Journal of
Eye Movement Research

MDPI

Article

Gaze-to-Intent: Integrating Eye Tracking and a Large Language Model for Task Intention Understanding

Weifeng Zhang, Yangfei Zhao, Renbo Liu and Kaijiang Li *

School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China; zhangwf2588@outlook.com (W.Z.)
* Correspondence: riversky@126.com

Abstract

Existing gaze-based methods typically infer predefined target objects or task categories from gaze positions, areas of interest, or temporal eye-movement features, but few first select task-relevant objects and then infer a more complete task intention from object semantics. We recruited 12 participants to perform task-guided visual search in simulated warehouse-scene images with multiple objects and developed a staged Gaze-to-Intent framework. The proposed Heatmap-Based Gaze Localization method converts discrete gaze samples into a spatial gaze heatmap and combines it with object-instance regions to obtain task-relevant targets and relative gaze strengths. The selected target objects are represented by their natural-language names; a large language model (LLM) adapted using Low-Rank Adaptation (LoRA) then infers a short task-intention label and generates structured fields for the target object, required tool, and destination. The Heatmap method achieved an overall Target-Set Selection Accuracy of 94.1%, 8.1 and 5.9 percentage points higher than the DBSCAN and OPTICS baselines, respectively. The adapted LLM improved task inference and showed some generalization to object names unseen during adaptation. End-to-end results showed that gaze-derived candidates supported task intention understanding and that the gaze-informed candidate expansion strategy provided complementary evidence for relevance assessment. Thus, task-relevant object semantics can link gaze-based target selection to task intention inference in the present controlled task-guided visual-search setting.

Keywords: eye tracking; gaze-based target selection; task intention understanding; gaze heatmap; large language model; human–computer interaction

Check for updates

Academic Editors: Naser Al Madi and Michael Burch

Received: 24 September 2026
Revised: 16 November 2026
Accepted: 23 November 2026
Published: 25 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

When people interact with their environment, they typically first acquire environmental information through vision, locate objects relevant to the current task, and then act accordingly [1–4]. Consequently, understanding which objects a user is currently attending to and inferring the user's current task intention from visual behavior is an important research problem in human–computer interaction and assistive systems [5–9]. Eye tracking continuously records the locations of visual attention and their changes during visual search, thereby providing direct behavioral cues about the environmental objects attended by a user. Moreover, these attended objects and their combinations are closely related to the task that the user intends to accomplish [2,7,10–12]. Gaze data therefore provide a natural source of information for linking visual target selection to high-level task intention understanding [6,7,10,13]. If an intelligent system can select task-relevant targets based on a user's gaze behavior and further understand the behavioral goal associated with those targets, it can

obtain auxiliary information about the user's needs without requiring complex explicit commands, thereby supporting natural human–computer interaction, assistive operation, and human–robot collaboration [6,8,9,13–15].

Gaze reflects the allocation of visual attention in an environment. Determining the object that a user is looking at is therefore a fundamental problem in eye-tracking research and human–computer interaction [14,16–20]. Previous studies have commonly characterized visual behavior using gaze position, fixation duration, and areas of interest (AOIs), and have subsequently used this information for target selection, interface control, and interaction-state understanding. Plopski et al. classified gaze interaction in extended reality into three categories: explicit gaze interaction, implicit adaptive interaction, and collaborative gaze interaction [6].

In explicit gaze interaction, gaze is used directly as a pointing, selection, or manipulation signal. Fono and Vertegaal proposed EyeWindows, which selects and enlarges windows according to the user's visual focus [16]; Ahn and Lee combined gaze with touch input for text entry on smart glasses [17]; and Sidenmark and Gellersen exploited the coordination between eye and head movements to support rapid target pointing, preselection, and confirmation [18]. Unlike direct command execution, implicit gaze interaction primarily uses gaze as a source of information about user state and context. Nakano and Ishii estimated user engagement during interaction with an agent from gaze-transition patterns [19]; Gebhardt et al. learned personalized interaction policies from gaze data [13]; and Lindlbauer et al. combined user state with environmental context to adapt the position and amount of information in mixed-reality interfaces online [15]. In collaborative gaze interaction, gaze additionally serves as a cue for shared attention and task coordination. Vertegaal et al. investigated gaze patterns in conversation and their relationship to communicative attention [21]; Andrist et al. enhanced coordinated interaction between humans and virtual characters through bidirectional gaze mechanisms [22]; Gupta et al. showed that sharing a user's gaze location in remote collaborative tasks improves task coordination and co-presence [14]; and Erickson et al. further examined the effect of depth information on visual object recognition in shared-gaze environments [20].

Overall, existing studies can use two- or three-dimensional gaze information to localize regions of visual attention and establish correspondences between gaze and specific objects in a scene. However, visual search involves substantial natural viewing behavior [6,23,24], and users may attend to multiple objects unrelated to the current task. Looking at an object therefore does not necessarily mean that it is task-relevant, and selecting targets with clear task relevance from continuous gaze data remains a key challenge.

Once the objects attended by a user have been identified, inferring the task being performed from gaze behavior constitutes another important direction in eye-tracking research. Studies in psychology and human–computer interaction have shown that visual search is jointly influenced by the current task, goal, and environmental information [25–28]. Based on this observation, early work recognized tasks or activities primarily from gaze positions, scanpaths, and their temporal patterns [29–34]. Bulling et al. used eye-movement features, including fixations and saccades, to recognize everyday activities such as reading, writing, and web browsing [10]; Haji-Abolhassani et al. inferred an observer's visual task from fixation trajectories using a hidden Markov model [34]; and Boisvert et al. further investigated the contributions of the spatial distribution and dynamics of fixations and of image features to task prediction [32]. In first-person daily activities, Fathi et al. jointly modeled gaze location, scene objects, and action labels, showing that information about objects near the gazed-at region can effectively support activity recognition [31]. These studies demonstrate that gaze not only reflects visual-search patterns but also provides important behavioral cues for inferring the user's current task state.

As the field has developed, intention prediction has gradually shifted from task recognition based on gaze trajectories toward object-level reasoning that incorporates the semantics of attended objects. Koochaki et al. used gaze clustering to identify regions of interest and combined object recognition, object visitation sequences, and deep models to predict user intention, thereby mapping gaze targets to task categories [7]. More recently, Yin et al. combined gaze-region modeling with OPTICS clustering to identify objects of interest for subsequent intention prediction in real-world 3-D scenarios [35]. Multimodal information has also been used for intention prediction, for example by combining gaze and electromyographic signals to predict manipulation targets [36]. Chadalavada et al. used gaze to support the recognition of navigation intention in human–robot collaboration [9], while Zunino et al. predicted future behavioral intentions from two- and three-dimensional motion cues [37].

Existing approaches to task and intention inference can be broadly grouped into feature-based discriminative methods, probabilistic inference methods, and hybrid or multimodal methods. Feature-based methods use eye-movement descriptors to predict activity, task, or search-target labels. Probabilistic methods model latent states, temporal transitions, and dependencies among gaze, objects, and actions. Hybrid or multimodal methods combine spatial attention representations, object information, physiological signals, or motion cues. Table 1 summarizes representative studies, their theoretical foundations, and their main characteristics.

Table 1. Comparison of representative approaches to task and intention inference.

Method Category	Representative Studies	Theoretical Foundations	Main Characteristics
Feature-based discriminative methods	Bulling et al. [10]; Kanan et al. [29]; Boisvert et al. [32]; Borji et al. [33]	Task-dependent eye-movement statistics, feature selection, and supervised classification, including support vector machines.	Extract fixation, saccade, spatial, or dynamic features to predict activities, tasks, or search targets.
Probabilistic inference methods	Haji-Abolhassani and Clark [34]; Coutrot et al. [30]; Fathi et al. [31]	Hidden Markov models, Bayesian inference, and joint probabilistic generative models.	Represent latent states, temporal transitions, and gaze–object–action dependencies for task inference, scanpath classification, or action recognition.
Hybrid or multimodal methods	Koochaki et al. [7]; Krausz et al. [36]; Zunino et al. [37]; Yin et al. [35]	Spatial clustering, object recognition, temporal or deep models, and multimodal fusion, including Kalman-based fusion.	Koochaki et al. use DBSCAN to extract attention regions; Yin et al. use OPTICS and include DBSCAN as a comparator; Krausz et al. combine EMG and gaze; Zunino et al. combine 2-D and 3-D motion cues for intention prediction.

DBSCAN is selected as a target-selection baseline because it has been used to extract gaze-based attention regions [7], identifies dense regions and sparse noise, and requires no predefined cluster count. OPTICS provides an additional density-based baseline following Yin et al. [35].

Nevertheless, most existing methods address specific scenarios and a limited set of predefined task categories, and therefore have limited ability to accommodate unseen object names and task-intention labels.

This study develops a staged Gaze-to-Intent framework for task-guided visual search. The framework selects task-relevant objects from gaze data and uses their semantic information to infer task intentions and generate structured task information. First, gaze data collected under task-guided conditions are used to determine the task-relevant objects currently attended by a user. The proposed Heatmap method represents the spatial distribution of two-dimensional gaze samples as a continuous heatmap and accumulates the heatmap response within each object-instance region to select task-relevant targets and obtain their relative gaze strengths. DBSCAN and OPTICS serve as target-selection baselines by grouping spatially neighboring gaze samples according to density and matching the resulting attention regions to object instances. The selected target objects are then mapped to natural-language names, and a task-adapted large language model (LLM) uses its pretrained semantic knowledge and the functional relationships among the objects to infer the user's current task intention. The model produces a short task-intention label and structured information describing the target object, required tool, and destination, thereby enabling a progressive transformation from visual-attention locations to target objects and then to high-level task semantics. In addition, target selection and semantic reasoning are integrated into the complete Gaze-to-Intent framework to evaluate the feasibility of task intention inference from gaze input and the effect of gaze-informed candidate expansion in the present controlled task-guided visual-search setting.

On the basis of the preceding analysis, this study addresses the following three research questions concerning the progressive relationship between gaze-based target selection and task intention understanding:

RQ1: Can task-guided gaze data effectively support the selection of task-relevant targets, and how do different spatial processing methods for gaze affect target-selection performance?

RQ2: Under task-guided visual-search conditions, can a task-adapted LLM use task-relevant object semantics to infer the current task intention and generate the corresponding structured task information?

RQ3: In the present controlled task-guided visual-search setting, can the complete Gaze-to-Intent framework support task intention understanding from gaze input, and can gaze-informed candidate expansion further improve end-to-end performance?

2. Materials and Methods

2.1. Experimental Overview

This study used a task-guided eye-tracking paradigm to investigate how information about human visual attention can progressively support task-relevant target selection and high-level task intention understanding. During the experiment, participants viewed warehouse-scene images containing multiple candidate objects and received a task instruction corresponding to the current scene. They performed task-guided visual search according to the task requirements and selected targets by actively directing their gaze toward task-relevant objects, while the experimental system continuously recorded gaze data throughout each task.

Three spatial processing methods were first applied to the gaze data to determine the objects attended by the participants: a local Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method [7], an Ordering Points To Identify the Clustering Structure (OPTICS) method [35], and the proposed Heatmap-Based Gaze Localization method. All three methods used the same gaze data as input but differed primarily in how they represented the spatial gaze distribution and determined targets. The spatial localization results produced by the three methods were then matched to object instances in the scene

and converted into the corresponding natural-language object names, thereby establishing a mapping from gaze position to task-relevant object semantics.

Task intention inference was subsequently performed with an LLM on the basis of these object semantics. First, given the ground-truth task-relevant object names, we independently evaluated the model's ability to infer the task intention and structured fields from object semantics, and analyzed the effects of task adaptation and the model's semantic generalization ability under the unseen condition. The recorded gaze data, Heatmap-based target selection, and LLM-based inference were then integrated into the complete Gaze-to-Intent framework. We compared the Name-only method, which used only the names of the top three candidate objects, with the Gaze-informed method, which expanded the candidate set and retained relative gaze strength, to evaluate the effect of the gaze-informed candidate expansion strategy on subsequent task intention understanding. Overall, the framework first uses gaze information to determine the objects attended by a user and then uses object semantics to infer the corresponding high-level task intention.

2.2. Participants

Twelve participants were recruited for this study, including eight men and four women. They were 22–37 years old, with a mean age of 27.55 years, and were primarily university students. All participants had normal or corrected-to-normal vision, and eight had prior experience using an eye tracker. Before the experiment, the researchers explained its purpose and procedure to all participants, and data collection began after informed consent had been obtained. All experimental data were stored and analyzed anonymously. The study was reviewed and approved by the Ethics Committee of Zhengzhou University (approval no. ZZUIRB2025-106).

2.3. Eye-Tracking Apparatus

The eye-tracking apparatus, its main specifications, and the operating conditions are summarized in Table 2.

Table 2. Apparatus specifications and operating conditions.

Item	Specification or Condition
Eye tracker	aSeePro desktop eye tracker; sampling rate: 250 Hz; nominal spatial accuracy: 0.5°
Display	24-inch monitor; resolution: 1920 × 1080 pixels
Viewing distance	Approximately 55–80 cm
Lighting	Indoor fluorescent lighting; direct sunlight avoided
Head movement	No head restraint; participants were instructed to minimize head movement

Before the formal experiment, all participants completed eye-tracker calibration. A five-point calibration procedure was used: participants directed their gaze toward the calibration points presented sequentially on the screen, after which the correspondence between gaze position and target position was verified from the calibration results. Following calibration, the eye tracker continuously recorded two-dimensional gaze positions on the scene image and their corresponding timestamps for subsequent gaze localization and visual-attention analysis.

During data preprocessing, the raw gaze records were first checked for validity. Samples produced by participant blinks, brief loss of tracking by the eye tracker, or gaze positions outside the valid screen area were removed directly from the raw data without interpolation. All remaining valid two-dimensional gaze samples from the scene-viewing stage were re-

tained by trial and used for subsequent DBSCAN and OPTICS clustering and gaze-heatmap computation, without further subsampling.

2.4. Task Design and Experimental Procedure

The experimental tasks were constructed from images of warehouse environments. Each image contained multiple interactive objects and was associated with tasks appropriate to the scene. Three task-relevant objects were predefined for each task and corresponded, respectively, to the target object (`target`), required tool (`tool`), and destination (`to`) in the three structured task fields.

The experimental task set consisted of 60 scene-specific task instances, with three tasks defined for each of the 20 warehouse scenes. Three representative task examples are presented in Table 3.

Table 3. Representative warehouse tasks, intent labels, and associated key objects.

Task Description	Intent Label	Key Objects
Move the box to the temporary buffer area	Move to Buffer	Box, Forklift, Temporary Buffer
Expedite the pallet shipment to Shipping Area A	Rush Outbound	Pallet, Robot, Shipping Area A
Move the wooden crate to Return Area A	Return Inbound	Wooden Crate, AGV, Return Area A

These examples represent three typical warehouse operations: temporary buffering, expedited outbound handling, and return handling. Each task is described by a concise intent label and its associated key objects. The task instructions, LLM prompts, and object names used as LLM inputs were in Chinese; English translations are provided in this paper for presentation.

Participants independently searched the scene according to the given task and directed their gaze toward the targets they considered relevant to that task. No fixed viewing order was imposed on the task-relevant objects. Participants could complete target selection according to their own visual-search habits and were asked to minimize unnecessary repeated switching between targets, allowing task-guided visual-search and target-selection behavior to be recorded.

Before the formal experiment, participants viewed the scene images used in the study to familiarize themselves with the scene content and experimental format, after which they completed eye-tracker calibration. At the beginning of each trial, the task instruction was displayed for up to 15 s, allowing participants to read and understand the requirements of the current trial while taking a brief rest. Participants could proceed to the scene image earlier when they were ready. The scene image was then presented while gaze data were recorded synchronously. Participants searched the scene and looked at relevant objects according to the task requirements, and the maximum duration of the scene image was 15 s. This limit was based on preliminary testing indicating that participants could complete the visual-search and target-selection task within this period. When participants believed that target selection was complete, they could end the current trial early and proceed to the next task. The experimental procedure is illustrated in Figure 1.

The scene images were generated from a simulated warehouse environment. Each participant was randomly assigned nine images from the image pool and completed the corresponding tasks sequentially. All tasks were completed in two rounds separated by a 5 min break; thus, each participant was scheduled to complete 54 trials, yielding 648 trial records in total. Twenty-one trials were excluded because target selection had not been completed when the maximum viewing time was reached, leaving 627 valid trials. The recorded gaze data were organized by trial and used for subsequent gaze-based target selection and task intention understanding.

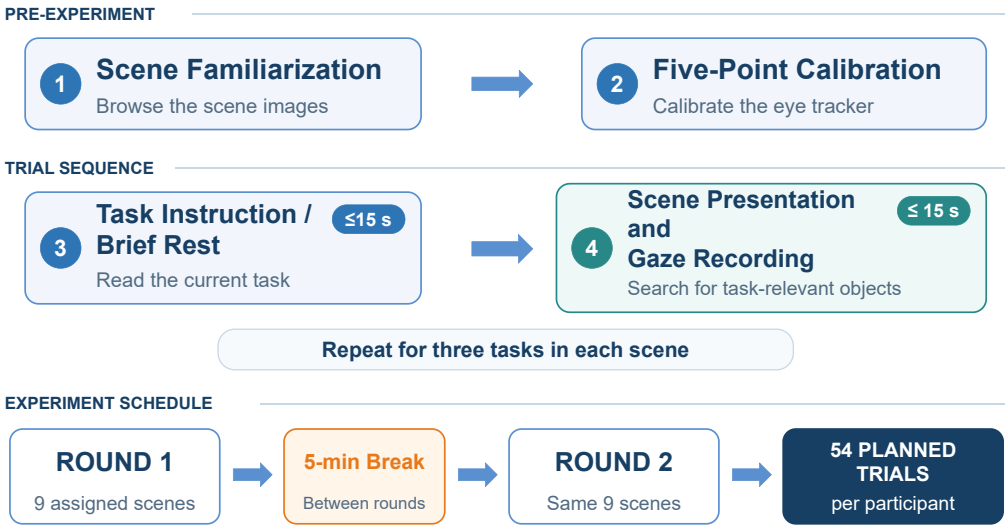


Figure 1. Procedure used for eye-tracking data collection.

2.5. Gaze-Based Target Selection Methods

Throughout the experiment, the eye tracker continuously recorded the participants' two-dimensional gaze positions in the scene images and the corresponding timestamps. These gaze records were first used to determine the principal regions of visual attention in the current task. The gaze positions were then associated with object instances in the scene to obtain task-relevant candidate objects.

To compare the effects of different spatial representations of gaze on target selection, the same gaze records were processed using DBSCAN-Based Gaze Localization, OPTICS-Based Gaze Localization, and Heatmap-Based Gaze Localization. The DBSCAN and OPTICS methods form discrete regions of attention according to the local spatial density of gaze samples, whereas the Heatmap method spatially accumulates discrete gaze samples to construct a continuous gaze heatmap representing the spatial distribution of visual attention. The spatial localization results produced by the three methods were subsequently matched to object instances in the scene and mapped to the corresponding natural-language object names, thereby completing the transformation from gaze information to task-relevant object semantics.

2.5.1. Gaze Localization

During each task trial, the eye tracker continuously recorded two-dimensional gaze positions in the scene image and their sampling times. The spatial distribution of the gaze samples collected during a trial was analyzed using DBSCAN and OPTICS clustering and a continuous gaze heatmap to determine the participant's principal regions of visual attention in the scene.

DBSCAN-Based Gaze Localization

Following [7], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was applied to the two-dimensional gaze samples collected within a trial; this approach is hereafter referred to as the DBSCAN method. The method automatically forms high-density clusters according to the spatial proximity and local density of the gaze samples and treats sparsely distributed samples as noise, thereby extracting the principal regions of visual attention without requiring the number of clusters to be specified in advance. The DBSCAN parameter settings are provided in Appendix A. Here, ε defines the neighborhood of a gaze sample, and MinPts specifies the minimum number of samples required to form a

high-density region. After clustering, candidate targets were determined from the spatial correspondences between the clustered regions and the object instances in the scene, and objects matched to valid clusters were treated as the candidate objects that received the participant's primary attention in the current trial.

OPTICS-Based Gaze Localization

OPTICS was applied to the same two-dimensional gaze samples as an additional clustering baseline, motivated by its use in gaze-based visual-attention modeling [35]. The method uses reachability distances to construct an ordering of the gaze samples and extracts clusters representing spatial attention regions. Samples not assigned to any cluster are labeled as noise. The resulting attention regions are associated with object instances in the scene to determine candidate targets.

Heatmap-Based Gaze Localization

Unlike the DBSCAN and OPTICS methods, which divide gaze samples into discrete clusters, the proposed approach represents the spatial distribution of gaze information using a continuous gaze heatmap; this approach is hereafter referred to as the Heatmap method. Suppose that a trial contains N two-dimensional gaze samples (x_i, y_i). A two-dimensional Gaussian kernel centered on each sample is superimposed to construct the corresponding gaze heatmap:

$$H(x,y) = \sum_{i=1}^{N} \exp\left(-\frac{(x-x_i)^2 + (y-y_i)^2}{2\sigma^2}\right), \quad (1)$$

where σ controls the spatial spread of an individual gaze sample in image space. Spatially concentrated gaze samples produce a high accumulated response in the corresponding region, whereas brief or spatially dispersed samples produce weaker local responses. The bandwidth was fixed at $\sigma = 50$ pixels, with each kernel truncated at $\pm 4\sigma$ per coordinate axis. Parameter sensitivity is evaluated in Section 3.1.

The heatmap is then normalized over the complete image domain Ω:

$$\widetilde{H}(x,y) = \frac{H(x,y)}{\sum_{(u,v)\in\Omega} H(u,v)}, \quad (2)$$

yielding the relative spatial gaze distribution for the current trial. This process transforms the discrete gaze samples into a continuous spatial response, allowing the degree of visual attention at different locations to be compared within a common spatial representation and providing a basis for matching gaze to object instances in the scene.

2.5.2. Object Semantic Mapping

After spatial gaze localization, the visual-attention information was associated with specific objects in the scene. EfficientSAM-S [38] was used to obtain instance-segmentation regions for the candidate objects, with an instance mask describing the spatial extent of each object in the image. YOLO-World [39] was used to recognize the corresponding object categories and obtain their natural-language class names. Each scene object was thus represented by both an instance region and a semantic name, providing the basis for transforming spatial gaze information into task-relevant object semantics.

For the Heatmap method, let R_i denote the instance-mask region of the ith candidate object. Its gaze score is defined as the accumulated normalized heatmap response within that region:

$$s_i = \sum_{(x,y)\in R_i} \widetilde{H}(x,y), \qquad i = 1, \ldots, M, \quad (3)$$

where M is the number of object instances in the current scene, and s_i is the relative gaze response received by the ith object region. Candidate objects are ranked according to s_i, and

the top m instances with the highest gaze scores are selected as candidate targets, with the value of m determined by the subsequent experimental setting. Global sum normalization scales all object scores within a trial by the same positive factor, preserving their ranking and hence Target-Set Selection Accuracy. The instance masks define the object regions, and target-set evaluation retains the three highest-scoring objects; no heatmap-value threshold is used for region extraction.

For DBSCAN and OPTICS, gaze clusters are matched to object instances using a point-in-mask rule. A cluster is associated with an object when the mean position of its gaze samples lies within the corresponding instance mask. The target instances selected by all three methods are subsequently mapped to their natural-language object names, completing the transformation from spatial gaze information to task-relevant object semantics. For the Heatmap method, the gaze score s_i associated with each candidate object is also retained as relative gaze strength for subsequent end-to-end task intention understanding.

2.6. LLM-Based Task Intention Inference

After the objects attended by a participant have been selected, their semantic information is used to infer the participant's current task intention. Gaze data primarily indicate where the user is looking in the scene, whereas the natural-language object names obtained through target selection and semantic mapping specify the objects being attended. The semantic combination of multiple target objects provides high-level cues for determining the current task, enabling a further transformation from visual-attention targets to task intention.

Qwen2.5-7B-Instruct was used as the task intention inference model and adapted to the task using Low-Rank Adaptation (LoRA) [40]. The model takes the names of the task-relevant objects associated with the current trial as its primary semantic input and infers the task that the user is likely to be performing from the functional relationships and task associations among the objects. The model outputs a short task-intention label and also produces structured fields specifying the target object (`target`), required tool (`tool`), and destination (`to`), thereby transforming task-relevant object semantics into structured information for high-level task understanding.

In the basic input representation, only the names of the three candidate objects with the highest gaze scores are provided to the model, such that task intention inference relies primarily on the semantic relationships among the candidate objects. In the end-to-end analysis, the Gaze-informed method expands the candidate set to the four objects with the highest gaze scores and provides both their names and relative gaze strengths to the LLM. Candidate expansion reduces the risk that a key task-relevant target will be omitted, while relative gaze strength provides complementary evidence for assessing the relevance of an additional candidate object to the current task.

2.7. Evaluation Protocol

Three experiments were conducted to separately evaluate gaze-based target selection, task intention inference from object semantics, and the complete Gaze-to-Intent framework. They corresponded, respectively, to the transformations from gaze information to target objects, from task-relevant object semantics to task intention, and from recorded gaze input to task-intention output.

2.7.1. Target Selection

The first experiment evaluated the effect of different spatial gaze representations on task-relevant target selection. The DBSCAN, OPTICS, and Heatmap methods were applied to the same set of experimental gaze records, and their spatial localization results were

matched to object instances in the scene to produce candidate-target sets. The three task-relevant objects predefined for each experimental task were used as the ground-truth target set. A trial was counted as correct only when all three ground-truth target objects were selected and no additional objects were included. Target-Set Selection Accuracy was calculated as the percentage of valid trials meeting this criterion. This metric was used to compare the ability of the three spatial processing methods to select task-relevant targets.

Overall accuracy was calculated across all valid trials, and participant-level accuracies were summarized as mean ± SD. The 95% confidence intervals for overall accuracy were estimated using cluster bootstrap with participants as the resampling units. Participant-level accuracies were compared using a Friedman test, followed by pairwise Wilcoxon signed-rank tests with Holm correction. Adjusted p values are reported for the pairwise comparisons. Effect sizes for the pairwise comparisons were quantified using the matched-pairs rank-biserial correlation (r_{rb}).

2.7.2. LLM Semantic Reasoning

To independently evaluate the model's ability to infer task intentions from task-relevant object semantics, this experiment excluded errors from the preceding gaze-based target-selection stage and directly used the ground-truth object names associated with each trial as model input. The data were divided by scene into training and test sets: 16 scenes (80%) provided 500 samples for LoRA adaptation, and four scenes (20%) provided 120 samples for testing. Seven additional valid trials were not included in the LLM training or seen test sets to obtain round sample counts.

The performance of the unadapted Qwen2.5-7B-Instruct model was first compared with that of the LoRA-adapted model to evaluate the effect of task adaptation on semantic reasoning. Adaptation used the prompt and structured-response format presented for the adapted model in Section 3.2 and was performed for 50 epochs. The LoRA configuration is presented in Appendix A.

The conventional test set was defined as the seen condition, allowing overlap in individual object names and intention labels with the adaptation data, while sharing no complete input–output samples. The held-out test samples were used only for evaluation. Each sample was submitted independently with its corresponding system prompt. The unadapted model's prompt explicitly specified the JSON output structure, whereas the adapted model used the simpler prompt presented in Section 3.2. A total of 120 textual test samples in which the object names were not observed during task adaptation were then used to form the unseen condition and evaluate the model's semantic generalization ability for unseen object names.

Intent Accuracy, All-Fields Accuracy, and Joint Accuracy were used as evaluation metrics. Intent Accuracy is the proportion of trials in which the predicted short task-intention label is semantically consistent with the ground-truth intention. All-Fields Accuracy requires all three structured fields—`target`, `tool`, and `to`—to be predicted correctly, whereas Joint Accuracy requires both a semantically correct task-intention label and correct predictions for all three structured fields.

2.7.3. End-to-End Gaze-to-Intent Evaluation

Following the preceding independent evaluations, the recorded gaze input, Heatmap method, object semantic mapping, and LoRA-adapted LLM were integrated into the complete Gaze-to-Intent framework. For each trial, candidate targets were first determined from the gaze scores calculated by the Heatmap method for the object-instance regions, and their natural-language object names were provided to the model to predict the task intention and structured task fields. To further analyze the role of gaze information in subsequent se-

mantic reasoning, two end-to-end input methods were considered. The Name-only method retained the three candidate objects with the highest gaze scores and provided only their natural-language names to the LLM. The Gaze-informed method expanded the candidate set to the four objects with the highest gaze scores and additionally retained the relative gaze strength of each candidate, reducing the risk that a key task-relevant target would fall outside the top three and providing evidence for assessing the relevance of the additional candidate object. The two methods were compared to evaluate the overall effect of the gaze-informed candidate expansion strategy on end-to-end task intention understanding. For the end-to-end evaluation, participant-level Joint Accuracy was compared between Name-only and Gaze-informed using a two-sided Wilcoxon signed-rank test.

3. Results

3.1. Gaze-Based Target Selection

The Heatmap, DBSCAN, and OPTICS methods were applied to the same gaze records to compare their performance in selecting task-relevant target sets.

As shown in Figure 2, the Heatmap method achieved the highest overall target-set selection accuracy across all valid trials, followed by OPTICS and DBSCAN. The Heatmap method outperformed DBSCAN and OPTICS by 8.1 and 5.9 percentage points, respectively. These results indicate higher target-set selection accuracy for the Heatmap method in the present task-guided visual-search setting.

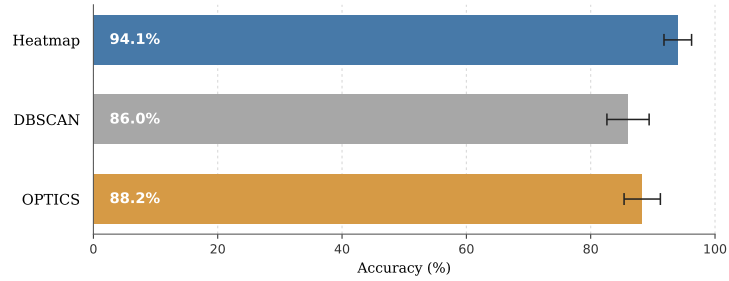


Figure 2. Overall target-set selection accuracy across all valid trials for the Heatmap, DBSCAN, and OPTICS methods. Bars represent overall accuracy, and error bars indicate 95% cluster-bootstrap confidence intervals obtained by resampling participants ($n = 12$).

At the participant level (Figure 3), mean target-set selection accuracy (mean ± SD) was highest for the Heatmap method (94.0% ± 4.1%), followed by OPTICS (88.2% ± 5.4%) and DBSCAN (85.9% ± 6.3%). The Heatmap method achieved higher accuracy than DBSCAN in 11 of the 12 participants and higher accuracy than OPTICS in 10 participants, with one tie. The advantage of the Heatmap method was observed in most participants.

A Friedman test of participant-level accuracies revealed a significant overall difference among the three methods ($\chi^2(2) = 15.27$, $p < 0.001$). Post hoc Wilcoxon signed-rank tests with Holm correction showed significant differences between Heatmap and DBSCAN (adjusted $p = 0.004$, $r_{rb} = 0.949$), Heatmap and OPTICS (adjusted $p = 0.010$, $r_{rb} = 0.909$), and OPTICS and DBSCAN (adjusted $p = 0.031$, $r_{rb} = 1.000$).

A sensitivity analysis of the paired Wilcoxon test used $n = 12$, 80% power, and a conservative two-sided significance threshold of $\alpha = 0.05/3$ for the three pairwise comparisons. Assuming normally distributed participant-level paired differences and using their observed standard deviations, simulations estimated minimum detectable differences of approximately 5.0 and 5.4 percentage points for Heatmap versus DBSCAN and OPTICS, respectively.

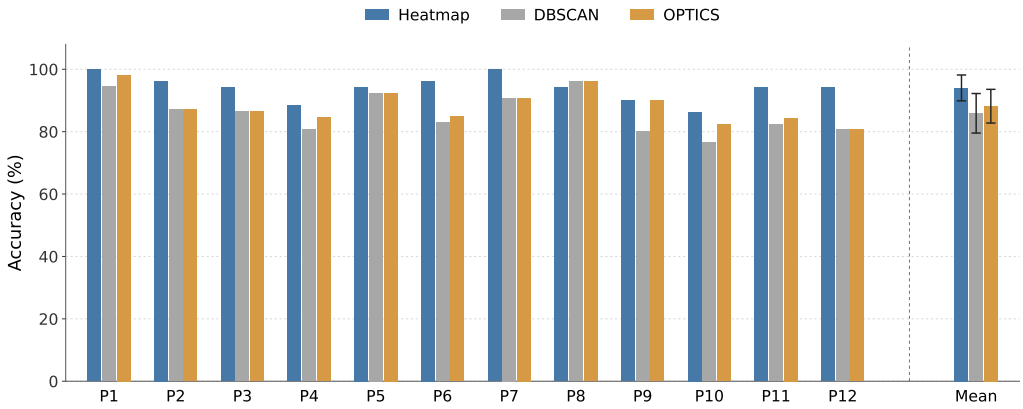


Figure 3. Participant-level Target-Set Selection Accuracy of the Heatmap, DBSCAN, and OPTICS methods. Bars for P1–P12 show each participant's accuracy. The Mean group shows the mean accuracy across the 12 participants, with error bars indicating the standard deviation.

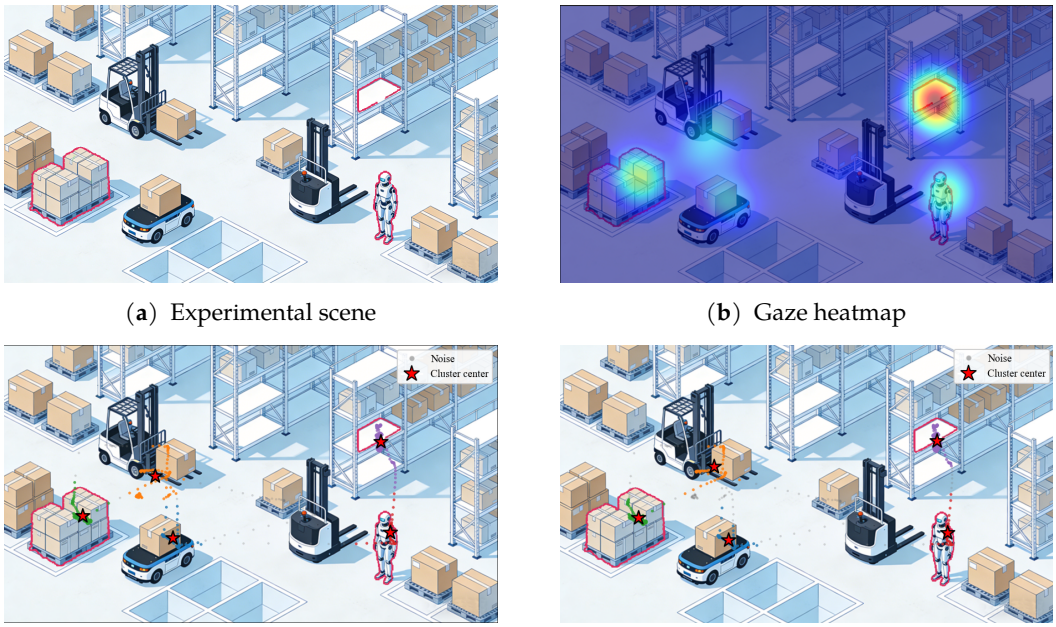
Figure 4 presents the target-localization results produced by the Heatmap, DBSCAN, and OPTICS methods for the same trial. Figure 4a shows the experimental scene, Figure 4b shows the continuous spatial gaze distribution generated by the Heatmap method, and Figure 4c shows the spatial clustering of gaze samples produced by the DBSCAN method. Figure 4d shows the spatial clustering results produced by the OPTICS method. The participant's gaze was concentrated primarily on task-relevant objects, although a small number of gaze samples were distributed over other regions during visual search. These samples outside the principal regions of visual attention were handled differently by the three spatial processing methods.

The DBSCAN method forms discrete clusters according to the local density of gaze samples and labels samples not assigned to any cluster as noise. When a task-relevant target receives a sufficiently large number of spatially concentrated gaze samples, the method can form a distinct cluster and match it to the target object. However, when the samples within a target region are sparse or dispersed, an effective cluster may not be formed. At the same time, a small number of relatively concentrated gaze samples may occur in a non-target region during visual search, producing an additional candidate region. As shown in Figure 4c, the target selections produced by the DBSCAN method are strongly influenced by the local density and spatial concentration of the gaze samples.

As shown in Figure 4d, OPTICS labels more gaze samples as noise and excludes them from the clusters compared with DBSCAN. The two clustering methods select the same candidate target objects in this trial.

In contrast, the Heatmap method transforms discrete gaze samples into a continuous spatial gaze distribution and accumulates the heatmap response within each object-instance region. As shown in Figure 4b, spatially concentrated gaze samples accumulated during sustained attention to task-relevant objects generate strong responses in the corresponding regions, whereas brief or dispersed samples in non-target regions generally produce only weak local responses. Candidate targets are selected by ranking the accumulated heatmap responses within the object-instance masks and retaining the highest-scoring objects.

Although the three methods process the same gaze records, they represent the spatial structure of visual attention differently. The DBSCAN and OPTICS methods represent gaze behavior as sets of discrete high-density regions, whereas the Heatmap method uses a continuous spatial response to characterize the overall distribution of gaze samples in the scene. On the present experimental data, the Heatmap method achieved higher Target-Set Selection Accuracy than both clustering baselines.

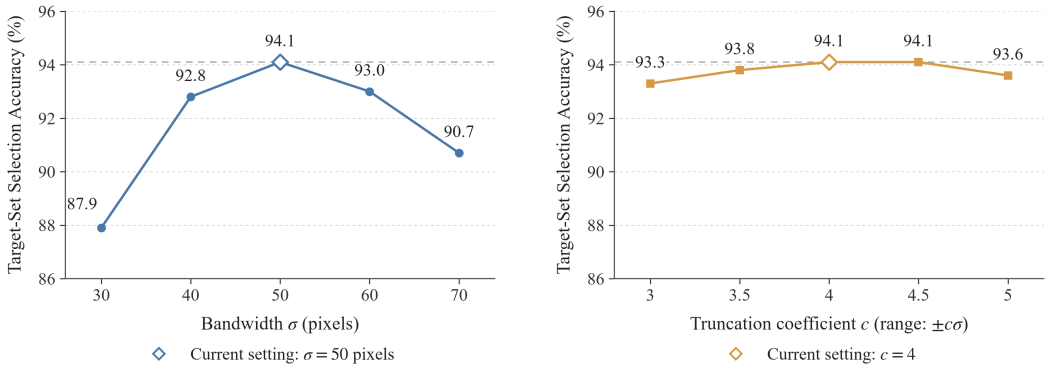


(**a**) Experimental scene (**b**) Gaze heatmap

(**c**) DBSCAN method (**d**) OPTICS method

Figure 4. Gaze-localization results for the same trial: (**a**) scene image; (**b**) Heatmap; (**c**) DBSCAN; (**d**) OPTICS. In (**c**,**d**), different dot colors indicate different gaze clusters. The red boxes in (**a**) indicate the task targets and were added for illustration only; participants viewed the scene without these boxes during the experiment.

Figure 5 presents the sensitivity of Target-Set Selection Accuracy to Gaussian bandwidth σ and kernel truncation coefficient c, with each parameter varied while the other settings were held fixed.



(**a**) Gaussian bandwidth ($c = 4$) (**b**) Kernel truncation ($\sigma = 50$ pixels)

Figure 5. Sensitivity of overall Target-Set Selection Accuracy to (**a**) Gaussian bandwidth and (**b**) kernel truncation. Diamonds indicate the current setting ($\sigma = 50$ pixels, $c = 4$), where the kernel is truncated at $\pm c\sigma$ per axis.

At $c = 4$, accuracy rose and then declined as bandwidth increased, peaking at $\sigma = 50$ pixels. Smaller bandwidths may leave responses overly localized around gaze samples displaced from the intended object regions, whereas larger bandwidths may spread responses into neighboring object regions and increase incorrect selections. At $\sigma = 50$ pixels, accuracy changed only slightly as c varied. Within the tested ranges, target-set selection accuracy was more sensitive to bandwidth than to kernel truncation.

3.2. Task Intention Understanding

After the gaze-based target-selection experiment, we further evaluated whether task-relevant object names provide effective semantic cues for LLM-based task intention understanding. For each trial, the ground-truth object names associated with the task were provided as textual input to Qwen2.5-7B-Instruct. The model inferred the corresponding task intention and produced a structured output containing the target object (`target`), required tool (`tool`), and destination (`to`). Intent Accuracy, All-Fields Accuracy, and Joint Accuracy were used to evaluate the predicted short task-intention label, all structured fields, and the complete task information, respectively.

To evaluate the model's ability to understand object semantics, we first compared prompt-based inference using the unadapted Qwen2.5-7B-Instruct model with inference using the model after LoRA-based task adaptation. To further examine semantic generalization beyond the task-adaptation data, textual test samples in which the object names had not appeared during task adaptation were constructed with reference to the domestic tidying scenarios reported in [41]. The model was therefore evaluated in terms of both the effect of task adaptation and its semantic generalization ability under the unseen condition. The results are presented in Table 4.

Table 4. Task intention understanding results for the different model configurations under the seen and unseen conditions.

Setting	Intent Acc. (%)	All-Fields Acc. (%)	Joint Acc. (%)
Without LoRA (Seen)	63.3	74.2	63.3
LoRA (Seen)	97.5	98.3	97.5
LoRA (Unseen)	91.7	95.0	90.8

As shown in Table 4, LoRA-based task adaptation substantially improved all three metrics under the seen condition. Intent Accuracy and Joint Accuracy both increased by 34.2 percentage points, while All-Fields Accuracy increased by 24.1 percentage points. These results indicate that task adaptation enables the model to use the semantic information conveyed by task-relevant objects more effectively for task intention inference and improves the consistency and accuracy of the structured-field output.

Figure 6 provides a qualitative comparison of the inputs and responses before and after LoRA adaptation. The unadapted model received a system prompt specifying the JSON output structure and sometimes returned additional natural-language explanations alongside the requested output. The intention label `Transport` is semantically plausible for this example, although the label wording and accompanying explanations varied across responses. The LoRA-adapted model used the simpler system prompt shown in the figure. It produced the inferred intention and the `target`, `tool`, and `to` fields in the expected JSON format.

Figure 7 presents a qualitative example under the unseen condition, using red clothes, a robotic arm, and a washing machine as the object-name input to the LoRA-adapted model. The model inferred the intention "Clean" and assigned Red Clothes, Robotic Arm, and Washing Machine to the `target`, `tool`, and `to` fields, respectively. This example shows a successful response in the expected JSON format for object names and an intention label not encountered during task adaptation.

Performance under the unseen condition was lower than under the seen condition but remained high overall, with Joint Accuracy decreasing by 6.7 percentage points. Thus, when the object names in the test samples had not been observed during task adaptation, the model could still use object semantics to predict most task intentions and structured fields, demonstrating a degree of semantic generalization to unseen object names.

(a) Without LoRA

System prompt (provided with each input): You are an AR assistant. Please infer the user's intention based on the set of objects they have gazed at. Return the output in the following format: {"intentid":"...","target":"...","tool":"...","to":"..."}

User input: User gaze: Pallet, Forklift, Buffer Area.

Model response: Based on the objects the user has gazed at, it can be inferred that the user may intend to use the forklift to move the pallet to the buffer area. The intention output is therefore as follows:
{"intentid":"Transport","target":"Pallet","tool":"Forklift","to":"Buffer Area"}

(b) With LoRA

System prompt: You are an AR assistant. Please infer the user's intention based on the set of objects they have gazed at.

User input: User gaze: Pallet, Forklift, Buffer Area.

Model response: {"intentid":"Move to Buffer","target":"Pallet","tool":"Forklift", "to":"Buffer Area"}

Figure 6. Comparison of model responses before and after LoRA adaptation using the same object-name input.

With LoRA: Unseen Scenario

System prompt: You are an AR assistant. Please infer the user's intention based on the set of objects they have gazed at.

User input: User gaze: Red Clothes, Robotic Arm, Washing Machine.

Model response: {"intentid":"Clean","target":"Red Clothes","tool":"Robotic Arm", "to":"Washing Machine"}

Figure 7. Illustrative response of the LoRA-adapted model to object names not encountered during task adaptation.

3.3. End-to-End Gaze-to-Intent Understanding

The preceding experiments separately demonstrated the role of gaze information in task-relevant target selection and the ability of task-relevant object semantics to support task intention inference by an LLM. On this basis, an end-to-end Gaze-to-Intent experiment was conducted to evaluate the complete process from gaze input to task-intention output. For each trial, the Heatmap method first processed the recorded gaze data and ranked the object instances according to the heatmap response accumulated within their regions. The three objects with the highest gaze scores were selected as candidate targets, and their names were provided to the LoRA-adapted Qwen2.5-7B-Instruct model, which inferred the current task intention and generated structured task fields, including the target object (`target`), required tool (`tool`), and destination (`to`). The complete pipeline combines gaze-based target selection, object-semantic mapping, and LLM-based task intention inference and was evaluated using gaze data recorded in the present controlled task-guided visual-search setting.

During task-guided visual search, brief glances or incidental attention to neighboring objects in regions containing densely arranged task targets may prevent some key task-relevant targets from entering the three candidates with the highest gaze scores. To reduce the risk of omitting a key target, we further adopted a gaze-informed candidate expansion strategy and considered two end-to-end input methods. The first method retained the three candidate objects with the highest gaze scores and provided only their object names to the LLM; this method is denoted as Name-only. The second method expanded the candidate set to the four

objects with the highest gaze scores and provided both the name and relative gaze strength of each candidate; this method is denoted as Gaze-informed. Candidate expansion increases the likelihood that key task-relevant targets enter the candidate set, while relative gaze strength provides complementary evidence for distinguishing task-relevant targets from an additional candidate object. The two methods were compared to evaluate the overall effect of the gaze-informed candidate expansion strategy on end-to-end task intention understanding.

As shown in Table 5, the Name-only method maintained high end-to-end performance when using only the names of the three candidate objects selected by the Heatmap method for task intention inference. Candidate-selection errors can affect subsequent task-intention inference, as illustrated by the following error analyses.

Table 5. End-to-end evaluation results for the Gaze-to-Intent framework.

Input Method	Intent Acc. (%)	All-Fields Acc. (%)	Joint Acc. (%)
Name-only	92.7	92.2	92.2
Gaze-informed	94.9	94.6	94.6

After the Gaze-informed method expanded the candidate set from three to four objects and retained their relative gaze strengths, all three metrics improved, including an increase of 2.4 percentage points in Joint Accuracy. This improvement suggests that candidate expansion and relative gaze strength jointly provide useful information for subsequent semantic reasoning in end-to-end task intention understanding.

At the participant level, mean Joint Accuracy (mean ± SD) was 92.1% ± 4.7% for Name-only and 94.5% ± 2.9% for Gaze-informed. Gaze-informed achieved higher Joint Accuracy in 10 of the 12 participants, with two ties. A two-sided Wilcoxon signed-rank test showed a significant difference between the two methods ($p = 0.002$, $r_{rb} = 1.000$).

A sensitivity analysis of the paired Wilcoxon test used $n = 12$, 80% power, and a two-sided significance threshold of $\alpha = 0.05$. Assuming normally distributed participant-level paired differences and using their observed standard deviation, simulations estimated a minimum detectable difference of approximately 2.1 percentage points in Joint Accuracy.

Across all end-to-end trials, the top-three candidates failed to cover all ground-truth objects in 37 trials. Within this subset, the fourth-ranked candidate recovered a missing ground-truth object in 25 trials (67.6%), while a non-ground-truth object was added in the remaining 12 (32.4%). As shown in Table 6, candidate expansion restored coverage of all three ground-truth objects in 19 trials. Gaze-informed produced outputs with all three structured fields correct in 14 of these trials, while field errors remained in the other five. In contrast, the Name-only inputs still lacked key object information, limiting the generation of correct structured fields. In the remaining 18 trials with incomplete candidate coverage, each method correctly predicted some task-intention labels, but neither produced an output with all three structured fields correct.

Table 6. Candidate coverage and model outputs after top-three omissions.

		Name-Only		Gaze-Informed	
Top-Four Coverage	**Trials**	**Intent**	**All-Fields**	**Intent**	**All-Fields**
All targets covered	19	2	0	14	14
Targets still missing	18	1	0	2	0
Total	37	3	0	16	14

Values are trial counts. Intent and All-Fields report the numbers of trials meeting the correctness criteria used for Intent Accuracy and All-Fields Accuracy, respectively.

For the full set of end-to-end trials, Table 7 summarizes the structured-field errors. Gaze-informed produced fewer errors than Name-only in the target-object, tool, and destination

fields. When task-relevant information is missing from the candidate input, the model may generate field values that are absent from the input and inconsistent with the ground truth, or leave the corresponding fields empty. The model may also assign other candidate objects to incorrect task roles.

Table 7. Structured-field errors in end-to-end evaluation.

Error Type	Name-Only	Gaze-Informed
Target-object error	21	17
Tool error	17	11
Destination error	21	13
Any field error	49	34

Values are trial counts across 627 end-to-end trials. Field-specific error categories may overlap. The final row counts each trial once if at least one field is incorrect, regardless of the number of incorrect fields.

Figure 8a,b illustrates a failure case and a successful outcome using candidates from the same Heatmap result. The ground-truth task objects were Box, AGV, and Return Area. The three highest-ranked candidates were Box, Return Area, and Shelf, while AGV ranked fourth. With these three object names as input, the Name-only model inferred the correct intention, "Move to Return Area", but generated "Manual" in the tool field. This hallucinated value was unsupported by the input objects and inconsistent with the ground-truth tool. The output therefore satisfied the intention-label criterion but failed both the All-Fields and Joint Accuracy criteria, showing how an upstream selection omission can propagate into an incorrect task field even when the intention label remains correct.

(**a**) **Name-only**

System prompt: You are an AR assistant. Please infer the user's intention based on the set of objects they have gazed at.

User input: User gaze: Box, Shelf, Return Area.

Model response: {"intentid":"Move to Return Area","target":"Box","tool":"Manual", "to":"Return Area"}

(**b**) **Gaze-informed**

System prompt: You are an AR assistant. Please infer the user's intention based on the set of objects they have gazed at and their gaze strengths.

User input: User gaze: Box (0.27), Shelf (0.16), Return Area (0.21), AGV (0.13).

Model response: {"intentid":"Move to Return Area","target":"Box","tool":"AGV", "to":"Return Area"}

(**c**) **Gaze-informed**

System prompt: You are an AR assistant. Please infer the user's intention based on the set of objects they have gazed at and their gaze strengths.

User input: User gaze: Parcel Bag (0.25), Robot (0.18), Drone (0.21), Flowers (0.16).

Model response: {"intentid":"Rush to Robot","target":"Parcel Bag","tool":"Drone", "to":"Robot"}

Figure 8. Examples of error propagation from Heatmap-based target selection to task intention inference: (**a**) a tool-field error with Name-only input; (**b**) a correct output with Gaze-informed input; (**c**) destination and intention-label errors when the candidate set omits the destination. Panels (**a**,**b**) show outputs from the same trial.

The Gaze-informed input in Figure 8b included the fourth-ranked AGV together with the relative gaze strengths of all four candidates. In this case, the model correctly assigned Box, AGV, and Return Area to the target, tool, and destination fields, respectively, while leaving Shelf out of the structured output. Both the intention label and all task fields were correct. The expanded candidate set thus supplied the tool information missing from the Name-only input.

Figure 8c presents an additional failure case. The ground-truth intention was "Move to Return Area", with Parcel Bag, Drone, and Return Area as the target object, tool, and destination, respectively. The four input candidates omitted Return Area. The model correctly selected Parcel Bag and Drone as the target and tool, but assigned Robot to the destination field and predicted "Rush to Robot". This example illustrates how missing destination information can propagate into errors in both the destination field and the intention label.

4. Discussion

This study investigated how task-guided gaze can support target selection and task intention understanding, yielding three principal findings. First, task-guided gaze information effectively supported the selection of task-relevant targets. Given the same gaze input, the Heatmap method achieved higher Target-Set Selection Accuracy than the DBSCAN and OPTICS baselines in the multi-target visual-search scenario considered in this study. Second, given task-relevant object semantics, the LoRA-adapted LLM accurately inferred task intentions and generated structured task information describing the target object, required tool, and destination. The model also exhibited a degree of semantic generalization to object names that were not observed during task adaptation. Finally, in the complete Gaze-to-Intent framework, candidate targets derived from recorded gaze data supported task intention understanding. Model performance improved further with the gaze-informed candidate expansion strategy. This improvement indicates that expanding the candidate set while retaining relative gaze strength provided a more informative candidate representation for subsequent task reasoning.

4.1. Interpretation of the Research Questions

4.1.1. RQ1: Gaze-Based Target Selection

The results for RQ1 show that task-guided gaze information can effectively support the selection of task-relevant targets. When the same gaze records were processed, the Heatmap method achieved higher Target-Set Selection Accuracy than the DBSCAN and OPTICS baselines. The three methods differ primarily in how they represent the spatial distribution of gaze. The DBSCAN and OPTICS methods represent gaze samples as discrete clusters and determine candidate targets from the resulting attention regions. The Heatmap method instead spatially accumulates gaze samples to form a gaze heatmap with responses of varying strength and then calculates the response within each object region to rank candidate targets according to their gaze scores. Task-relevant objects generally receive high accumulated responses, whereas responses produced by brief or dispersed gaze samples are relatively weak. The response-strength differences retained by the Heatmap method therefore provide more fine-grained information for distinguishing candidate targets, which may explain its higher target-selection performance in the present multi-target visual-search scenario.

4.1.2. RQ2: Object Semantics for Task Intention Understanding

For RQ2, the LoRA-adapted model substantially outperformed the unadapted base model in predicting task intentions and structured task information. This result indicates

that task adaptation enables an LLM to use task-relevant object semantics more consistently for task intention inference. Although a general-purpose LLM possesses knowledge about objects and tasks, adaptation to the semantic relationships among task-relevant objects and to the required output format helped transform these relationships into high-level task information more consistently. Rather than outputting only a fixed task label, the proposed representation describes a task in terms of its task intention, target object, required tool, and destination, allowing the model to express more complete task content. The model also maintained high performance under the unseen condition, suggesting that semantic representations based on natural-language object names provide a degree of generalization and that task inference does not depend entirely on direct mappings between fixed object categories and task labels observed during task adaptation.

4.1.3. RQ3: End-to-End Gaze-to-Intent Understanding

The end-to-end results for RQ3 support the feasibility of task intention inference from gaze input in the present controlled task-guided visual-search setting. During actual visual search, brief glances, attention to neighboring objects, or the presence of multiple semantically related objects can affect the ranking of candidate targets. A key task-relevant target may therefore fall outside the three candidates with the highest gaze scores. The Gaze-informed method expands the candidate set to four objects and simultaneously retains the relative gaze strength of each candidate. In the present end-to-end evaluation, the Gaze-informed configuration achieved higher Joint Accuracy than the Name-only method. The example in Figure 8a,b illustrates how a candidate omission can lead to a hallucinated tool field despite a correct intention label; including the missing object among the expanded candidates supported a correct output in that example.

4.2. Limitations and Future Work

This study establishes a pipeline from gaze-based target selection to structured task intention understanding. The evaluation involved 12 participants and a set of 20 simulated warehouse-scene images under task-guided visual-search conditions. The sample size and controlled task conditions limit the generalizability and ecological validity of the findings.

Predefined tasks and task-guided viewing reduced uncertainty about participants' goals and relevant objects. Under free viewing, gaze strength may correspond less closely to task relevance. With open-ended intentions, even correctly selected objects may support several plausible tasks, potentially reducing the accuracy of intention labels and structured fields. During multi-step activities or interruptions that redirect attention, accumulated gaze may mix objects from different subgoals, causing current targets to be omitted and selection errors to propagate into task inference. Performance under free viewing and continuous real-world interaction requires further evaluation.

The generated task information, including the target object, required tool, and destination, provides structured semantic inputs for subsequent planning and control. However, the current evaluation mainly focuses on intention understanding rather than physical task execution. Future work will explore the integration of the proposed framework with robot platforms to examine its effectiveness in real-world interaction scenarios.

Future work will integrate robot task-planning and control modules by associating the target objects, tools, and destinations generated by the model with physical objects, spatial locations, and manipulation commands in the environment, enabling a robot to perform operations such as grasping, transporting, and placing objects. On this basis, a closed-loop interaction system spanning gaze input, target selection, task intention understanding, and physical robot execution can be developed, extending the proposed Gaze-to-Intent framework to human–robot collaboration in real-world tasks.

5. Conclusions

This study developed a staged Gaze-to-Intent framework that transforms task-guided gaze information into task-relevant object semantics and subsequently into high-level task intentions. The Heatmap method first models the spatial distribution of gaze samples and combines it with object-instance regions to select the task-relevant targets attended by the user. The selected objects are then mapped to natural-language semantics, and a task-adapted LLM infers the task intention and generates structured task information describing the target object, required tool, and destination.

In the present controlled task-guided visual-search setting, the Heatmap method achieved higher Target-Set Selection Accuracy than the DBSCAN and OPTICS baselines. The LoRA-adapted LLM mapped sets of task-relevant object names to short intention labels and structured fields, with some generalization to object names not encountered during task adaptation. In the present end-to-end evaluation, the Gaze-informed method achieved higher Joint Accuracy than the Name-only method.

The findings provide preliminary evidence for the feasibility of combining gaze-based selection of task-relevant targets with LLM-based task intention inference and structured task information generation in the present controlled task-guided visual-search setting. Future integration of the target-object, tool, and destination information with robot task-planning and control modules may enable a closed-loop interaction process from gaze input and task understanding to physical task execution.

Author Contributions: Conceptualization, W.Z. and K.L.; methodology, W.Z. and K.L.; software, Y.Z.; validation, Y.Z. and W.Z.; formal analysis, W.Z.; investigation, W.Z. and R.L.; resources, K.L.; data curation, R.L.; writing—original draft preparation, W.Z.; writing—review and editing, W.Z. and K.L.; visualization, Y.Z.; supervision, K.L.; project administration, R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was reviewed and approved by the Ethics Committee of Zhengzhou University (approval no. ZZUIRB2025-106; date of approval: 14 March 2025).

Informed Consent Statement: Written informed consent was obtained from all participants before data collection. All participants were fully informed about the experimental procedures and voluntarily agreed to participate in this study.

Data Availability Statement: The data supporting this study are available from the corresponding author upon reasonable request. Raw eye-tracking recordings and participant-related behavioral data contain potentially identifiable behavioral traces and cannot be made publicly available under the approved consent and data-management arrangements. Processed gaze data, experimental task descriptions, and analysis results may be available upon reasonable request.

Acknowledgments: The authors would like to thank all participants who contributed to the eye-tracking experiment and provided the experimental data used in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AOI	Area of Interest
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
LLM	Large Language Model
LoRA	Low-Rank Adaptation
SAM	Segment Anything Model
SD	Standard Deviation
YOLO	You Only Look Once

Appendix A

The explicitly specified hyperparameter configurations of the gaze localization methods and LLM adaptation are summarized in Table A1. Other unspecified parameters of DBSCAN and OPTICS were set to the default values of the corresponding scikit-learn implementations.

Table A1. Hyperparameter configurations of the methods used in this study.

Method/Model	Hyperparameter Configuration
DBSCAN	$\varepsilon = 90$ pixels, MinPts = 50
OPTICS	max_eps = 90 pixels, min_samples = 50, $\xi = 0.1$, min_cluster_size = None
Heatmap	$\sigma = 50$ pixels, kernel=2D Gaussian, numerical truncation= $\pm 4\sigma$ per axis
Qwen2.5-7B-Instruct (LoRA)	$r = 8$, $\alpha = 16$, dropout= 0, learning rate= 1×10^{-4}, epochs= 50, target modules=`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`

References

1. Kytö, M.; Ens, B.; Piumsomboon, T.; Lee, G.A.; Billinghurst, M. Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2018; pp. 1–14. [CrossRef]
2. Toyama, T.; Kieninger, T.; Shafait, F.; Dengel, A. Gaze Guided Object Recognition Using a Head-Mounted Eye Tracker. In *Proceedings of the Symposium on Eye Tracking Research and Applications*; ACM: New York, NY, USA, 2012; pp. 91–98. [CrossRef]
3. Jacob, R.J.K. The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look At Is What You Get. *ACM Trans. Inf. Syst.* **1991**, *9*, 152–169. [CrossRef]
4. Land, M.F.; Furneaux, S. The Knowledge Base of the Oculomotor System. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **1997**, *352*, 1231–1239. [CrossRef] [PubMed]
5. Majaranta, P.; Bulling, A. Eye Tracking and Eye-Based Human–Computer Interaction. In *Advances in Physiological Computing*; Springer: London, UK, 2014; pp. 39–65. [CrossRef]
6. Plopski, A.; Hirzle, T.; Norouzi, N.; Qian, L.; Bruder, G.; Langlotz, T. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-Worn Extended Reality. *ACM Comput. Surv.* **2022**, *55*, 1–39. [CrossRef]
7. Koochaki, F.; Najafizadeh, L. A Data-Driven Framework for Intention Prediction via Eye Movement with Applications to Assistive Systems. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 974–984. [CrossRef] [PubMed]
8. Novak, D.; Riener, R. Enhancing Patient Freedom in Rehabilitation Robotics Using Gaze-Based Intention Detection. In *Proceedings of the IEEE International Conference on Rehabilitation Robotics*; IEEE: Piscataway, NJ, USA, 2013; pp. 1–6. [CrossRef] [PubMed]
9. Chadalavada, R.T.; Andreasson, H.; Schindler, M.; Palm, R.; Lilienthal, A.J. Bi-Directional Navigation Intent Communication Using Spatial Augmented Reality and Eye-Tracking Glasses for Improved Safety in Human–Robot Interaction. *Robot.-Comput.-Integr. Manuf.* **2020**, *61*, 101830. [CrossRef]
10. Bulling, A.; Ward, J.A.; Gellersen, H.; Tröster, G. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 741–753. [CrossRef] [PubMed]
11. Ou, J.; Oh, L.M.; Fussell, S.R.; Blum, T.; Yang, J. Predicting Visual Focus of Attention from Intention in Remote Collaborative Tasks. *IEEE Trans. Multimed.* **2008**, *10*, 1034–1045. [CrossRef]
12. Giannopoulos, I.; Kiefer, P.; Raubal, M. GazeNav: Gaze-Based Pedestrian Navigation. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*; ACM: New York, NY, USA, 2015; pp. 337–346. [CrossRef]
13. Gebhardt, C.; Hecox, B.; van Opheusden, B.; Wigdor, D.; Hillis, J.; Hilliges, O.; Benko, H. Learning Cooperative Personalized Policies from Gaze Data. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*; ACM: New York, NY, USA, 2019; pp. 197–208. [CrossRef]
14. Gupta, K.; Lee, G.A.; Billinghurst, M. Do You See What I See? The Effect of Gaze Tracking on Task Space Remote Collaboration. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 2413–2422. [CrossRef] [PubMed]
15. Lindlbauer, D.; Feit, A.M.; Hilliges, O. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*; ACM: New York, NY, USA, 2019; pp. 147–160. [CrossRef]
16. Fono, D.; Vertegaal, R. EyeWindows: Evaluation of Eye-Controlled Zooming Windows for Focus Selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2005; pp. 151–160. [CrossRef]
17. Ahn, S.; Lee, G. Gaze-Assisted Typing for Smart Glasses. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*; ACM: New York, NY, USA, 2019; pp. 857–869. [CrossRef]

18. Sidenmark, L.; Gellersen, H. Eye&Head: Synergetic Eye and Head Movement for Gaze Pointing and Selection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*; ACM: New York, NY, USA, 2019; pp. 1161–1174. [CrossRef]
19. Nakano, Y.I.; Ishii, R. Estimating User's Engagement from Eye-Gaze Behaviors in Human-Agent Conversations. In *Proceedings of the International Conference on Intelligent User Interfaces*; ACM: New York, NY, USA, 2010; pp. 139–148. [CrossRef]
20. Erickson, A.; Norouzi, N.; Kim, K.; LaViola, J.J.; Bruder, G.; Welch, G.F. Effects of Depth Information on Visual Target Identification Task Performance in Shared Gaze Environments. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 1934–1944. [CrossRef] [PubMed]
21. Vertegaal, R.; Slagter, R.; van der Veer, G.; Nijholt, A. Eye Gaze Patterns in Conversations: There Is More to Conversational Agents than Meets the Eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2001; pp. 301–308. [CrossRef]
22. Andrist, S.; Gleicher, M.; Mutlu, B. Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2017; pp. 2571–2582. [CrossRef]
23. Reddy, G.S.R.; Proulx, M.J.; Hirshfield, L.M.; Ries, A.J. Towards an Eye-Brain-Computer Interface: Combining Gaze with the Stimulus-Preceding Negativity for Target Selections in XR. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2024; pp. 1–17. [CrossRef]
24. Chiossi, F.; Imamaliyev, E.; Bleichner, M.; Mayer, S. Anticipation Before Action: EEG-Based Implicit Intent Detection for Adaptive Gaze Interaction in Mixed Reality. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2026; pp. 1–24. [CrossRef]
25. Yarbus, A.L. *Eye Movements and Vision*; Plenum Press: New York, NY, USA, 1967. [CrossRef]
26. Rayner, K. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychol. Bull.* **1998**, *124*, 372–422. [CrossRef] [PubMed]
27. Duchowski, A.T. *Eye Tracking Methodology: Theory and Practice*; Springer: Berlin/Heidelberg, Germany, 2007. [CrossRef]
28. Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; van de Weijer, J. *Eye Tracking: A Comprehensive Guide to Methods and Measures*; Oxford University Press: Oxford, UK, 2011.
29. Kanan, C.; Ray, N.A.; Bseiso, D.N.; Hsiao, J.H.; Cottrell, G.W. Predicting an Observer's Task Using Multi-Fixation Pattern Analysis. In *Proceedings of the Symposium on Eye Tracking Research and Applications*; ACM: New York, NY, USA, 2014; pp. 287–290. [CrossRef]
30. Coutrot, A.; Hsiao, J.H.; Chan, A.B. Scanpath Modeling and Classification with Hidden Markov Models. *Behav. Res. Methods* **2018**, *50*, 362–379. [CrossRef] [PubMed]
31. Fathi, A.; Li, Y.; Rehg, J.M. Learning to Recognize Daily Actions Using Gaze. In *Proceedings of the Computer Vision–ECCV 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 314–327. [CrossRef]
32. Boisvert, J.F.G.; Bruce, N.D.B. Predicting Task from Eye Movements: On the Importance of Spatial Distribution, Dynamics, and Image Features. *Neurocomputing* **2016**, *207*, 653–668. [CrossRef]
33. Borji, A.; Lennartz, A.; Pomplun, M. What Do Eyes Reveal About the Mind?: Algorithmic Inference of Search Targets from Fixations. *Neurocomputing* **2015**, *149*, 788–799. [CrossRef]
34. Haji-Abolhassani, A.; Clark, J.J. An Inverse Yarbus Process: Predicting Observers' Task from Eye Movement Patterns. *Vis. Res.* **2014**, *103*, 127–142. [CrossRef] [PubMed]
35. Yin, Z.; Wan, Z.; Yang, M.; Xiong, Y.; Wang, W.; Wu, S. Robust Gaze-Based Intention Prediction for Real-World Scenarios. *IEEE Trans. Cogn. Dev. Syst.* **2025**, *17*, 835–846. [CrossRef]
36. Krausz, N.E.; Lamotte, D.; Batzianoulis, I.; Hargrove, L.J.; Micera, S.; Billard, A. Intent Prediction Based on Biomechanical Coordination of EMG and Vision-Filtered Gaze for End-Point Control of an Arm Prosthesis. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2020**, *28*, 1471–1480. [CrossRef] [PubMed]
37. Zunino, A.; Cavazza, J.; Koul, A.; Cavallo, A.; Becchio, C.; Murino, V. Predicting Human Intentions from Motion Cues Only: A 2D+3D Fusion Approach. In *Proceedings of the 25th ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2017; pp. 591–599. [CrossRef]
38. Xiong, Y.; Varadarajan, B.; Wu, L.; Xiang, X.; Xiao, F.; Zhu, C.; Dai, X.; Wang, D.; Sun, F.; Iandola, F.; et al. EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 16111–16121.
39. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 16901–16911. [CrossRef]

40. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
41. Wu, J.; Antonova, R.; Kan, A.; Lepert, M.; Zeng, A.; Song, S.; Bohg, J.; Rusinkiewicz, S.; Funkhouser, T. TidyBot: Personalized Robot Assistance with Large Language Models. *Auton. Robot.* **2023**, *47*, 1087–1102. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.