We would ask them about the overall difference between the field notes generated by the two systems and about their view on how gaze information helps or

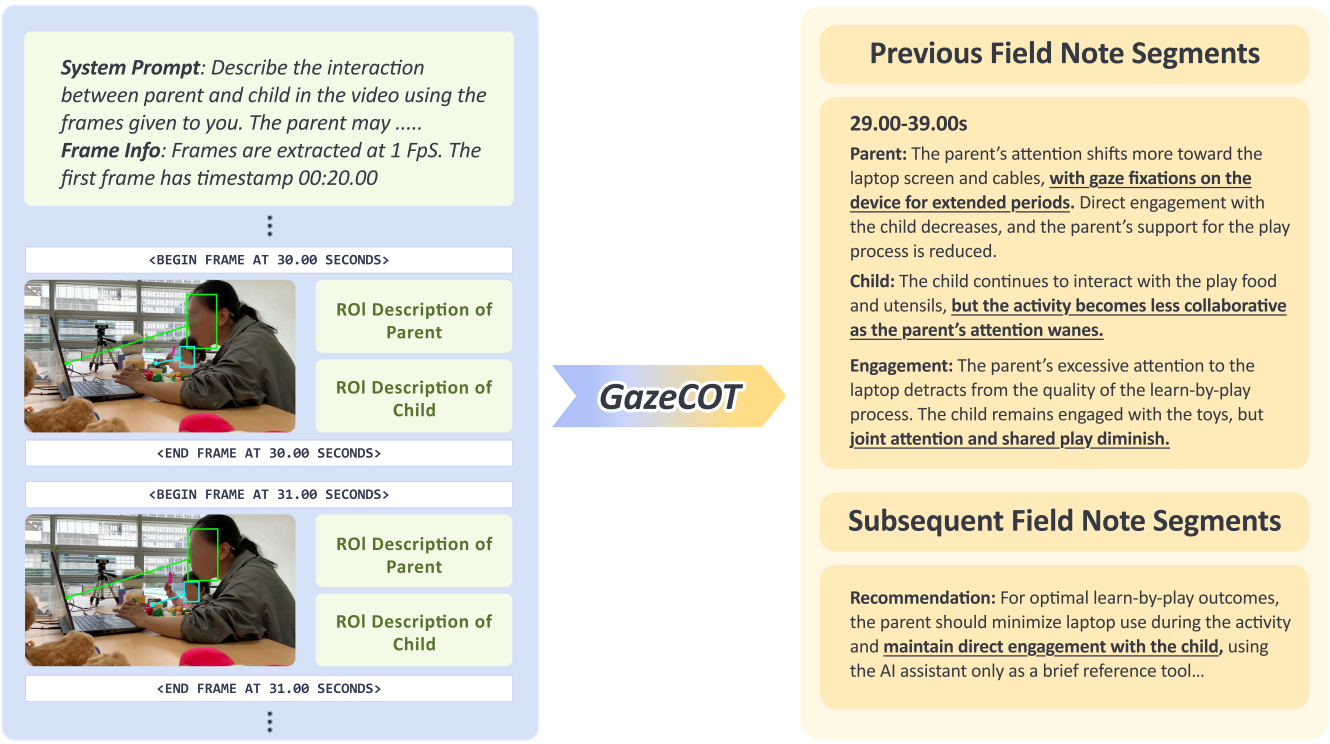


Figure 12: An example of field note generation with GazeCoT. The modified reasoning context is on the left, while a segment of the field note is on the right. Gaze-related lines in the field notes are underlined and bolded. GazeCoT enables accurate joint attention analysis in the field note, and produces actionable advice for the parent. The faces are blurred to protect the participants' privacy.

hinders accurate social perception in this scenario. If the participant mentions a specific clip where the field note left deep good or bad impressions, we would record that case for further analysis. The interview script is provided in Appendix H.

Statistical Analysis. For each user, we average their ratings over the 11 field notes. Therefore, for each metric (e.g. accuracy, usefulness, etc.), each user has one average rating for GazeCoT and another for the baseline. We treat these average ratings as continuous values. Since the difference between GazeCoT and the baseline passes the Shapiro-Wilk normality test ($p>0.05$) for all metrics, we use the paired t-test. We use $p<0.05$ as the threshold of statistical significance.

6.5 Results and Analysis

6.5.1 GazeCoT Improves Output Quality. As shown in Figure 13, GazeCoT brings significant improvements in output quality metrics. This result demonstrates GazeCoT's potential to improve MLLMs' performance in complex social perception tasks, making previously out-of-reach tasks like parent-child JME analysis possible. This further confirms the findings in Section 5.2 and provides additional insight on **RQ2**.

GazeCoT Provides Accurate and Comprehensive Attention Analysis. The participants provided various reasons for the perceived improvement in field note quality. Several participants (P1, P2, P4, P6, P7) found that GazeCoT is better at capturing the shift in both individual and joint attention, leading to better description of the dynamics of parent-child interaction and higher ratings in accuracy (GazeCoT 5.38±0.72, Baseline 4.03±0.58, p=0.0006), comprehensiveness (GazeCoT 5.24±0.56, Baseline 4.45±0.45, p=0.0100), and usefulness (GazeCoT 5.20±0.80, Baseline 4.29±0.46, p=0.0032). P6 attributed this to GazeCoT's focus on gaze and told us "*gaze is one of the most important signals of shifts in attention*". Specifically, as noted by P1, P2 and P4, GazeCoT is better at identifying and

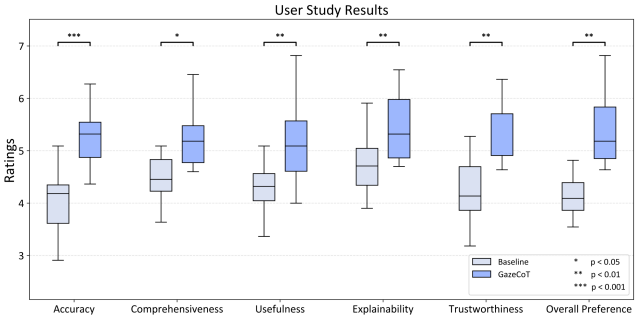


Figure 13: Results of the user study. GazeCoT achieved statistically significant improvements in all 6 metrics.

analyzing cases where joint attention is weak or nonexistent, leading to more insightful and useful overall analysis of parent-child JME. P1 commended GazeCoT's ability to capture brief gaps in joint attention, saying "*Sometimes I would miss short gaps in joint attention when watching the clips on my own. This system (GazeCoT) did a great job finding them. I can then go back to the video and analyze the joint attention gaps myself.*" This shows that explicitly integrating third-person gaze with MLLMs leads to improvements in downstream applications, especially those requiring accurate and comprehensive analysis of human attention. These results also confirm that GazeCoT indeed met design goal **G1** (accurately injecting gaze information).

GazeCoT Reduces Hallucination. P4 mentioned the level of hallucination as the definitive factor in her ratings, saying "*The field notes generated by this system (GazeCoT) have much less made-up content. The other system (baseline) adds too much fictional content and that instantly ruins my impression.*" P7 also stated that "*It (GazeCoT) does not invent non-existent shifts in joint attention, while the other system (baseline) sometimes does.*" This can be attributed to the accurate gaze-grounded attention analysis generated by GazeCoT. In addition, some participants (P4, P6, P7) also found that GazeCoT provides more precise timestamps for the events in the clip, leading to higher accuracy ratings. This can be explained by the structured reasoning context we used in GazeCoT, which reduces hallucination in multi-image situations. These results also indicate that design goal **G2** (reducing hallucination) is fulfilled.

6.5.2 GazeCoT is More Explainable and Trustworthy. The results in Figure 13 demonstrate that, in addition to having better quality, GazeCoT's output is more explainable (GazeCoT 5.45±0.66, Baseline 4.73±0.56, p=0.0018) and trustworthy (GazeCoT 5.24±0.60, Baseline 4.26±0.65, p=0.0041) than the baseline MLLM. Explainability and trust are essential to human-AI interaction, and these results confirm that gaze-informed social reasoning improves both aspects. This addresses **RQ4**.

Gaze as a Common Ground for Explainable Social Perception. Some participants (P1, P2, P4, P10) perceived the field notes generated by GazeCoT as more explainable due to the grounding of conclusions on joint attention in gaze (e.g. there is a lack of joint attention because the parent is looking at the laptop while the child plays on his own, looking down at the toy in his hand). P1 mentioned that "*Gaze provides a common ground between me and the AI, since both of us rely on it to analyze joint attention. This makes the field note feel more explainable.*" This is consistent with prior observation that AI explanations that match human explanatory norms are perceived as better [88].

However, other participants (P6, P7, P9) did not perceive a significant difference in explainability. P7, who has extensive experience with LLMs, said "*I do not think the two are much different, as both systems back their conclusions with seemingly plausible explanations. I am more concerned about the inherent explainability of LLMs, which are essentially black boxes.*" This reflects the perception gap in MLLM explainability between ordinary users and experts, consistent with previous findings on explainable AI [58, 142]. In conclusion, some users with expertise in LLMs remain (rightly) concerned about the gap between perceived plausibility of MLLMs' self-explanations and actual explainability [1, 23, 127].

Building Trust with Gaze-Informed Social Reasoning. As mentioned before, the accuracy of the field notes is the most important factor in determining the participants' impression of a system. The same is true for building trust in the system and is consistent with existing literature on trust in automated systems [48, 69]. Many participants (P1, P3, P4, P5) expressed that the higher accuracy of GazeCoT's field notes is the primary reason they trust it more than the baseline. P1 said "*Its (GazeCoT's) capability to accurately capture shifts in attention, as well as the reasoning process based on gaze observations, that made me trust it more.*" P1's perspective shows that, in addition to accuracy, GazeCoT's more human-like social reasoning also played a part in improving trust. In another example, P2 commented "*It (GazeCoT) uses gaze to analyze the joint engagement of the parent and the child, which is consistent with my own way of analyzing these clips. I trust it more because I know it uses the same analysis techniques as I do.*" P2's comments suggest that, similar to how common ground between users and GazeCoT in social perception improves perceived explainability, common ground in social reasoning and mental model improves trust. P9 also remarked "*The system (GazeCoT) using gaze and posture to analyze parent-child interaction is similar to my mental model.*" The effect of human-AI mental model alignment on trust is similar to that found in non-MLLM automated systems [49, 86].

Despite the higher trust in GazeCoT in general, one participant (P6) found the baseline more trustworthy. P6 said in the interview "*It feels like the system (baseline) wants to tell me more about the clip. The other system (GazeCoT) is not giving me all the details and I trust it less.*" After more in-depth discussion with P6, we identified the cause to be GazeCoT's tendency to prioritize gaze information and joint attention, while the baseline selects more randomly from a broader set of social cues. This led to P6's perceived transparency of the baseline. Therefore, apart from accuracy, the perceived transparency and openness of the AI is also a factor in building trust [155] in MLLM-based artificial social intelligence. However, it is noteworthy that the level of transparency should be balanced, as the overflow of details could induce blind trust in the system [99], leading to worse outcomes.

7 Discussion

7.1 Towards Human-Aligned Social Reasoning in MLLMs

An important observation from our user study is that common ground between human and AI in social perception and social reasoning improves perceived explainability and trust. This leads to a key question: ***how does GazeCoT create such common ground?*** Since a major function of human gaze as a social cue is to indicate a person's attention [12, 34, 37, 40, 65, 92], we provide a qualitative answer from the perspective of guiding and aligning the attention mechanism [130] of MLLMs with human attention, a useful lens for both vision tasks [39, 100, 152] and human-AI interaction [41, 137, 144, 156]. In Figure 14, we visualize how the visual attention for the token *"looking"* changes when presented with visual gaze prompts. The MLLM used here is LLAVA 1.5 7B, whose attention map can be readily visualized with the VaLSe tool[2] proposed by Chen et al.

[2]https://github.com/Ziwei-Zheng/VaLSe

[19]. As shown in the figure, GazeCoT's visual prompts are able to direct more of LLAVA 1.5's attention for the token *looking* to areas humans associate with the concept of *looking*. This provides insight into how GazeCoT works: it guides and aligns MLLM attention with human attention, increasing the influence of human attention on the reasoning process and making it more aligned with human norms. Therefore, visual prompting allows for a deeper, more fundamental integration of gaze information and MLLMs compared to existing "gaze ⇒ text ⇒ LLM" pipelines in HCI research [67, 68].

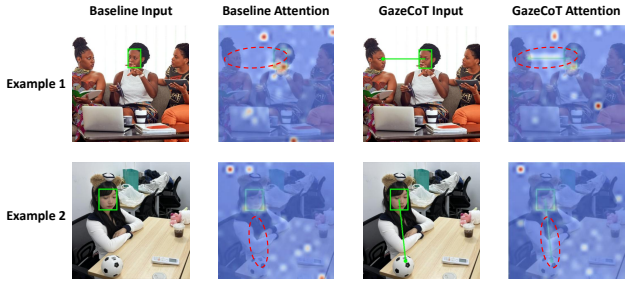


Figure 14: Attention heatmaps for the token "*looking*" in LLAVA 1.5, with and without visual prompting. The model is prompted with the question "*What is the person marked by the bounding box looking at.*" The attention values are from the final transformer block of LLAVA 1.5. The visual prompt (gaze line) is able to guide the MLLM to focus on the direction and target of the person's gaze.

A natural follow-up question would be ***how to expand this common ground in social perception and reasoning***. Fortunately, gaze is not the only non-verbal social cue that conveys human attention, intent, emotion and belief. Other social cues can also be injected into the reasoning context of MLLMs to guide model attention. For example, the GazeCoT pipeline can be directly adapted to work on *pointing*, leveraging existing pointing direction estimation models [90, 128]. Human pose and gestures can also be added to the pipeline, although these would lean more on detailed text description of the pose [57, 164] or gesture [57, 109] rather than direct visual prompting. Similar to gaze, all of these non-verbal social cues require fine-grained visual perception and spatial reasoning, areas where specialized vision models can complement MLLMs.

By empowering MLLMs with the ability to perceive these non-verbal social cues, we can create even more common ground between humans and MLLMs in social perception and social reasoning. This opens up a new paradigm for **human-aligned social reasoning in MLLMs**, which enables MLLMs to understand social cues like humans, aligns social perception and social reasoning between MLLMs and humans, and eventually leads to smoother and more productive human-AI interaction.

7.2 Evolution of GazeCoT: Gaze-based MLLM Social Intelligence for the Future

A key advantage of GazeCoT is the **interchangeability of its pipeline components**. Both the gaze estimation model and the MLLM can be easily switched to newer, more powerful models down the line. Therefore, GazeCoT evolves together with vision encoder backbones, gaze estimation algorithms, and MLLMs.

What Do Better Features Look Like? Better vision encoders should provide smoother, cleaner and more fine-grained dense features. This is evident in the comparison between DINOv2 features and DINOv3 features in Figure 15. Development on this front mainly involve larger-scale training data and novel training techniques [118]. For more dense feature visualizations and discussions on *how* to improve dense feature maps with Gram anchoring during training, we refer to the DINOv3 paper [118]. We expect further progress in general-purpose vision backbones to improve gaze estimation and GazeCoT, as vision encoders are the foundation for both GazeLLE-style gaze estimation and MLLMs.

Figure 15: Comparison of the dense feature maps generated by DINOv2 and DINOv3. The images are the same ones from Figure 3. The feature maps are reduced from 1024 (DINOv2-ViT-L) or 1280 (DINOv3-ViT-H+) dimensions to 3 dimensions through PCA, and visualized as RGB images. DINOv3 features are smoother, more aligned with people and objects, has clearer representation of head and eye features, and has fewer artifacts in the background.

Aligning Gaze Estimation With Human Perception. In previous sections, we discussed aligning MLLM perception with human perception. A similar perspective can be applied to gaze estimation. Despite achieving metrics that *appear* superhuman (AUC and L2), GazeLLE-v3-H still significantly underperformed humans in the gaze target recognition benchmark. This suggests a misalignment between scene-agnostic metrics like L2 and human perception.

The case in Figure 16 demonstrates that the effect of L2 errors on downstream tasks (product-level and object-level gaze target recognition) is **anisotropic**, varying greatly depending on the direction of the error. This calls for the wider adoption of metrics such as task-level accuracy and object-level accuracy. The same principles can be applied to model training as well. The attempt by Tafasca et al. [121] to incorporate object labels into the loss function showed some success, but much work remains to be done. One potential approach is to use semantic segmentation models such as SAM [61, 102] to complement the scene-agnostic loss function with scene-specific object segmentation information, rewarding

predictions that align with the object-level and task-level semantics. Progress in this direction would lead to gaze estimation models that are more capable of decoding human attention, benefiting downstream HCI applications.

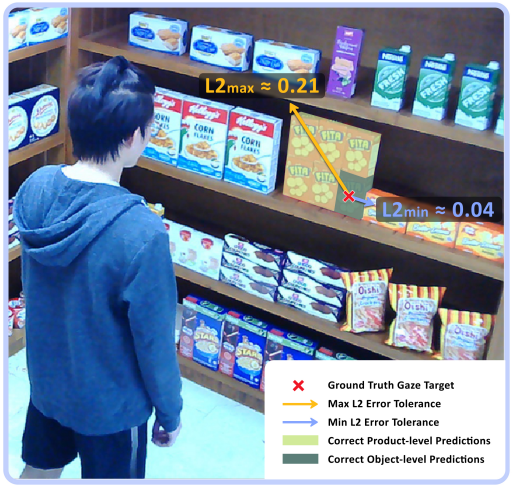


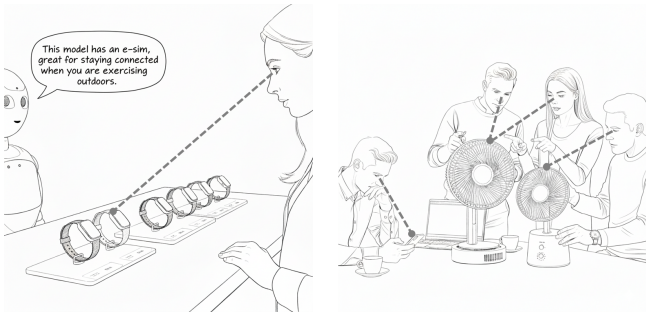
Figure 16: How L2 metrics deviate from human expectations in gaze target estimation. In this example from the GOO-Real dataset [125], we consider gaze target predictions within the light green area (on any of the 6 FITA boxes) to be correct for the task of *recognizing which product the customer is looking at*, and those within the dark green area (on the ground truth FITA box) to be correct for the task of *recognizing which object the customer is looking at*. The tolerance for L2 error varies significantly across different directions: tiny errors in the blue direction would be catastrophic, while errors in the orange direction are tolerated more by downstream tasks.

GazeCoT for MLLM Training. As discussed in Section 2.2, a major bottleneck in MLLM perception is the lack of image-text pairs with fine-grained captions. GazeCoT can be adapted to provide fine-grained captions on gaze and social interaction in images. This is relevant to the generation of synthetic training data, which has been widely used in LLM and MLLM training [7, 76, 112, 163]. If GazeCoT is able to mitigate the lack of gaze-related training data, we can expect future MLLMs to be better at perceiving gaze and understanding social interactions, which in turn improves GazeCoT. This would usher in a virtuous cycle in gaze-aware MLLM development.

7.3 Adapting GazeCoT to Downstream HCI Applications

A wide range of HCI applications rely on socially aware multimodal AI, and benefit from the gaze-informed social reasoning afforded by GazeCoT. Examples include:

- **Human-robot interaction (HRI)**. Robots often need to understand user instructions or infer implicit user intent (Figure 17a). In these cases, GazeCoT can help the robot decode the user's intent directly from gaze, as envisioned by previous work [83]. This reduces the need for excessive verbal instructions and leads to smarter and more intuitive HRI.
- **Workplace collaboration**. Prior work on using first person gaze to improve collaboration in VR [10, 11] demonstrated the potential of gaze in facilitating better collaboration. Gaze-based co-located collaboration has also been explored but was limited to screens with attached eye trackers [159]. GazeCoT can expand the modeling of joint attention and mutual awareness to MLLM-based applications (Figure 17b) in generic physical spaces [16, 162], enabling more intelligent moderation of conferences, discussions and brainstorming without the need for eye tracking hardware.
- **Early childhood education**, which we explored in our user study, can also benefit from GazeCoT. Prior work [63] directly combines gaze and close-ended object recognition to describe play context. Extending this paradigm, GazeCoT allows for open-ended and versatile interpretation of gaze and surrounding visual context, enabling more fine-grained, socially-aware content generation for early childhood education.
- **MLLM-powered accessibility applications** for blind and low-vision (BLV) users. Prior work like WorldScribe [18] uses MLLMs to generate adaptive real-world visual descriptions for BLV users. GazeCoT can expand such systems to describe non-verbal social dynamics.
- **Creativity and performative applications** [30]. GazeCoT may be used to model the attention of the audience and performer, providing feedback on improving the performance. We hope GazeCoT can inspire future work exploring the integration of third-person gaze and MLLMs in diverse HCI applications.



(a) (b)

Figure 17: Two potential downstream applications of GazeCoT. (a) GazeCoT can support natural human-robot interaction by improving gaze understanding, enabling the robot to provide shopping advice proactively. (b) GazeCoT can analyze workplace discussions by identifying each person's focus, allowing an MLLM-based AI moderator to guide the discussion to be more comprehensive and productive and to provide detailed personalized recaps afterwards.

7.4 Ethical Implications

The primary focus of this paper is on the technical and user experience aspects of integrating third-person gaze into MLLM social

reasoning. However, we acknowledge that this topic also has profound ethical implications, especially on privacy and surveillance [134, 136]. Specialized models for tasks like gaze estimation and pose estimation can perceive social cues but cannot interpret them [64, 132]. MLLMs can interpret text-based cues but fail at social perception [73, 161]. GazeCoT combines the strengths of the two, potentially enabling invasive and unethical forms of fully automated *cognitive surveillance*.

The potential applications of GazeCoT fall into two categories in terms of ethical concerns. The first type of application is ***cooperative***, where the individuals whose attention and intent are analyzed by GazeCoT are the users who proactively initiated the application and stand to benefit directly. Some examples of cooperative applications are HRI (especially in home settings) and personal AI assistants. In these cases, informed consent is paramount to mitigating ethical concerns. Furthermore, on-device inference leveraging lightweight MLLMs (e.g. Qwen2.5 VL 3B, Phi 4, etc.) could largely eliminate the risk of sending away sensitive data. For applications where state-of-the-art proprietary MLLMs are needed, GazeLLE-v3 can be hosted locally, with the cloud API receiving only gaze estimations and processed images in which the faces are anonymized. Finally, even in cooperative use cases, bystanders' privacy needs to be considered.

The second type of application is ***non-cooperative***. This refers to cases where individuals' gaze and cognitive state are analyzed without their proactive input or direct benefits to them. Examples include advanced security surveillance and employee surveillance. **We strongly caution against such cases of non-cooperative cognitive surveillance, which pose severe privacy and ethical risks.** We advocate for strict governance that limits gaze-aware MLLMs to user-authorized cooperative scenarios with informed consent and proper privacy protections. As things stand currently, massive use of MLLMs on such dense tasks like cognitive surveillance remains very costly compute-wise [149], which limits real-world applications. This provides academia, the public, and regulators a time window to establish the type of robust governance needed for the ethical and responsible use of gaze- and socially-aware MLLMs as part of the broader vision of ethical AI [28, 138, 146].

7.5 Limitations and Future Work

7.5.1 Latency, Cost, and Pipeline Optimization. GazeCoT requires an additional MLLM inference step due to the ROI description tool. On the gaze target recognition benchmark, the GPT-4.1 version of the GazeCoT pipeline took on average 11.0 seconds per question. Removing the ROI description step reduces that to 5.9 seconds. In comparison, the baseline MLLM took 5.2 seconds. GazeCoT also introduces more image tokens in the Task Agent's context, resulting in higher token usage. In the same experiment, GazeCoT used on average 182 prompt tokens and 97 completion tokens (182P/97C) for ROI description, and 812P/125C for the Task Agent per question. Removing ROI description reduces token usage to 658P/67C. The baseline used 365P/79C, with the savings primarily coming from the absence of the visually prompted image. For reference, for GPT-4.1, OpenAI charges $2 for 1 million prompt tokens, and $8 for 1 million completion tokens at the time of writing.

Therefore, latency-sensitive applications, especially those working in relatively simple environments, can skip the ROI description step. Another way to mitigate the higher latency of GazeCoT is to use locally-hosted open source MLLMs such as Qwen2.5 VL and switching to streaming output. This way, it is possible to integrate gaze estimation into real-time MLLM-powered interactions. Since the main focus of this paper is on enabling general-purpose third-person gaze understanding in MLLMs, we leave streamlining GazeCoT to more efficiently meet the demands of downstream applications as future work.

7.5.2 From Face Detection to Head Detection. GazeCoT uses a face detection model (RetinaFace[3]), which does not work well for individuals not facing the camera. This is problematic in scenarios with certain camera angles (e.g. ceiling-mounted camera, a robot following a person, etc.). A fix would be switching to head detection, potentially by training a YOLO detection model [124] on head detection datasets [97, 133]. GazeLLE-v3 can be used as-is, as it is capable of dealing with individuals not facing the camera.

7.5.3 User Studies in More Diverse Scenarios. We acknowledge that the user experiment is limited to a single scenario, parent-child JME analysis during play. While it is a complex scenario difficult for standalone MLLMs, it is far from covering the full scope of artificial social intelligence. In addition, our users were experts in parent-child interaction. Their perception of explainability and trust, though valuable, may differ from that of ordinary users. Future work on integrating third-person gaze directly into MLLM-based interactive systems, such as those outlined in Section 7.3, would be highly valuable, and would provide insight on how gaze-informed MLLMs change the interaction process and user experience.

8 Conclusion

We introduced GazeCoT, a plug-and-play pipeline that combines the strengths of gaze estimation models and MLLMs. GazeCoT works by injecting third-person gaze predictions made by our state-of-the-art GazeLLE-v3 model into the CoT reasoning context of MLLMs in the form of hybrid visual and text gaze prompts, significantly improving MLLMs' ability to perceive human gaze and engage in socially intelligent CoT reasoning. We demonstrated GazeCoT's ability to improve MLLMs' social intelligence through experiments on two benchmarks and a user study. The results showed improved social intelligence, explainability, and trustworthiness, which can be partially attributed to gaze acting as a common ground between human and MLLM social perception and social reasoning. Finally, we discussed how GazeCoT aligns human and AI social perception and social reasoning, how introducing other non-verbal social cues can further strengthen that alignment, how GazeCoT can be improved and adapted to more HCI applications, and the ethical implications of socially-aware MLLMs. We hope GazeCoT can inspire more work on enabling MLLMs to truly understand people and social interactions, as well as work on socially intelligent interactive systems.

[3] https://github.com/serengil/retinaface

Acknowledgments

This work is supported by National Key R&D Program of China under Grant No. 2024YFB4505500 & 2024YFB4505501, Beijing Key Lab of Networked Multimedia, Institute of Artificial Intelligence, Tsinghua University (THUAI), College of AI, Tsinghua University, and Beijing National Research Center for Information Science and Technology (BNRist).

References

[1] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. arXiv:2402.04614 [cs.CL] https://arxiv.org/abs/2402.04614

[2] Deepti Aggarwal, Robyn Garnett, Bernd Ploderer, Frank Vetere, Patricia Eadie, and Bronwyn Joy Davidson. 2015. Understanding Video based Parent Training Intervention for Children with Autism. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction* (Parkville, VIC, Australia) *(OzCHI '15)*. Association for Computing Machinery, New York, NY, USA, 10–19. doi:10.1145/2838739.2838770

[3] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Can LLMs Reason Like Humans? Assessing Theory of Mind Reasoning in LLMs for Open-Ended Questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 34–44. doi:10.1145/3627673.3679832

[4] William Sims Bainbridge, Edward E. Brent, Kathleen M. Carley, David R. Heise, Michael W. Macy, Barry Markovsky, and John Skvoretz. 1994. Artificial Social Intelligence. *Annual Review of Sociology* 20, Volume 20, 1994 (1994), 407–436. doi:10.1146/annurev.so.20.080194.002203

[5] Jun Bao, Buyu Liu, and Jun Yu. 2022. ESCNet: Gaze Target Detection With the Understanding of 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14126–14135.

[6] Andrew P. Bayliss, Jessica Bartlett, Claire K. Naughtin, and Ada Kritikos. 2011. A direct link between gaze perception and social attention. *Journal of Experimental Psychology: Human Perception and Performance* 37, 3 (2011), 634–644. doi:10.1037/a0020559

[7] Aidan Bell, James Gore, and Behrooz Mansouri. 2025. Augmenting Vision-Language Retrieval: The Role of Multimodal LLMs as Synthetic Data Generators. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 3050–3054. doi:10.1145/3726302.3730167

[8] Margrit Betke. 2010. *Intelligent Interfaces to Empower People with Disabilities*. Springer US, Boston, MA, 409–432. doi:10.1007/978-0-387-93808-0_15

[9] Riccardo Bovo, Steven Abreu, Karan Ahuja, Eric J Gonzalez, Li-Te Cheng, and Mar Gonzalez-Franco. 2025. EmBARDiment: an Embodied AI Agent for Productivity in XR. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 708–717. doi:10.1109/VR59515.2025.00093

[10] Riccardo Bovo, Daniele Giunchi, Muna Alebri, Anthony Steed, Enrico Costanza, and Thomas Heinis. 2022. Cone of Vision as a Behavioural Cue for VR Collaboration. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 502 (Nov. 2022), 27 pages. doi:10.1145/3555615

[11] Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Joshua Newn, Hans Gellersen, Enrico Costanza, and Thomas Heinis. 2023. Speech-Augmented Cone-of-Vision for Exploratory Data Analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 162, 18 pages. doi:10.1145/3544548.3581283

[12] Rechele Brooks and Andrew N. Meltzoff. 2005. The development of gaze following and its relation to language. *Developmental Science* 8, 6 (2005), 535–543. doi:10.1111/j.1467-7687.2005.00445.x

[13] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The Revolution of Multimodal Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13590–13618. doi:10.18653/v1/2024.findings-acl.807

[14] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. ViP-LLaVA: Making Large Multimodal Models Understand Arbitrary Visual Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12914–12923.

[15] Runze Cai, Nuwan Janaka, Hyeongcheol Kim, Yang Chen, Shengdong Zhao, Yun Huang, and David Hsu. 2025. AiGet: Transforming Everyday Moments into Hidden Knowledge Discovery with AI Assistance on Smart Glasses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 631, 26 pages. doi:10.1145/3706598.3713953

[16] Zhenyao Cai, Seehee Park, Nia Nixon, and Shayan Doroudi. 2024. Advancing Knowledge Together: Integrating Large Language Model-based Conversational AI in Small Group Collaborative Learning. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 37, 9 pages. doi:10.1145/3613905.3650868

[17] Yiming Cao, Zhen Li, Lizhen Cui, and Chunyan Miao. 2025. Adaptive Human-LLMs Interaction Collaboration: Reinforcement Learning driven Vision-Language Models for Medical Report Generation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 62, 6 pages. doi:10.1145/3706599.3719852

[18] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 140, 18 pages. doi:10.1145/3654777.3676375

[19] Boxu Chen, Ziwei Zheng, Le Yang, Zeyu Geng, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025. Seeing It or Not? Interpretable Vision-aware Latent Steering to Mitigate Object Hallucinations. arXiv:2505.17812 [cs.CV] https://arxiv.org/abs/2505.17812

[20] Huili Chen, Yubin Kim, Kejia Patterson, Cynthia Breazeal, and Hae Won Park. 2025. Social robots as conversational catalysts: Enhancing long-term human-human interaction at home. *Science Robotics* 10, 100 (2025), eadk3307. doi:10.1126/scirobotics.adk3307

[21] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models. arXiv:2503.09567 [cs.AI] https://arxiv.org/abs/2503.09567

[22] Wenhe Chen, Hui Xu, Chao Zhu, Xiaoli Liu, Yinghua Lu, Caixia Zheng, and Jun Kong. 2022. Gaze Estimation via the Joint Modeling of Multiple Cues. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2022), 1390–1402. doi:10.1109/TCSVT.2021.3071621

[23] Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2024. Do models explain themselves? counterfactual simulatability of natural language explanations. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 310, 25 pages.

[24] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv:2412.05271 [cs.CV] https://arxiv.org/abs/2412.05271

[25] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15959–15983. doi:10.18653/v1/2024.acl-long.847

[26] Jouh Yeong Chew and Xiaohan Wang. 2024. Joint Attention Estimation during Multi-party Facilitation Using Multi-Modal Fusion. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 322–326. doi:10.1145/3610978.3640669

[27] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. 2020. Detecting Attended Visual Targets in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5396–5406.

[28] Nicholas Kluge Corrêa, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, Rodrigo Mambrini, Luiza Galvão, Edmund Terem, and Nythamar de Oliveira. 2023. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns* 4, 10 (2023). doi:10.1016/j.patter.2023.100857

[29] Mustafa Doga Dogan, Eric J Gonzalez, Karan Ahuja, Ruofei Du, Andrea Colaço, Johnny Lee, Mar Gonzalez-Franco, and David Kim. 2024. Augmented Object Intelligence with XR-Objects. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 19, 15 pages. doi:10.1145/3654777.3676379

[30] Riccardo Drago, Yotam Sechayk, Mustafa Doga Dogan, Andrea Sanna, and Takeo Igarashi. 2025. ImprovMate: Multimodal AI Assistant for Improv Actor

Training. In *Companion Publication of the 2025 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, New York, NY, USA, 526–532. https://doi.org/10.1145/3715668.3736363

[31] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 340, 20 pages.

[32] Zhizhao Duan, Hao Cheng, Duo Xu, Xi Wu, Xiangxie Zhang, Xi Ye, and Zhen Xie. 2024. CityLLaVA: Efficient Fine-Tuning for VLMs in City Scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 7180–7189.

[33] Robert M Emerson, Rachel I Fretz, and Linda L Shaw. 1995. Writing ethnographic fieldnotes. Chicago guides to writing, editing. *and publishing. Chicago, IL: University of Chicago Press. Feld, S., & Brenneis, D.(2004). Doing anthropology in sound. American Ethnologist* 31, 4 (1995), 461–474.

[34] N.J. Emery. 2000. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews* 24, 6 (2000), 581–604. doi:10.1016/S0149-7634(00)00025-7

[35] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. 2018. Inferring Shared Attention in Social Scene Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6460–6468.

[36] Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, and Yixin Zhu. 2021. Learning Triadic Belief Dynamics in Nonverbal Communication From Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7312–7321.

[37] Lifeng Fan, Manjie Xu, Zhihao Cao, Yixin Zhu, and Song-Chun Zhu. 2022. Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research* 1, 2 (2022), 144–160.

[38] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. 2021. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11390–11399.

[39] Thomas FEL, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. 2022. Harmonizing the object recognition strategies of deep neural networks with humans. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 9432–9446. https://proceedings.neurips.cc/paper_files/paper/2022/file/3d681cc4487b97c08e5aa67224dd74f2-Paper-Conference.pdf

[40] Alexandra Frischen, Andrew P. Bayliss, and Steven P. Tipper. 2007. Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin* 133, 4 (2007), 694–724. doi:10.1037/0033-2909.133.4.694 PsycINFO Database Record (c) 2025 APA, all rights reserved.

[41] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022. Aligning Eyes between Humans and Deep Neural Network through Interactive Attention Alignment. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 489 (Nov. 2022), 28 pages. doi:10.1145/3555590

[42] Khashayar Ghamati, Maryam Banitalebi Dehkordi, and Abolfazl Zaraki. 2025. Which AI Sees Like Us? Investigating the Cognitive Plausibility of Language and Vision Models via Eye-Tracking in Human-Robot Interaction. *Sensors* 25, 15 (2025). doi:10.3390/s25154687

[43] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2025. A survey on LLM-as-a-Judge. *The Innovation* (2025). doi:10.1016/j.xinn.2025.101253 doi: 10.1016/j.xinn.2025.101253.

[44] Ananya Gubbi Mohanbabu and Amy Pavel. 2024. Context-Aware Image Descriptions for Web Accessibility. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 62, 17 pages. doi:10.1145/3663548.3675658

[45] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. 2024. RegionGPT: Towards Region Understanding Vision Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13796–13806.

[46] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2025. LLaVA-UHD: An LMM Perceiving Any Aspect Ratio and High-Resolution Images. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 390–406.

[47] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. 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 (CVPR) Workshops*. 5041–5050.

[48] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[49] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* Volume 5 - 2023 (2023). doi:10.3389/fcomp.2023.1096257

[50] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030

[51] Baosheng James Hou, Joshua Newn, Ludwig Sidenmark, Anam Ahmad Khan, and Hans Gellersen. 2024. GazeSwitch: Automatic Eye-Head Mode Switching for Optimised Hands-Free Pointing. *Proc. ACM Hum.-Comput. Interact.* 8, ETRA, Article 227 (May 2024), 20 pages. doi:10.1145/3655601

[52] Xiyun Hu, Dizhi Ma, Fengming He, Zhengzhe Zhu, Shao-Kang Hsia, Chenfei Zhu, Ziyi Liu, and Karthik Ramani. 2025. GesPrompt: Leveraging Co-Speech Gestures to Augment LLM-Based Interaction in Virtual Reality. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 59–80. doi:10.1145/3715336.3735769

[53] Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 5880–5895. doi:10.18653/v1/2025.findings-acl.306

[54] Runhui Huang, Xinpeng Ding, Chunwei Wang, Jianhua Han, Yulong Liu, Hengshuang Zhao, Hang Xu, Lu Hou, Wei Zhang, and Xiaodan Liang. 2025. HiRes-LLaVA: Restoring Fragmentation Input in High-Resolution Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 29814–29824.

[55] Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. 2023. NLIP: Noise-Robust Language-Image Pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 1 (Jun. 2023), 926–934. doi:10.1609/aaai.v37i1.25172

[56] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. MotionGPT: Human Motion as a Foreign Language. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 20067–20079. https://proceedings.neurips.cc/paper_files/paper/2023/file/3fbf0c1ea0716c03dea93bb6be78dd6f-Paper-Conference.pdf

[57] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. 2019. Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in a Triadic Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10873–10883.

[58] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376219

[59] Jinwook Kim, Sangmin Park, Qiushi Zhou, Mar Gonzalez-Franco, Jeongmi Lee, and Ken Pfeuffer. 2025. PinchCatcher: Enabling Multi-selection for Gaze+Pinch. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 853, 16 pages. doi:10.1145/3706598.3713530

[60] Yunsoo Kim, Jinge Wu, Yusuf Abdulle, Yue Gao, and Honghan Wu. 2024. Enhancing Human-Computer Interaction in Chest X-Ray Analysis Using Vision and Language Model with Eye Gaze Patterns. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (Eds.). Springer Nature Switzerland, Cham, 184–194.

[61] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4015–4026.

[62] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2024. You'll Never Walk Alone: A Sketch and Text Duet for Fine-Grained Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16509–16519.

[63] Taeahn Kwon, Minkyung Jeong, Eon-Suk Ko, and Youngki Lee. 2022. Captivate! Contextual Language Guidance for Parent–Child Interaction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 219, 17 pages. doi:10.1145/3491102.3501865

[64] Dennis Küster, Eva G. Krumhuber, Lars Steinert, Anuj Ahuja, Marc Baker, and Tanja Schultz. 2020. Opportunities and Challenges for Using Automatic Human Affect Analysis in Consumer Research. *Frontiers in Neuroscience* Volume 14 - 2020 (2020). doi:10.3389/fnins.2020.00400

[65] Stephen R. H. Langton, Roger J. Watt, and Vicki Bruce. 2000. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences* 4, 2 (Feb. 2000), 50–59. doi:10.1016/S1364-6613(99)01436-9

[66] Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Achin Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, David Kim, and Ruofei Du. 2025. Sensible Agent: A Framework for Unobtrusive Interaction with Proactive AR Agents. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 119, 22 pages. doi:10.1145/3746059.3747748

[67] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 408, 20 pages. doi:10.1145/3613904.3642230

[68] Jaewook Lee, Tianyi Wang, Jacqui Fashimpaur, Naveen Sendhilnathan, and Tanya R. Jonker. 2025. Walkie-Talkie: Exploring Longitudinal Natural Gaze, LLMs, and VLMs for Query Disambiguation in XR. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 586, 9 pages. doi:10.1145/3706599.3720236

[69] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[70] Phillip Y. Lee, Taehoon Yoon, and Minhyuk Sung. 2024. GrounDiT: Grounding Diffusion Transformers via Noisy Patch Transplantation. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 58610–58636. doi:10.52202/079017-1868

[71] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 17044–17068. doi:10.52202/079017-0542

[72] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326 [cs.CV] https://arxiv.org/abs/2408.03326

[73] Hao Li, Hao Fei, Zechao Hu, Zhengwei Yang, and Zheng Wang. 2025. VEGAS: Towards Visually Explainable and Grounded Artificial Social Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 5 (Apr. 2025), 4707–4715. doi:10.1609/aaai.v39i5.32497

[74] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19730–19742. https://proceedings.mlr.press/v202/li23q.html

[75] Jiale Li, Mingrui Wu, Zixiang Jin, Hao Chen, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. 2025. MIHBench: Benchmarking and Mitigating Multi-Image Hallucinations in Multimodal Large Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 3143–3152. doi:10.1145/3746027.3754993

[76] Ziyan Li, Jianfei Yu, Jia Yang, Wenya Wang, Li Yang, and Rui Xia. 2024. Generative Multimodal Data Augmentation for Low-Resource Multimodal Named Entity Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 7336–7345. doi:10.1145/3664647.3681598

[77] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

[78] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A Survey on Hallucination in Large Vision-Language Models. arXiv:2402.00253 [cs.CV] https://arxiv.org/abs/2402.00253

[79] Yang Liu, Thorbjørn Mikkelsen, Zehai Liu, Gengchen Tian, Diako Mardanbegi, Qiushi Zhou, Hans Gellersen, and Ken Pfeuffer. 2025. At a Glance to Your Fingertips: Enabling Direct Manipulation of Distant Objects Through SightWarp. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. doi:10.1145/3746059.3747653

[80] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Skq89Scxx

[81] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. WildVision: Evaluating Vision-Language Models in the Wild with Human Preferences. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 48224–48255. doi:10.52202/079017-1528

[82] Mathias N. Lystbæk, Thorbjørn Mikkelsen, Roland Krisztandl, Eric J Gonzalez, Mar Gonzalez-Franco, Hans Gellersen, and Ken Pfeuffer. 2024. Hands-on, Hands-off: Gaze-Assisted Bimanual 3D Interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 80, 12 pages. doi:10.1145/3654777.3676331

[83] Hanfang Lyu, Xiaoyu Wang, Nandi Zhang, Shuai Ma, Qian Zhu, Yuhan Luo, Fugee Tsung, and Xiaojuan Ma. 2025. Signaling Human Intentions to Service Robots: Understanding the Use of Social Cues during In-Person Conversations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 603, 21 pages. doi:10.1145/3706598.3714235

[84] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 107–110. doi:10.1145/1357054.1357072

[85] Aoran Mei, Guo-Niu Zhu, Huaxiang Zhang, and Zhongxue Gan. 2024. ReplanVLM: Replanning Robotic Tasks With Visual Language Models. *IEEE Robotics and Automation Letters* 9, 11 (2024), 10201–10208. doi:10.1109/LRA.2024.3471457

[86] Michael Merry, Pat Riddle, and Jim Warren. 2021. A mental models approach for defining explainable artificial intelligence. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 344. doi:10.1186/s12911-021-01703-7

[87] Alaeddine Mihoub and Grégoire Lefebvre. 2017. Social Intelligence Modeling using Wearable Devices. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 331–341. doi:10.1145/3025171.3025195

[88] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. doi:10.1016/j.artint.2018.07.007

[89] Seyed Mahed Mousavi, Edoardo Cecchinato, Lucia Hornikova, and Giuseppe Riccardi. 2025. Garbage In, Reasoning Out? Why Benchmark Scores are Unreliable and What to Do About It. arXiv:2506.23864 [cs.CL] https://arxiv.org/abs/2506.23864

[90] Shu Nakamura, Yasutomo Kawanishi, Shohei Nobuhara, and Ko Nishino. 2023. DeePoint: Visual Pointing Recognition and Direction Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20577–20587.

[91] Debosmit Neogi, Nataraj Das, and Suman Deb. 2022. *Eye-Gaze Based Hands Free Access Control System for Smart City Public Interfaces*. Springer Nature Singapore, Singapore, 139–156. doi:10.1007/978-981-16-7498-3_9

[92] Lauri Nummenmaa and Andrew J. Calder. 2009. Neural mechanisms of social attention. *Trends in Cognitive Sciences* 13, 3 (March 2009), 135–143. doi:10.1016/j.tics.2008.12.006

[93] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=a68SUt6zFt Featured Certification.

[94] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. 2016. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5048–5054. doi:10.1109/IROS.2016.7759741

[95] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 456, 20 pages. doi:10.1145/3706598.3713726

[96] Kevin A Pelphrey, Jeffrey D Singerman, Truett Allison, and Gregory McCarthy. 2003. Brain activation evoked by perception of gaze shifts: the influence of context. *Neuropsychologia* 41, 2 (2003), 156–170. doi:10.1016/S0028-3932(02)00146-X The cognitive neuroscience of social behavior.

[97] Dezhi Peng, Zikai Sun, Zirong Chen, Zirui Cai, Lele Xie, and Lianwen Jin. 2018. Detecting Heads using Feature Refine Net and Cascaded Multi-scale Architecture. In *2018 24th International Conference on Pattern Recognition (ICPR)*. 2528–2533. doi:10.1109/ICPR.2018.8545068

[98] Melina Petsolari, Seray B Ibrahim, and Petr Slovak. 2024. Socio-technical Imaginaries: Envisioning and Understanding AI Parenting Supports through Design Fiction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing*

Systems (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 98, 27 pages. doi:10.1145/3613904.3642619

[99] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. doi:10.1145/3411764.3445315

[100] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring Human-Like Attention Supervision in Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). doi:10.1609/aaai.v32i1.12272

[101] Argenis Ramirez Ramirez Gomez, Christopher Clarke, Ludwig Sidenmark, and Hans Gellersen. 2021. Gaze+Hold: Eyes-only Direct Manipulation with Continuous Gaze Modulated by Closure of One Eye. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Full Papers)*. Association for Computing Machinery, New York, NY, USA, Article 10, 12 pages. doi:10.1145/3448017.3457381

[102] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. 2025. SAM 2: Segment Anything in Images and Videos. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=Ha6RTeWMd0

[103] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/ec8956637a99787bd197eacd77acce5e-Paper.pdf

[104] Jun Rekimoto. 2025. GazeLLM: Multimodal LLMs incorporating Human Visual Attention. In *Proceedings of the Augmented Humans International Conference 2025 (AHs '25)*. Association for Computing Machinery, New York, NY, USA, 302–311. doi:10.1145/3745900.3746075

[105] Evan F Risko, Daniel C Richardson, and Alan Kingstone. 2016. Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science* 25, 1 (2016), 70–74.

[106] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M. Rehg. 2025. Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 28874–28884.

[107] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4463–4473. doi:10.18653/v1/D19-1454

[108] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. 2018. Human Gaze Following for Human-Robot Interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 8615–8621. doi:10.1109/IROS.2018.8593580

[109] Debajit Sarma and Manas Kamal Bhuyan. 2021. Methods, databases and recent advancement of vision-based hand gesture recognition for hci systems: A review. *SN Computer Science* 2, 6 (2021), 436. https://doi.org/10.1007/s42979-021-00827-x

[110] Shreshth Saxena, Areez Visram, Neil Lobo, Zahid Mirza, Mehak Khan, Biranugan Pirabaharan, Alexander Nguyen, and Lauren K Fink. 2025. SocialEyes: Scaling Mobile Eye-tracking to Multi-person Social Settings. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 751, 19 pages. doi:10.1145/3706598.3713910

[111] Fiona Louise Scott. 2022. Family mediation of preschool children's digital media practices at home. *Learning, Media and Technology* 47, 2 (2022), 235–250. doi:10.1080/17439884.2021.1960859

[112] Amrith Setlur, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. 2024. RL on Incorrect Synthetic Data Scales the Efficiency of LLM Math Reasoning by Eight-Fold. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 43000–43031. doi:10.52202/079017-1361

[113] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 8612–8642. doi:10.52202/079017-0275

[114] Aditya Sharma, Michael Saxon, and William Yang Wang. 2024. Losing Visual Needles in Image Haystacks: Vision Language Models are Easily Distracted in Short and Long Contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 5429–5451. doi:10.18653/v1/2024.findings-emnlp.312

[115] Weiyan Shi, Hai Viet Le, and Kenny Tsu Wei Choo. 2025. Towards Multimodal Large-Language Models for Parent-Child Interaction: A Focus on Joint Attention. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 535, 6 pages. doi:10.1145/3706599.3720215

[116] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does CLIP know about a red circle? Visual prompt engineering for VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11987–11997.

[117] Ludwig Sidenmark, Franziska Prummer, Joshua Newn, and Hans Gellersen. 2023. Comparing Gaze, Head and Controller Selection of Dynamically Revealed Targets in Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics* 29, 11 (2023), 4740–4750. doi:10.1109/TVCG.2023.3320235

[118] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 2025. DINOv3. arXiv:2508.10104 [cs.CV] https://arxiv.org/abs/2508.10104

[119] Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. 2025. VLM-Social-Nav: Socially Aware Robot Navigation Through Scoring Using Vision-Language Models. *IEEE Robotics and Automation Letters* 10, 1 (2025), 508–515. doi:10.1109/LRA.2024.3511409

[120] Lisa J Stephenson, S Gareth Edwards, and Andrew P Bayliss. 2021. From gaze perception to social cognition: The shared-attention system. *Perspectives on Psychological Science* 16, 3 (2021), 553–576.

[121] Samy Tafasca, Anshul Gupta, Victor Bros, and Jean-Marc Odobez. 2024. Toward Semantic Gaze Target Detection. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 121422–121448. doi:10.52202/079017-3858

[122] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. 2023. ChildPlay: A New Benchmark for Understanding Children's Gaze Behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20935–20946.

[123] Yilin Tang, Yuyang Fang, Tianle Wang, Lingyun Sun, and Liuqing Chen. 2025. "This is My Fault", Really? Understanding Blind and Low-Vision People's Perception of Hallucination in Large Vision Language Models. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 44, 20 pages. doi:10.1145/3746059.3747597

[124] Yunjie Tian, Qixiang Ye, and David Doermann. 2025. YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv:2502.12524 [cs.CV] https://arxiv.org/abs/2502.12524

[125] Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. 2021. GOO: A Dataset for Gaze Object Prediction in Retail Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 3125–3133.

[126] Hsin-Ruey Tsai, Shih-Kang Chiu, and Bryan Wang. 2025. GazeNoter: Co-Piloted AR Note-Taking via Gaze Selection of LLM Suggestions to Match Users' Intentions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 32, 22 pages. doi:10.1145/3706598.3714294

[127] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 74952–74965. https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf

[128] Satoshi Ueno, Sei Naito, and Tsuhan Chen. 2014. An efficient method for human pointing estimation for robot interaction. In *2014 IEEE International Conference on Image Processing (ICIP)*. 1545–1549. doi:10.1109/ICIP.2014.7025309

[129] Franz A. Van-Horenbeke and Angelika Peer. 2021. Activity, Plan, and Goal Recognition: A Review. *Frontiers in Robotics and AI* Volume 8 - 2021 (2021). doi:10.3389/frobt.2021.643010

[130] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[131] Roel Vertegaal. 1999. The GAZE groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA,

294–301. doi:10.1145/302979.303065

[132] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 12 (2009), 1743–1759. doi:10.1016/j.imavis.2008.11.007 Visual and multimodal analysis of human spontaneous behaviour:.

[133] Tuan-Hung Vu, Anton Osokin, and Ivan Laptev. 2015. Context-Aware CNNs for Person Head Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2893–2901.

[134] Sandra Wachter. 2024. Limitations and loopholes in the EU AI Act and AI Liability Directives: what this means for the European Union, the United States, and beyond. *Yale Journal of Law and Technology* 26, 3 (2024), 671–718.

[135] Junqi Wang, Chunhui Zhang, Jiapeng Li, Yuxi Ma, Lixing Niu, Jiaheng Han, Yujia Peng, Yixin Zhu, and Lifeng Fan. 2024. Evaluating and Modeling Social Intelligence: A Comparative Study of Human and AI Capabilities. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 46. https://escholarship.org/uc/item/2j53v5nv

[136] Xukang Wang, Ying Cheng Wu, Mengjie Zhou, and Hongpeng Fu. 2024. Beyond surveillance: privacy, ethics, and regulations in face recognition technology. *Frontiers in Big Data* Volume 7 - 2024 (2024). doi:10.3389/fdata.2024.1337465

[137] Zeyu Wang, Yuanchun Shi, Yuntao Wang, Yuchen Yao, Kun Yan, Yuhan Wang, Lei Ji, Xuhai Xu, and Chun Yu. 2024. G-VOILA: Gaze-Facilitated Information Querying in Daily Scenarios. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 78 (May 2024), 33 pages. doi:10.1145/3659623

[138] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. arXiv:2112.04359 [cs.CL] https://arxiv.org/abs/2112.04359

[139] Glenn E. Weisfeld and Jody M. Beresford. 1982. Erectness of posture as an indicator of dominance or success in humans. *Motivation and Emotion* 6, 2 (1982), 113–131. doi:10.1007/BF00992459

[140] Ethan Wilson, Naveen Sendhilnathan, Charlie S. Burlingham, Yusuf Mansour, Robert Cavin, Sai Deep Tetali, Ajoy Savio Fernandes, and Michael J. Proulx. 2025. Eye Gaze as a Signal for Conveying User Attention in Contextual AI Systems. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications (ETRA '25)*. Association for Computing Machinery, New York, NY, USA, Article 119, 7 pages. doi:10.1145/3715669.3727349

[141] Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. 2025. ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 207–229.

[142] Claire Woodcock, Brent Mittelstadt, Dan Busbridge, and Grant Blank. 2021. The Impact of Explanations on Layperson Trust in Artificial Intelligence–Driven Symptom Checker Apps: Experimental Study. *J Med Internet Res* 23, 11 (3 Nov 2021), e29386. doi:10.2196/29386

[143] Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. 2025. Mitigating Hallucinations in Multimodal Large Language Models via Image Token Attention-Guided Decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 1571–1590. doi:10.18653/v1/2025.naacl-long.75

[144] Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. 2024. Voila-A: Aligning Vision-Language Models with User's Gaze Attention. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 1890–1918. doi:10.52202/079017-0060

[145] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10632–10643.

[146] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing* 4, 2 (2024), 100211. doi:10.1016/j.hcc.2024.100211

[147] Zhoutong Ye, Mingze Sun, Huan ang Gao, Xutong Wang, Xiangyang Wang, Yu Mei, Chang Liu, Qinwei Li, Chengwen Zhang, Qinghuan Lan, Chun Yu, and Yuanchun Shi. 2025. MOAT: Evaluating LMMs for Capability Integration and Instruction Grounding. arXiv:2503.09348 [cs.CL] https://arxiv.org/abs/2503.09348

[148] Kate Yen, Yeqi Chen, Yi Cheng, Sijin Chen, Ying-Yu Chen, Yiran Ni, and Alexis Hiniker. 2018. Joint Media Engagement between Parents and Preschoolers in the U.S., China, and Taiwan. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 192 (Nov. 2018), 19 pages. doi:10.1145/3274461

[149] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (11 2024), nwae403. doi:10.1093/nsr/nwae403

[150] Junnan Yu, Xiang Qi, and Siqi Yang. 2024. Parent-Child Joint Media Engagement Within HCI: A Scoping Analysis of the Research Landscape. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 121, 21 pages. doi:10.1145/3613904.3642307

[151] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. MM-Vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2381, 25 pages.

[152] Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. 2017. Supervising Neural Attention Models for Video Captioning by Human Gaze Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 490–498.

[153] Sojeong Yun and Youn-kyung Lim. 2025. What If Smart Homes Could See Our Homes?: Exploring DIY Smart Home Building Experiences with VLM-Based Camera Sensors. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 945, 22 pages. doi:10.1145/3706598.3713265

[154] Xin Zeng, Xiaoyu Wang, Tengxiang Zhang, Chun Yu, Shengdong Zhao, and Yiqiang Chen. 2024. GestureGPT: Toward Zero-Shot Free-Form Hand Gesture Understanding with Large Language Model Agents. *Proc. ACM Hum.-Comput. Interact.* 8, ISS, Article 545 (Oct. 2024), 38 pages. doi:10.1145/3698145

[155] John Zerilli, Umang Bhatt, and Adrian Weller. 2022. How transparency modulates trust in artificial intelligence. *Patterns* 3, 4 (April 2022). doi:10.1016/j.patter.2022.100455

[156] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. 2021. Human gaze assisted artificial intelligence: a review. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Yokohama, Japan) *(IJCAI'20)*. Article 689, 8 pages.

[157] Xu Zhang, Kailun Yang, Jiacheng Lin, Jin Yuan, Zhiyong Li, and Shutao Li. 2024. PVPUFormer: Probabilistic Visual Prompt Unified Transformer for Interactive Image Segmentation. *IEEE Transactions on Image Processing* 33 (2024), 6455–6468. doi:10.1109/TIP.2024.3492713

[158] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Computational Linguistics* 51, 4 (12 2025), 1373–1418. doi:10.1162/COLI.a.16

[159] Yanxia Zhang, Ken Pfeuffer, Ming Ki Chong, Jason Alexander, Andreas Bulling, and Hans Gellersen. 2017. Look together: using gaze for assisting co-located collaborative search. *Personal Ubiquitous Comput.* 21, 1 (Feb. 2017), 173–186. doi:10.1007/s00779-016-0969-x

[160] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024. Why are Visually-Grounded Language Models Bad at Image Classification?. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 51727–51753. doi:10.52202/079017-1639

[161] Zory Zhang, Pinyuan Feng, Bingyang Wang, Tianwei Zhao, Suyang Yu, Qingying Gao, Hokin Deng, Ziqiao Ma, Yijiang Li, and Dezhi Luo. 2025. Can Vision Language Models Infer Human Gaze Direction? A Controlled Study. arXiv:2506.05412 [cs.CV] https://arxiv.org/abs/2506.05412

[162] Zheng Zhang, Weirui Peng, Xinyue Chen, Luke Cao, and Toby Jia-Jun Li. 2025. LADICA: A Large Shared Display Interface for Generative AI Cognitive Assistance in Co-located Team Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 147, 22 pages. doi:10.1145/3706598.3713289

[163] Henry Hengyuan Zhao, Pan Zhou, and Mike Zheng Shou. 2025. GENIXER: Empowering Multimodal Large Language Model as a Powerful Data Generator. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 129–147.

[164] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep Learning-based Human Pose Estimation: A Survey. *ACM Comput. Surv.* 56, 1, Article 11 (Aug. 2023), 37 pages. doi:10.1145/3603618

[165] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. 2022. RegionCLIP: Region-Based Language-Image Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16793–16803.

[166] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large

Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=1tZbq88f27

A Training GazeLLE-v3

The original GazeLLE model is trained for 15 epochs on the GazeFollow dataset with the Adam optimizer. The initial learning rate (LR) is set at $\eta = 10^{-3}$, with cosine annealing LR scheduling [80] over the 15 epochs ending with $\eta_{min} = 10^{-7}$. Three data augmentation techniques were used to train GazeLLE, namely random cropping, horizontal flipping, and bounding box jittering. For GazeLLE-v3-L, we used the exact same training procedure. For GazeLLE-v3-H, we trained the model for 20 instead of 15 epochs to take advantage of the higher quality dense features extracted by DINOv3. A modified version of the publicly released code of GazeLLE[4] is used to train GazeLLE-v3. In total, GazeLLE-v3-L took around 13 RTX 5090 GPU hours to train, while GazeLLE-v3-H took around 32 RTX 5090 GPU hours. We used the checkpoint from the last epoch in GazeCoT.

B GazeLLE-v3 Failures

Figure 18: Failures by GazeLLE-v3-H. A green bounding box indicates the person whose gaze is being estimated, and a green dot indicates the predicted fixation point (the point with the largest heatmap value). The three bad cases on the left are caused by insufficient understanding of head and eye orientation, while the case on the right is caused by both poor fine-grained feature understanding and the inability to consider the action's (drinking from a cup) impact on gaze by blocking the line of sight. The leftmost image is from Zhang et al. [161] while the other three are from GazeFollow [103].

While GazeLLE-v3 represents a significant improvement over the original GazeLLE model, it still has weaknesses in understanding fine-grained head and eye features, as well as the influence of human activity on gaze. We present some of the bad cases in Figure 18. We expect these weaknesses to be addressed by progress in general purpose vision encoders and gaze estimation algorithms. Addressing these failures would improve the performance of GazeCoT and benefit HCI applications in general.

C From Heatmap to ROI

We provide the detailed ROI extraction procedure used in the ROI description tool. Given a gaze heatmap $\mathbf{H}$, the ROI is determined through a multi-stage process. First, the heatmap $\mathbf{H} \in \mathbb{R}^{64\times64}$ is binarized into a mask $\mathbf{M} \in \{0,1\}^{64\times64}$, using a threshold θ to isolate the region around the fixation point:

$$\mathbf{M}(i,j) = \begin{cases} 1 & \text{if } \mathbf{H}(i,j) \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

Second, the minimal bounding box $\mathbf{B} = (x_{\min}, y_{\min}, w, h)$ for this region is extracted from $\mathbf{M}$, defined by its top-left corner $(x_{\min}, y_{\min})$ and its dimensions (w, h). Finally, to incorporate local context, this box is uniformly expanded by a factor of n around its center to yield the final ROI, $\mathbf{B}' = (x'_{\min}, y'_{\min}, w', h')$:

$$w' = n \cdot w, \quad h' = n \cdot h,$$
$$c_x = x_{\min} + \frac{w}{2}, \quad c_y = y_{\min} + \frac{h}{2},$$
$$x'_{\min} = c_x - \frac{w'}{2}, \quad y'_{\min} = c_y - \frac{h'}{2}$$

We empirically set the binarization threshold at $\theta = 0.15$ and the expansion factor at $n = 2.0$. If the need arises, θ and n can be adjusted to fit specific downstream applications of GazeCoT. Decreasing θ or increasing n would lead to larger ROIs, while increasing θ or decreasing n would lead to smaller ROIs.

D GazeCoT System Prompts

We present the system prompts used in GazeCoT. These are also available in the supplementary materials. We will release a public repository for GazeCoT once the paper is accepted for publication.

Task Agent Prompt (For VQA)

```
{
    "task": "You are given images overlaid with the focus
    point of the gaze of person(s) of interest and a
    question. Answer the question to the best of your
    ability.",
    "requirements": [
        "There are n images in total, each paired with a
        detailed text description. Each image-text pair
        corresponds to one person of interest: the image
        is annotated with a colored bounding box around
        the person of interest's head and a dot of the
        same color marking the fixation point of their
        gaze. The text is a detailed description of the
        area on and around the fixation point.",
        "For each image pair, first analyze the annotated
        original image. Then analyze the text
        description and its implications for the question.
        Finally, answer the question.",
        "If you are given a list of options, you must
        choose from it and repeat the choice verbatim.
        You must not give an answer not from the list
        of options.",
        "You must answer in the following json format:
        {
            "analysis": "(write your analysis here)",
            "answer": "(your answer)
        }"
    ]
}
```

Task Agent Prompt (For Parent-Child JME Analysis)

```
Describe the interaction between parent and child in
the video using the frames given to you. Analyze their
engagement with the activity. Write your analysis in
```

[4] https://github.com/fkryan/gazelle

```
segments with clear timestamps. The FPS is {FPS}. The
first frame is at {START_TIMESTAMP} seconds.
```

ROI Description Prompt

```
this is a zoomed-in image with a small red circle annotating
the focus point. Find where the focus point is
(Important!) and very briefly describe it and
surrounding objects in a single short sentence.
Do not describe the background.
```

E Play Sequence Collection Procedure

The play segments mentioned in Section 6 were gathered from a user study investigating an AI assistant for parent-child play-based English as a foreign language (EFL) learning. The experiment was designed to evaluate the system's impact on learning outcomes and triadic (parent-child-AI) interaction patterns.

In that study, 16 parent-child dyads (Parents: 14F 2M, Children: 11F 5M) were recruited online. The children were between 3 and 6 years old (4.6±1.1). There were no specific language proficiency requirements for the parents and children, as that study intended to capture the usage patterns of parents and children with various language proficiency levels. The experimenter informed the parent and child about the terms for data storage and usage, privacy, and potential risks. All participants signed an informed consent form. **The usage of the recordings in this paper (GazeCoT) strictly follows the terms agreed to by participants.**

The study took place in a lab environment. Following an informed consent process that covered data usage, privacy, and risks, each dyad completed two 20-minute play sessions. The main task for the parent was to integrate language learning into their play while maintaining the child's engagement. The room where the experiments took place has a table with toy sets. These include one set of kitchen toys and one set of clinic toys, used for the two play sessions, respectively. The toy sets were chosen to contain potential vocabulary items for the parents to teach. Apart from the toys, the parent was provided with a laptop, tablet, or phone with an AI assistant application. The parent can use the AI to provide either language support (e.g. asking about vocabulary and expressions) or game progression support (asking for advice on what to do next). The experimenter exited the room after the preparation to ensure natural parent–child interaction.

F Field Note Example

Below is the full version of the GazeCoT-generated field note shown as an example in Figure 12.

0.00–19.00s

Parent: The parent is primarily focused on the laptop and its cables, with hands on the keyboard and gaze directed toward the device. There are brief moments when the parent glances at play items (toy ice cream cone, play food), but overall, their attention is dominated by the laptop, likely consulting the AI assistant for information.

Child: The child's gaze is mostly on their own hand, forearm, and play materials (pink marker, green toy figure, containers, toy frying pan). The child is engaged in solitary play, not interacting with the parent or participating in a joint activity.

Engagement: Throughout this period, there is minimal evidence of joint attention or meaningful parent-child engagement. The parent's over-reliance on the laptop results in limited support for the child's play, and the child remains independently focused on their own activities.

19.00–22.00s

Parent: The parent becomes more actively involved, reaching toward the play area and objects (glass jar, toy vegetables) while still consulting the laptop for language support.

Child: The child is engaged with the play food, holding a pink toy knife and interacting with the items on the table.

Engagement: Both parent and child are focused on the play objects and each other, with gaze fixations on the shared activity space. Joint attention and collaborative play are evident.

22.00–28.00s

Parent: The parent continues to interact with the child, focusing on the child's hand and the pot/pan lid, supporting the play scenario. Gaze alternates between the child's actions and the laptop.

Child: The child manipulates the pot lid and other toy food items, maintaining active participation in the play.

Engagement: The parent uses the AI assistant as a support tool but not as a distraction. The child remains engaged with the play process, and joint attention is maintained.

29.00–39.00s

Parent: The parent's attention shifts more toward the laptop screen and cables, with gaze fixations on the device for extended periods. Direct engagement with the child decreases, and the parent's support for the play process is reduced.

Child: The child continues to interact with the play food and utensils, but the activity becomes less collaborative as the parent's attention wanes.

Engagement: The parent's excessive attention to the laptop detracts from the quality of the learn-by-play process. The child remains engaged with the toys, but joint attention and shared play diminish.

40.00–42.00s

Parent: The parent is seated beside the child, with one hand on the laptop and gaze directed toward the laptop's cable/connector and screen edge. The parent may be consulting the AI assistant but is partially engaged with the child.

Child: The child is actively engaged with the play food set, holding a toy spatula and interacting with toy vegetables and containers.

Engagement: The child is focused on the play activity, responding to the parent's prompts. The parent's attention is split but not to the exclusion of the child. Mild flag for parent attention, but joint attention is present.

43.00–59.00s

Parent: The parent's gaze shifts from the laptop to the child's hands and play food items. The parent's hand is often on the laptop, but their gaze is increasingly on the child's actions. The parent guides and responds to the child's play choices, with gestures indicating active participation.

Child: The child is highly engaged, using utensils, holding vegetables, and responding to the parent's comments. The child's gaze and hand movements show active participation in the play scenario.

Engagement: The interaction is collaborative, with both parties focused on the play food set and utensils. Joint attention and shared activity are strong, and the parent is not overly distracted by the laptop. The play-by-learning process is maintained, with both parent and child focused on the activity.

—

Overall Assessment

Parent's Engagement: The parent's attention fluctuates throughout the session. In the first segment (0.00–19.00s), the parent is overly focused on the laptop, resulting in minimal engagement with the child and limited joint attention. From 19.00–28.00s, the parent balances attention between the child and the AI assistant, supporting collaborative play. However, from 29.00–39.00s, the parent's attention again shifts excessively to the laptop, reducing direct support for the child's play. In the final segment (40.00–59.00s), the parent's engagement improves, with gaze and gestures directed toward the child and the play activity, fostering joint attention and collaborative learning.

Child's Engagement: The child is consistently engaged with the play materials, but the quality of engagement depends on the parent's involvement. When the parent is distracted by the laptop, the child plays independently. When the parent is actively involved, the child participates in shared play and responds to prompts, demonstrating joint attention.

Gaze, Attention, and Joint Attention: Joint attention is limited in the early and middle segments due to the parent's excessive focus on the laptop. It improves significantly in the later segment (40.00–59.00s), with both parent and child attending to the shared activity space and each other. Flagged Issues: The main concern is the parent's excessive attention to the AI assistant/laptop during 0.00–19.00s and 29.00–39.00s, which detracts from the quality of the learn-by-play process and limits joint attention. The child's engagement is generally strong, but collaborative play is reduced when the parent is distracted.

Recommendation: For optimal learn-by-play outcomes, the parent should minimize laptop use during the activity and maintain direct engagement with the child, using the AI assistant only as a brief reference tool. Sustained joint attention and collaborative play are essential for effective language learning and parent-child interaction.

G User Study Questionnaire

We present the questionnaire used by the participants to rate the field notes in the user study.

Questionnaire

Rate the field note on the following 6 aspects on a scale of 1 to 7. 1 for *strongly disagree*, 7 for *strongly agree*.

(1) **Accuracy.** The observations in the field note are accurate and consistent with the video clip.
(2) **Comprehensiveness.** The field note is comprehensive and covers all useful details from the video clip.
(3) **Usefulness.** The field note is useful for your work or research. It offers insight and / or inspires your work or research.
(4) **Explainability.** The conclusions of the field note are explainable and grounded in specific scenes and events from the video clip.
(5) **Trustworthiness.** I trust the contents of this field note, and would trust other field notes generated by the same system.
(6) **Overall Preference.** I am overall satisfied with the field note.

H Interview Script

In the post-experiment interview, we asked the participants the following questions:

(1) What is your impression on the two types of field notes in general? Have you observed any general trends in terms of the difference between the two?
(2) What is your own method of analyzing the clips? How does each of the two systems align with it?
(3) Do you think directing the AI to focus on gaze interaction improves field note quality for the task of parent-child JME analysis?
(4) Do you have any suggestions for further improving the AI's ability to understand complex parent-child interactions?

Note that the questions were asked *after* we revealed to the participant which notes were generated by System A (baseline) and System B (GazeCoT). Apart from these questions, we also asked the participant to explain the reason behind their ratings. Finally, we discussed specific examples that left the participants with deep impressions.

I GazeLLE-v3-H vs GazeLLE-v3-L

We trained two variants of GazeLLE-v3, GazeLLE-v3-L and GazeLLE-v3-H. The former uses DINOv3-ViT-L (300M parameters) as the vision backbone, while the latter uses DINOv3-ViT-H+ (840M parameters). In addition to the quantitative metric comparison in Table 1, we also provide a qualitative comparison in Figure 19. GazeLLE-v3-L performed between GazeLLE-L and GazeLLE-v3-H. Therefore, we chose GazeLLE-v3-H as the gaze estimator for GazeCoT due to its superior performance. The increase in model size (840M vs 300M) is not a particular concern in our study, since both variants of GazeLLE-v3 can process dozens of frames per second on an RTX 5090 GPU. If the need arises, GazeLLE-v3-L can be used instead of GazeLLE-v3-H in resource-constrained environments, although a drop in output quality should be expected.

Figure 19: Comparison of GazeLLE-L (released by Ryan et al.) and our GazeLLE-v3-L and GazeLLE-v3-H models.