GIICS: A dataset for gaze-direction-based intention inference in complex visual scenes

Qianxi Yu
Hangzhou Collaborative Innovation Institute of Language Services
Hangzhou City University
Hangzhou, China
yqx912@gmail.com

Chen Ling
College of Biomedical Engineering and Instrument Sciences
Zhejiang University
Hangzhou, China
chen_ling@zju.edu.cn

Shitao Feng
School of Foreign Languages
Hangzhou City University
Hangzhou, China
32208074@stu.hzcu.edu.cn

Jiayi Hu
School of Foreign Languages
Hangzhou City University
Hangzhou, China
32208188@stu.hzcu.edu.cn

Yongtong Lu
School of Foreign Languages
Hangzhou City University
Hangzhou, China
32208090@stu.hzcu.edu.cn

Nai Ding*
College of Biomedical Engineering and Instrument Sciences
Zhejiang University
Hangzhou, China
ding_nai@zju.edu.cn

Abstract*—To precisely recognize a visual object embedded in a complex visual scene, a person often has to adjust his or her gaze to fixate on the object, i.e., placing the object into the center of the visual field. Therefore, correctly understanding gaze direction and, more importantly, which object is being fixated on, is necessary to understand the intention of a person. To test how well vision large language models (VLLMs) can understand gaze direction and which object is being fixated on in a complex visual scene, the current study presents a novel dataset, referred to as the Gaze-based Intention Inference in Complex Scenes (GIICS), that comprises over 2800 pictures taken in 6 indoor and 14 outdoor backgrounds. Each picture contains two visual objects embedded in the visual background, with a viewer fixating on one of them. To verify if VLLMs can directly understand gaze direction instead of inferring gaze direction based on body posture or head direction, we constructed the GIICS-BodyPosture subset of images featuring uncommon combinations of body and head directions. Moreover, we constructed the GIICS-ObjectAngle subset, where we manipulated the visual angles between competing objects (e.g., objects placed to the left or right of the viewer) to investigate how angular intervals affect gaze target detection. We paired each picture with a visual question answering (VQA) question (e.g., "what is the person looking at? (A) the yellow object; (B) the blue object; (C) neither.") and evaluated how well nine popular VLLMs, e.g., Llama and Qwen2.5-VL, could answer the questions in both Chinese and English. The median VQA accuracy across models was 37.1% (VQA in Chinese) and 40.6% (VQA in English) and the highest performance was 57.1% (VQA in Chinese) and 58.8% (VQA in English). These results reveal significant deficits of current VLLMs when inferring human intention based on gaze direction in complex visual scenes.

***Index Terms*—Gaze-based intention inference, VQA dataset, Vision large language models.**

*Corresponding author.

I. Introduction

Recent breakthroughs in vision large language models (VLLMs) have significantly enhanced the performance of multimodal LLMs in reasoning about a visual scene [27], [32]. The Visual Question Answering (VQA) paradigm, which links natural language processing with computer vision, is a powerful tool for evaluating these models through asking textual questions about images [3]. While numerous VQA datasets have been established, most of them focused on static scene understanding tasks, such as object detection [20], [26] or attribute description [1], [8], lacking explorations into reasoning about human behaviors which convey intentions, such as gaze.

Gaze is a powerful social cue for interpreting humans' intentions [9]. Gaze understanding has broad applications in areas including human-computer interaction [23], automatic driving [16], and attention analysis [25], [28]. While numerous datasets have supported the estimation of precise 2D/3D coordinates and angular degrees of gaze [17], [29], [33], they rarely addressed the semantic interpretation of gaze targets, such as identifying what the person in the picture looks at.

Semantic interpretation of gaze requires reasoning beyond coordinates, considering contextual elements, such as complex postures, presence of target and competitive objects and real-world backgrounds. Yet, most existing gaze-related datasets excluded gazed-at objects from the images, not to mention introducing competing objects [17], [29], [33]. Furthermore, most existing datasets typically focused on head or face regions, overlooking the influence of body posture. Even in a few datasets of full-body images, body-head alignment was typically assumed to be consistent [17]. However, the misalignment across body and head directions is common in

TABLE I
COMPARISON BETWEEN POPULAR GAZE DATASETS AND GIICS

Dataset	Backgrounds	Body In Images	Body–head Directional Inconsistency	Objects Presence In Images
MPIIGaze	Indoor	No	No	No
GazeCapture	Indoor; Outdoor	No	No	No
Columbia Gaze	Indoor	No	No	No
Gaze360	Indoor; Outdoor	Yes	No	No
ETH-XGaze	Indoor	No	No	No
GIICS	**Indoor; Outdoor**	**Yes**	**Yes**	**Yes**

everyday social interactions. For example, people could tilt their heads toward someone next to them with their bodies still facing the front, and their gazes may not be directed at the head-facing person. While humans could resolve such ambiguity, it remained unclear whether VLLMs could identify the same reliable cues for gaze perception as humans.

Therefore, the current study presented a specialized dataset of gaze images and an accompanying VQA task to evaluate VLLMs on semantic inference of gaze targets under real-world distractions, including complex body-head direction combinations and interference from competing objects. In this work, we first demonstrated the procedure of dataset construction, and then reported the performance of numerous popular VLLMs on our dataset. Our main contributions are as follows: (1) Gaze is an important social cue, but no existing dataset links gaze direction to the exact object being gazed at. We presented a novel dataset **GIICS** and benchmark task of semantic gaze interpretation, extending traditional gaze direction estimation problems to gaze-direction-based intention inference. (2) The dataset included 20 diverse background settings, competing objects embedded in the visual background, and complex combinations of body-head directions for the viewer, providing a test evaluating robustness of VLLMs in reasoning about gaze under ambiguity. Unlike previous gaze direction datasets, which typically show only the face area, our images cover both the face and upper body in multi-object natural scenes. Furthermore, our dataset is the first to include body-head mismatches to assess whether models can directly infer the gaze target without relying on posture cues. (3) We benchmarked numerous VLLMs (e.g., Llama 3 [6], LLaVA-1.5 [21], Gemma-3 [12], Gemini 2.0 Flash-Lite [13] and Qwen2.5-VL series [4] on our dataset, analyzing their accuracy and robustness across different gaze scenarios.

II. RELATED WORK

A. Multimodal LLMs and Visual Question Answering (VQA) Task

Unlike traditional LLMs which excel in processing textual input alone, multimodal LLMs can cope with inputs of various modalities, including text, speech and images [27], [30]. To reason about visual content, these models typically consist of at least three components: a visual encoder, a vision-to-language adapter and an LLM backbone [6], [30]. The visual encoder, often a frozen pre-trained visual model, is responsible for extracting image features. The adapter then aligns visual features with textual embeddings, enabling cross-modal interaction [2], [19]. With these vision-aligned textual embeddings, the LLM backbone further reasons about visual input and generate outputs based on textual prompts.

Considering the strong text reasoning capabilities of LLM backbones, the Visual Question Answering (VQA) paradigm has emerged as a key tool in evaluating performance of visual reasoning by multimodal LLMs [3]. The VQA task requires the model to process both the question in natural language and the visual input before generating an answer for the question according to the visual information of the inputted images. The questions in VQA tasks can either be close-ended (e.g., multiple choice) or open-ended (e.g., flexible text or caption generation) and are typically categorized into two types: (1) questions based on observed ground truth of images, including the shape, color and number of objects, (2) questions which require external knowledge, such as commonsense understanding [24], [31] or spatial relationships [22].

The growing interests in VQA have led to numerous benchmark datasets, with the most famous one being the VQA dataset itself, which collected 204721 images with both open-ended and multiple-choice questions for ground-truth answers [3]. More recent research has presented VQA datasets of human behaviors and intentions, such as the MotionLLM dataset for human motion reasoning [7] and the Image-Emotion-Social-Net dataset for emotion-related questions reasoning [35]. However, one crucial aspect of intention-conveying human behaviors remained largely overlooked: human gaze.

B. Gaze-related Datasets

Gaze is a crucial non-verbal cue for inferring humans' intentions and desires [9]. Accurate detection and understanding of human gaze could enable machines to better interact with humans and enhance their performance in areas such as automatic driving and smart-home robots [23]. As a result, gaze estimation has been a crucial topic in computer vision, yielding numerous gaze-related datasets [11], [17], [18], [29], [34].

Early gaze datasets adopted screen-based collection methods, capturing gaze data through built-in cameras on laptops or smartphones, with participants instructed to look at on-screen targets. For instance, the MPIIGaze dataset collected 213659 pictures of 15 participants through their laptop cameras [34] while GazeCapture recorded 2445504 frames of 1474 subjects, using mobile devices such as iPads or iPhones [18]. While these datasets achieved impressive scale contributed by the

Fig. 1. Equipment setup across 20 backgrounds. Faces in the image illustration are masked for privacy

efficiency of private equipment to reach a broad pool of participants, they typically suffered from limited variability of gaze range, head pose and environment contexts due to their fixed setup of camera and screen-based interaction.

To address these constraints, subsequent datasets resorted to external cameras to capture images, allowing for more variability in gaze range and head poses. The Columbia Gaze dataset, for example, collected 5880 gaze images of 56 participants, covering 5 head poses and 21 gaze directions [29]. However, its use of a chinrest to fix participants' head and a black laboratory background, limited its generalization beyond laboratory settings. Gaze360 further advanced the field by capturing 197,588 images of 238 subjects from both indoor and outdoor settings, providing both full-body and cropped face images, though participants mostly faced forward with aligned body-head directions [17]. The ETH-Xgaze further enhanced variations in head poses and gaze ranges, capturing over 1 million images of 110 participants [33]. However, this dataset was confined to a laboratory background and ignored the body part of photographed subjects, limiting its

representation of natural postures.

While the existing gaze datasets have contributed to the gaze research, they share several limitations that hinder their suitability for semantic intention inference. First, they primarily focused on gaze estimation tasks, predicting 2D coordinates (e.g., GazeCapture) or 3D directions (e.g., Gaze360, ETH-XGaze), without linking gaze to semantic targets within the context. Second, their focus on face or head regions and their assumption of body part alignment overlooked the influence of inconsistent body-head orientation combinations, which are common in real-world settings. Furthermore, prior datasets rarely considered the influence of competitive objects, particularly when the two objects are placed at close angular intervals, a factor that can introduce ambiguity in gaze target detection. Table I summarized key characteristics of these datasets, including background settings, inclusion of body in images, directional inconsistencies across body parts and the presence of gazed-at and competitive objects. These comparisons highlighted the need for a novel dataset which captures naturalistic body-head postures and gaze-object interactions in real-world environments, to support semantic reasoning of gaze target.

III. Method

This dataset provides images of human gaze in complex real-world backgrounds and comprises two subsets. The GIICS-BodyPosture subset contains images of both consistent and inconsistent body-head directions, while the GIICS-ObjectAngle subset further features varied angles between the gazed-at target and a competitive object. Using these images, we constructed a VQA task to assess VLLMs' ability to infer gaze targets despite multiple distractions. Specifically, the question format was unified as a forced-choice multiple-choice problem:"What is the person looking at?" Three options were always provided: two corresponding to the objects present in the scene and one "neither" option. The correct answer was determined by the instructed gaze direction of the subject, when the subject was instructed to look at the camera or at an area not aligned with any object, option "neither" was designated as the correct answer.

A. Data Collection Setup

The general setup for collecting the images in both subsets is illustrated in Fig. 1 To capture images, a camera secured on the tripod was used, with its height always aligned to the eyes of the photographed subject while its distance from subjects varying from 3 to 5 meters to ensure that subjects were centered on the images. For each image, a square aspect ratio was chosen to maximize the subject's prominence in the frame, resulting in images with a resolution of 3024*3024 pixels. A total of 20 campus contexts were included as the backgrounds of images, with 14 outdoor and 6 indoor settings.

Four adult subjects (three females, one male), aged from 21 to 24, were recruited as models for image collection. All subjects had normal vision, no color blindness, and could direct their gaze as instructed by the experimenter. None of the subjects wore glasses. Subjects were recruited on a voluntary basis and all of them signed the written informed consent, allowing their photographs to be included in the dataset. The dataset, with facial areas (excluding the eyes) masked, will be available upon request for non-commercial purposes. While all images used for illustration in this paper are masked to ensure privacy, the images used for model testing are not masked.

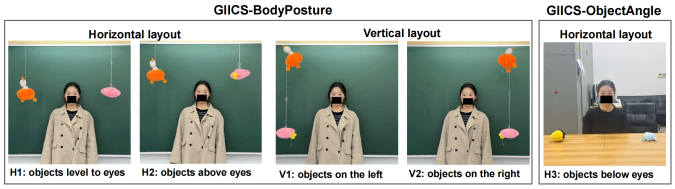

Fig. 2. Object layouts for GIICS-BodyPosture and GIICS-ObjectAngle subsets. Faces in the image illustration are masked for privacy.

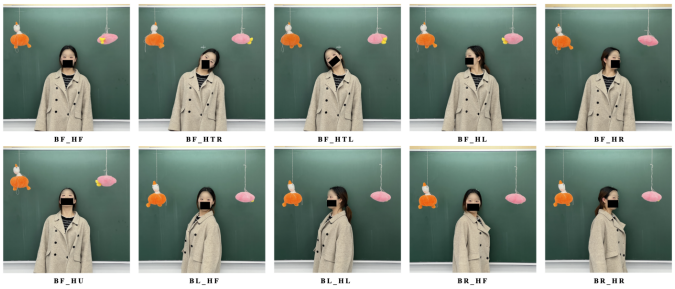

Fig. 3. Illustration of 10 body-head direction combinations for GIICS-BodyPosture subset. Faces in the image illustration are masked for privacy.

B. Object Position Layouts

To guide human gaze, two pairs of colored objects (yellow vs. blue and orange vs. pink) were employed, with one serving the target and the other serving as a potential competitor. To minimize position-related biases, we included both horizontal and vertical layouts of objects. In the horizontal layout, the two objects were horizontally positioned symmetrically around the subject at varying heights: at eye level (H1), above eye level (H2) and below eye level (H3). In the vertical layout, the two objects were placed along a vertical line, either to the left (V1) or to the right of the image (V2), with their middle point aligned with the eye level of subjects. These layouts correspond to distinct subsets of GIICS, each involving different settings, as summarized in Table II. The visual examples of each object layout are illustrated in Fig. 2.

For clarity, all directional descriptions in this dataset were defined relative to the image frames rather than the subject's perspective. For instance, the direction of "left" refers to the left frame of images. Moreover, the term "forefront" refers to the direction toward the camera, rather than the front of the subject's body.

C. GIICS-BodyPosture Subset

The GIICS-BodyPosture subset investigates the influence of body-head direction combinations on gaze detection. Considering both human mobility constraints and the visibility of

TABLE II
OVERVIEW OF TWO SUBSETS OF GIICS DATASET

Subset	Feature	Object Position Layout	Settings
GIICS-BodyPosture	Various combinations of body and head directions	Horizontal: level to eyes Horizontal: above eyes Vertical: on the left side of subject Vertical: on the right side of subject	14 outdoor backgrounds; 2 indoor backgrounds
GIICS-ObjectAngle	Graded changes in the angular interval between objects	Horizontal: below eyes	4 indoor backgrounds

Fig. 4. Illustration of 4 angular intervals between objects for GIICS-ObjectAngle subset. Faces in the image illustration are masked for privacy.

gaze limited by postures, we designed three body directions: (1) Body to Forefront (BF), where the person's body is facing the camera; (2) Body to Left (BL), where the person's body turns approximately 30 degrees to the left of the image; (3) Body to Right (BR), where the person's body turns approximately 30 degrees to the right of the image. For head directions, we designed six categories: (1) Head to Forefront (HF), head facing the camera; (2) Head Upward (HU), head facing forward, tilted approximately 30 degrees upward; (3) Head Tilted Left (HTL), head tilted to the left of the picture about 30 degrees; (4) Head Tilted Right (HTR), head tilted to the right of the picture about 30 degree; (5) Head Turned Left (HL), head rotated about 30 degrees to the left of the picture; (6) Head Turned Right (HR), head rotated about 30 degrees to the right of the picture.

To evaluate whether models could accurately infer gaze targets despite such directional inconsistencies, we combined body and head directions, creating nine complex posture combinations, as shown in Fig. 3 For this subset, we designed five types of gaze directions for both horizontal and vertical layouts of objects, with some directed at objects and others not. These included: (1)Gaze at Object 1 (e.g., gaze at the orange object (GO-O)); (2) Gaze at Object 2 (e.g., gaze at the pink object (GO-P)); (3) Gaze to Forefront (GF), with eyes fixating at the camera; (4) Gaze Upward (GU), with eyes looking upward above the head; (5) Gaze Downward (GD), with eyes looking down to the ground. Notably, for the vertical layout of objects conditions, we included two additional gaze directions to introduced ambiguity: gaze to the left of the midpoint line between the two objects (GLM) and gaze to the right of the midpoint (GRM).

TABLE III
MODELS' OVERALL ACCURACY ON THE GIICS DATASET

Model	Prompt language	Accuracy
Gemma-3-4B-IT	Chinese	0.24
	English	0.19
Gemma-3-12B-IT	Chinese	0.47
	English	0.43
Gemma-3-27B-IT	Chinese	0.48
	English	0.58
Llama 3.2 Vision	Chinese	0.21
	English	0.22
Llama 4 Scout	Chinese	0.23
	English	0.30
LLaVA-1.5	Chinese	0.30
	English	0.41
Qwen2.5-VL-7B-Instruct	Chinese	0.37
	English	0.41
Qwen2.5-VL-72B-Instruct	Chinese	0.57
	English	0.54
Gemini 2.0 Flash-Lite	Chinese	0.56
	English	0.59

D. GIICS-ObjectAngle Subset

The GIIC-ObjectAngle subset further explores how the angle between a target object and a competitive object influence gaze detection. In this setup, the subject was seated at the table where the two objects were placed. The subject was instructed to fixate on the target object while the competing object moved, with the angular intervals between them varying across 120°, 90°, 60°, and 30° (see Figure 4 for examples). The maximum angle was set to 120° due to the human ocular motor range of approximately ±55° [15], along with the expanded range of view contributed by body and head movement. The 30° angle interval was chosen as it was the smallest angular interval at which eye direction shifts across objects remained visible in the images.

Only two subjects participated in the image collection for this subset, and their bodies were always directed to forefront, resulting in only six body-head direction combinations in this subset. This was done to avoid changes of the perceived angle between the gaze target and distractor caused by turning the body.

TABLE IV
MODELS' ACCURACY ACROSS OBJECT LAYOUTS IN THE GIICS-BODYPOSTURE SUBSET

Model	Prompt language	Object spatial layouts				
		H1	H2	V1	V2	Overall
Gemma-3-4B-IT	Chinese	0.25	0.28	0.17	0.21	0.22
	English	0.17	0.21	0.13	0.15	0.16
Gemma-3-12B-IT	Chinese	0.53	0.57	0.47	0.42	0.49
	English	0.47	0.55	0.41	0.37	0.44
Gemma-3-27B-IT	Chinese	0.57	0.55	0.50	0.44	0.51
	English	0.64	0.61	0.67	0.58	0.62
Llama 3.2 Vision	Chinese	0.21	0.22	0.18	0.21	0.21
	English	0.23	0.25	0.19	0.22	0.22
Llama 4 Scout	Chinese	0.26	0.28	0.19	0.20	0.23
	English	0.33	0.35	0.25	0.28	0.30
LLaVA-1.5	Chinese	0.32	0.34	0.20	0.34	0.30
	English	0.45	0.44	0.41	0.41	0.43
Qwen2.5-VL-7B-Instruct	Chinese	0.46	0.47	0.32	0.35	0.39
	English	0.50	0.51	0.38	0.36	0.43
Qwen2.5-VL-72B-Instruct	Chinese	0.60	0.62	0.67	0.59	0.62
	English	0.56	0.62	0.60	0.53	0.58
Gemini 2.0 Flash-Lite	Chinese	0.64	0.59	0.64	0.58	0.61
	English	0.68	0.65	0.63	0.60	0.64

E. Image Validation for Dataset Construction

A total of 3873 images were collected, including 432 images featuring varied angles between objects. For each image, a forced-choice question "What is the person looking at?" was presented with three options (e.g., "1. the yellow object; 2. the blue object; 3. neither"), with object colors varying across images. The correct answer for each image was initially determined based on the intended gaze direction set by the experimenter during image collection.

The images and their corresponding questions were divided into five subsets, constructing five blocks of VQA tasks, each containing 770 to 778 images. To validate each question-answer pair, a total of 10 participants (all with normal vision and no color blindness) were recruited, with two participants assigned to finish each block. To ensure that images effectively demonstrate the intended gaze direction, only those for which both the participants provided the correct answer were retained. After the human validation, a total of 2809 images comprised the final dataset, with 2509 in the GIICS-BodyPosture subset and 300 in the GIICS-ObjectAngle subset.

IV. ANALYSIS ON VLLMS' PERFORMANCE

We tested nine VLLMs from five model families on our dataset of original, unmasked images, using both Chinese and English prompts. Table III demonstrated the mean accuracy of each model on the overall VQA task. The results showed that Gemini 2.0 Flash-Lite performed the best among all the models, with its accuracy over 55%. Most models from the Gemma-3 and Qwen2.5 families reached accuracy above chance level, but Gemma-3-4B-IT performed the worst, with its accuracy around 20%. Compared with the 100% human detection accuracy, VLLMs still faced challenges in semantically inferring gaze target.

A. VLLMs' Performance on GIICS-BodyPosture Subset

The GIICS-BodyPosture subset assessed the performance of VLLMs in inferring gaze targets when distracted by complex body-head direction combinations. The overall accuracy, as well as the accuracy breakdown for each spatial layout of objects, is shown in Table IV. The results revealed that Gemini 2.0 Flash-Lite performed the best on this subset, consistently achieving over 60% accuracy across all layouts of objects. In contrast, Gemma-3-4B-IT faced significant challenges in detecting gaze targets amid complex body-head direction combinations, with its accuracy dropping to 16%.

We further calculated the accuracy of models across different combinations of body-head directions, as presented in Table V. The results showed that most models performed the best when both the body and head faced the forefront (BF_HF), with Qwen2.5-VL-72B-Instruct and Gemini 2.0 Flash-Lite consistently achieving accuracy higher than 65% under this condition. However, for other consistent combinations of body-head directions (e.g., BL_HL and BR_HR), the accuracy did not differ much from that of inconsistent body-head directions. This may be attributed to the fact that the forward-facing posture provides the clearest view of the eyes' gaze direction. This further supported the idea that inconsistencies in body and head orientation can indeed affect models' detection of the gaze target.

B. VLLMs' Performance on GIICS-ObjectAngle subset

The GIICS-ObjectAngle subset evaluated the performance of VLLMs in gaze target detection, influenced by varied angles between objects. Table VI presented the overall accuracy of each model on this subset. The results revealed that gaze target detection was more challenging when angular variations between the target and the competitor were introduced, as most models performed worse (with below-chance accuracy), compared to the GIICS-BodyPosture subset, except for Gemma-3-4B-IT. The prevalent decline in accuracy indicated that changes in angles between objects increased difficulties in reasoning about human gaze by VLLMs.

Table VII further demonstrated the accuracy of models across four angular intervals between objects in the GIICS-ObjectAngle subset. The results showed that most models performed the best when the angle between objects was 120°, except for Gemma-3-4B-IT, which exhibited an accuracy pattern opposite to other models. Qwen2.5-VL-72B-Instruct consistently outperformed other models across both Chinese and English prompts; however, its accuracy was only 42%, indicating a significant gap from human-level competence in this regard.

C. Do VLLMs rely on body or head directions during inference?

We further explored **GIICS-BodyPosture** subset, focusing on scenes that the person gaze at an object. We removed scenes from the subset where the person is not gazing at any object and divided the rest of the scenes into following parts:

TABLE V
MODELS' ACCURACY ACROSS BODY–HEAD DIRECTION COMBINATIONS IN THE GIICS-BODYPOSTURE SUBSET

Model	Prompt language	Body–head directions									
		BF_HF	BL_HL	BR_HR	BF_HL	BF_HR	BF_HTL	BF_HTR	BF_HU	BL_HF	BR_HF
Gemma-3-4B-IT	Chinese	0.22	0.21	0.23	0.23	0.20	0.19	0.23	0.23	0.23	0.24
	English	0.16	0.17	0.15	0.18	0.14	0.16	0.18	0.15	0.15	0.17
Gemma-3-12B-IT	Chinese	0.61	0.38	0.44	0.42	0.44	0.49	0.50	0.59	0.49	0.50
	English	0.59	0.33	0.38	0.35	0.39	0.46	0.41	0.52	0.46	0.45
Gemma-3-27B-IT	Chinese	0.61	0.41	0.38	0.51	0.45	0.51	0.48	0.59	0.56	0.52
	English	0.68	0.50	0.55	0.59	0.60	0.64	0.57	0.65	0.70	0.66
Llama 3.2 Vision	Chinese	0.17	0.21	0.22	0.23	0.25	0.20	0.23	0.17	0.18	0.22
	English	0.23	0.20	0.19	0.23	0.24	0.24	0.23	0.26	0.16	0.21
Llama 4 Scout	Chinese	0.34	0.15	0.20	0.17	0.18	0.23	0.19	0.23	0.26	0.25
	English	0.54	0.20	0.22	0.18	0.17	0.31	0.30	0.25	0.34	0.37
LLaVA-1.5	Chinese	0.33	0.31	0.22	0.31	0.28	0.27	0.29	0.30	0.32	0.32
	English	0.41	0.45	0.45	0.35	0.45	0.41	0.45	0.42	0.42	0.45
Qwen2.5-VL-7B-Instruct	Chinese	0.45	0.27	0.30	0.33	0.39	0.42	0.43	0.38	0.43	0.44
	English	0.58	0.26	0.28	0.31	0.40	0.52	0.45	0.46	0.46	0.46
Qwen2.5-VL-72B-Instruct	Chinese	0.66	0.53	0.59	0.57	0.64	0.62	0.60	0.62	0.67	0.64
	English	0.65	0.40	0.51	0.49	0.61	0.56	0.63	0.56	0.65	0.62
Gemini 2.0 Flash-Lite	Chinese	0.65	0.64	0.64	0.54	0.64	0.60	0.53	0.61	0.64	0.61
	English	0.68	0.65	0.66	0.55	0.64	0.64	0.58	0.66	0.66	0.65

TABLE VI
MODELS' OVERALL ACCURACY IN THE GIICS-OBJECTANGLE SUBSET

Model	Prompt language	Accuracy
Gemma-3-4B-IT	Chinese	0.36
	English	0.41
Gemma-3-12B-IT	Chinese	0.26
	English	0.32
Gemma-3-27B-IT	Chinese	0.20
	English	0.18
Llama 3.2 Vision	Chinese	0.20
	English	0.22
Llama 4 Scout	Chinese	0.26
	English	0.29
LLaVA-1.5	Chinese	0.28
	English	0.24
Qwen2.5-VL-7B-Instruct	Chinese	0.19
	English	0.24
Qwen2.5-VL-72B-Instruct	Chinese	0.17
	English	0.22
Gemini 2.0 Flash-Lite	Chinese	0.16
	English	0.17

TABLE VII
MODELS' ACCURACY ACROSS DIFFERENT ANGLES BETWEEN OBJECTS IN THE GIICS-OBJECTANGLE SUBSET

Model	Prompt language	Angle between objects			
		120°	90°	60°	30°
Gemma-3-4B-IT	Chinese	0.26	0.46	0.41	0.38
	English	0.29	0.51	0.49	0.45
Gemma-3-12B-IT	Chinese	0.28	0.32	0.16	0.25
	English	0.33	0.39	0.18	0.38
Gemma-3-27B-IT	Chinese	0.38	0.14	0.03	0.10
	English	0.42	0.05	0.01	0.00
Llama 3.2 Vision	Chinese	0.31	0.15	0.15	0.08
	English	0.29	0.24	0.12	0.18
Llama 4 Scout	Chinese	0.29	0.30	0.19	0.20
	English	0.36	0.27	0.26	0.18
LLaVA-1.5	Chinese	0.28	0.32	0.26	0.20
	English	0.42	0.16	0.04	0.15
Qwen2.5-VL-7B-Instruct	Chinese	0.34	0.12	0.10	0.03
	English	0.42	0.18	0.12	0.03
Qwen2.5-VL-72B-Instruct	Chinese	0.42	0.03	0.01	0.00
	English	0.42	0.09	0.07	0.10
Gemini 2.0 Flash-Lite	Chinese	0.39	0.00	0.01	0.00
	English	0.40	0.04	0.01	0.00

- **Match**: Gaze direction aligns with both the body and head directions.
- **Partial match**: Gaze direction aligns with either the head G_{head} or the body G_{body} direction, but not both.
- **Mismatch**: Gaze direction aligns with neither the body nor the head direction.

We calculated the average accuary across both English and Chinese prompts of each model under these scenes in Table VIII. The result shows that most models perform significantly better (with a 10% increase in accuracy compare to mismatch condition) when gaze aligns with the directions of both head and body. In addition, most of the models also perform well when the gaze only aligns with head direction. However, when the gaze only aligns with body direction, the performance of most models does not differ much from the mismatch condition. The result indicates that models exhibit a slight dependency on both head and body directions when inferring gaze, with head direction exerting a greater influence on model performance rather than body direction.

D. Models' performance across different angles

We further experimented on **GIICS-ObjectAngle** subset, and plotted the psychometric curves [36] of different models as shown in Fig. 5 The figure illustrates the discrimination rates of different models across four angles. The discrimination rate of most models generally decreases as the gaze angle decreases from higher degrees to lower ones, showing that larger angle changes make it easier for the models to distinguish between objects. In addition, the majority of models show a sharp change in discrimination rate when moving from 120° to 90°, but the rate stabilizes or shows slight changes from 90° to

TABLE VIII
MODELS' ACCURACY OF GAZE TARGET DETECTION UNDER DIFFERENT ALIGNMENT CONDITIONS

Model	Match	Partial match		Mismatch
		G_{head}	G_{body}	
Gemma-3-4B-IT	0.45	0.46	0.37	0.44
Gemma-3-12B-IT	0.38	0.36	0.32	0.30
Gemma-3-27B-IT	0.40	0.28	0.28	0.24
Llama 3.2 Vision	0.55	0.52	0.52	0.52
Llama 4 Scout	0.72	0.59	0.53	0.42
LLaVA-1.5	0.36	0.38	0.36	0.33
Qwen2.5-VL-7B-Instruct	0.46	0.43	0.36	0.38
Qwen2.5-VL-72B-Instruct	0.43	0.38	0.23	0.29
Gemini 2.0 Flash-Lite	0.19	0.21	0.19	0.19
Mean	**0.44**	**0.40**	**0.35**	**0.34**

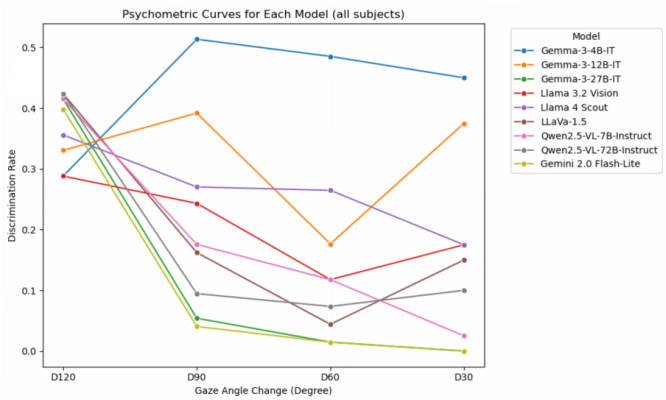

Fig. 5. Psychometric curves of different VLLMs on the GIICS-ObjectAngle subset

TABLE IX
MODELS' ACCURACY BEFORE VS. AFTER LORA FINE-TUNING ACROSS DIFFERENT SUBSETS

Model	Body-Head-Gaze direction		Angle between objects		
	Partial match	Mismatch	90°	60°	30°
Qwen2.5-VL-7B-Instruct	0.38	0.38	0.15	0.07	0.03
Qwen2.5-VL-7B-Instruct-LoRA	0.47	0.48	0.28	0.25	0.15

30°, indicating that models' ability to discriminate further is limited when the angle is reduced to a certain level.

E. Validation of Generalization through LoRA Fine-tuning

To investigate the generalization ability of VLLM across different subset metrics, we applied LoRA fine-tuning [37] to Qwen2.5-VL-7B-Instruct and evaluated it on subsets of GIICS. Table IX shows the results, indicating that LoRA fine-tuning yields clear generalization across conditions. In the Body-Head-Gaze Direction subset focusing on scenes that the person gaze at an object, fine-tuning on Match data led to consistent improvements under the more challenging Partial Match and Mismatch conditions, suggesting that discriminative features learned from highly aligned body-head scenarios can transfer to more complex cases with weaker or no alignment. Similarly, in the GIICS-ObjectAngle subset, fine-tuning with the most distinctive 120° condition resulted in substantial gains at 90°, 60°, and even 30°, demonstrating that features learned from large angular separations can generalize effectively to smaller and more ambiguous angles.

V. CONCLUSION

Gaze is an important social cue to infer intention. Here, we constructed the GIICS dataset to evaluate gaze-based intention inference. The results demonstrated that current state-of-the-art VLLMs faced significant challenges in accurately detecting gaze targets, particularly when distracted by complex body-head direction combinations, as seen in the GIICS-BodyPosture subset. VLLMs continued to face difficulties when the visual angle between two objects were as large as 120°, as seen in the GIICS-ObjectAngle subset. By including diverse object spatial positions, inconsistent body-head orientations and varied angles between objects, this dataset offered a novel challenge for models to reason about gaze direction in the real-world environments. In summary, the principled and diverse settings in GIICS made it a suitable benchmark to evaluate model performance on gaze-based intention inference in complex visual scenes.

ACKNOWLEDGMENT

This work was supported by the National Science and Technology Innovation 2030 Major Project 2021ZD0204100 (2021ZD0204105 to N.D.)

REFERENCES

[1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4971–4980, 2018. [Online]. Available: https://arxiv.org/abs/1712.00377

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: A visual language model for few-shot learning," *arXiv preprint arXiv:2208.02011*, 2022. [Online]. Available: https://arxiv.org/abs/2204.14198v1

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 2425–2433, 2015. [Online]. Available: https://arxiv.org/abs/1505.00468

[4] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, *et al.*, "Qwen2.5-VL technical report," *arXiv preprint arXiv:2502.13923*, 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[5] N. Burra, I. Mares, and A. Senju, "The influence of top-down modulation on the processing of direct gaze," *WIREs Cognitive Science*, 2019. [Online]. Available: https://doi.org/10.1002/wcs.1500

[6] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, M. Cornia, and R. Cucchiara, "The revolution of multimodal large language models: A survey," *arXiv preprint arXiv:2402.12451*, 2024. [Online]. Available: https://arxiv.org/abs/2402.12451

[7] L.-H. Chen, S. Lu, A. Zeng, H. Zhang, B. Wang, R. Zhang, and L. Zhang, "MotionLLM: Understanding human behaviors from human motions and videos," *arXiv preprint arXiv:2405.20340*, 2024. [Online]. Available: https://arxiv.org/abs/2405.20340

[8] M. Cimpoi, *et al.*, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3606–3613, 2014. [Online]. Available: https://doi.org/10.1109/CVPR.2014.461

[9] N. J. Emery, "The eyes have it: The neuroethology, function and evolution of social gaze," *Neuroscience & Biobehavioral Reviews*, vol. 24, no. 6, pp. 581–604, 2000. [Online]. Available: https://doi.org/10.1016/S0149-7634(00)00025-7

[10] R. Exline, D. Gray, and D. Schuette, "Visual behavior in a dyad as affected by interview content and sex of respondent," *J. Personality and Social Psychology*, vol. 1, no. 3, pp. 201–206, 1965. [Online]. Available: https://doi.org/10.1037/h0021865

[11] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 339–357, 2018. [Online]. Available: https://doi.org/10.1007/978-3-030-01249-6_21

[12] A. Gemma Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Rame, M. Rivière, *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[13] Google DeepMind, "Gemini 2.0," 2025. [Online]. Available: https://deepmind.google/technologies/gemini

[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[15] D. Guitton and M. Volle, "Gaze control in humans: Eye-head coordination during orienting movements to targets within and beyond the oculomotor range," *J. Neurophysiology*, vol. 58, no. 3, pp. 427–459, 1987.

[16] S. Hergeth, L. Lorenz, R. Vilimek, and J. F. Krems, "Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving," *Human Factors*, vol. 58, no. 3, pp. 509–519, 2016. [Online]. Available: https://doi.org/10.1177/0018720815625744

[17] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 6912–6921, 2019. [Online]. Available: https://arxiv.org/abs/1910.10088

[18] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2176–2184, 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.239

[19] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 19730–19742, 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 740–755, 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_48

[21] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 26296–26306, 2024.

[22] P. Lu, L. Ji, W. Zhang, N. Duan, M. Zhou, and J. Wang, "R-VQA: Learning visual relation facts with semantic attention for visual question answering," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 1880–1889, 2018. [Online]. Available: https://arxiv.org/abs/1805.09701

[23] P. Majaranta and A. Bulling, "Eye tracking and eye-based human–computer interaction," in *Advances in Physiological Computing*, S. H. Fairclough and K. Gilleade, Eds. Springer, pp. 39–65, 2014. [Online]. Available: https://doi.org/10.1007/978-1-4471-6392-3_3

[24] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3195–3204, 2019. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00331

[25] B. Massé, S. Ba, and R. Horaud, "Tracking gaze and visual focus of attention of people involved in social interaction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2711–2724, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2782819

[26] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 74–93, 2015. [Online]. Available: https://doi.org/10.1109/ICCV.2015.303

[27] K. Sanderson, "GPT-4 is here: What scientists think," *Nature*, vol. 615, no. 7954, p. 773, 2023. [Online]. Available: https://www.nature.com/articles/d41586-023-00816-5

[28] A. Senju and T. Hasegawa, "Direct gaze captures visuospatial attention," *Visual Cognition*, vol. 12, no. 1, pp. 127–144, 2005. [Online]. Available: https://doi.org/10.1080/13506280444000157

[29] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: Passive eye contact detection for human-object interaction," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, pp. 271–280, 2013.

[30] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal large language models: A survey," in *Proc. IEEE Int. Conf. Big Data (BigData)*, pp. 2247–2256, 2023. [Online]. Available: https://doi.org/10.1109/BigData59044.2023.10386743

[31] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6720–6731, 2019. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00688

[32] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[33] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 365–381, 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-58558-7_22

[34] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, 2019. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2778103

[35] S. Zhao, H. Yao, Y. Gao, G. Ding, and T. S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE Trans. Affective Comput.*, vol. 9, no. 4, pp. 526–540, 2016. [Online]. Available: https://doi.org/10.1109/TAFFC.2016.2628787

[36] R. Yssaad-Fesselier and K. Knoblauch, "Modeling psychometric functions in R," *Behavior Research Methods*, vol. 38, no. 1, pp. 28–41, 2006. [Online]. Available: https://doi.org/10.3758/BF03192747

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022. [Online]. Available: https://arxiv.org/abs/2106.09685