

Plug-and-Play Clarifier: A Zero-Shot Multimodal Framework for Egocentric Intent Disambiguation

Sicheng Yang¹, Yukai Huang², Weitong Cai³, Shitong Sun³, You He^{1*},
Jiankang Deng⁴, Hang Zhang², Jifei Song⁵, Zhensong Zhang²

¹Shenzhen International Graduate School, Tsinghua University ²Independent Researcher

³Queen Mary University of London ⁴Imperial College London ⁵University of Surrey
yangsc25@mails.tsinghua.edu.cn, heyou@mail.tsinghua.edu.cn

*Corresponding author

Abstract

The performance of egocentric AI agents is fundamentally limited by multimodal intent ambiguity. This challenge arises from a combination of underspecified language, imperfect visual data, and deictic gestures, which frequently leads to task failure. Existing monolithic Vision-Language Models (VLMs) struggle to resolve these multimodal ambiguous inputs, often failing silently or hallucinating responses. To address these ambiguities, we introduce the **Plug-and-Play Clarifier**, a zero-shot and modular framework that decomposes the problem into discrete, solvable sub-tasks. Specifically, our framework consists of three synergistic modules: (1) a text clarifier that uses dialogue-driven reasoning to interactively disambiguate linguistic intent, (2) a vision clarifier that delivers real-time guidance feedback, instructing users to adjust their positioning for improved capture quality, and (3) a cross-modal clarifier with grounding mechanism that robustly interprets 3D pointing gestures and identifies the specific objects users are pointing to. Extensive experiments demonstrate that our framework improves the intent clarification performance of small language models (4–8B) by approximately 30%, making them competitive with significantly larger counterparts. We also observe consistent gains when applying our framework to these larger models. Furthermore, our vision clarifier increases corrective guidance accuracy by over 20%, and our cross-modal clarifier improves semantic answer accuracy for referential grounding by 5%. Overall, our method provides a plug-and-play framework that effectively resolves multimodal ambiguity and significantly enhances user experience in egocentric interaction.

Introduction

Egocentric AI agents, particularly those embedded in wearable devices such as AI glasses, are emerging as a new and significant area of human-computer interaction. The goal is to develop an always-on cognitive partner that perceives the world from the user’s first-person perspective. Such an agent could understand user goals and provide seamless assistance for daily physical tasks, including assembling furniture, cooking complex recipes, or navigating unfamiliar environments. By operating from this viewpoint, these agents can deliver contextually-rich, proactive support that is tightly integrated with the user’s activities (Li et al. 2025a).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

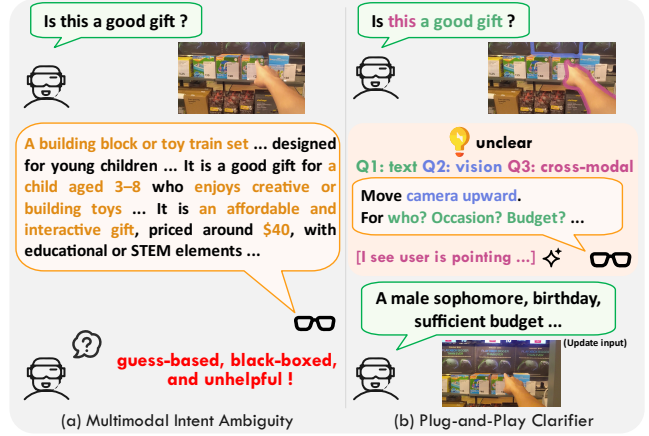


Figure 1: Our Clarifier framework resolves the multimodal ambiguous query, “Is this a good gift?” (a) Multimodal Intent Ambiguity: A standard AI defaults to a guess, making assumptions about the recipient (e.g., a child), their interests, and the user’s budget. This “black box” approach is unhelpful because if the assumptions are wrong, the recommendation is useless. (b) Plug-and-Play Clarifier: Our system avoids guessing. It first identifies that key information (recipient, occasion, budget), missing visual context, and pointing gestures between modalities. It then proactively asks clarifying questions and provides camera feedback (“For who? Move camera upward...”). Once the user provides the necessary context, the system can deliver a relevant and genuinely helpful recommendation.

However, this vision faces a fundamental challenge: **multimodal intent ambiguity**. Unlike interactions with text-based chatbots, egocentric interaction is inherently noisy, fast-paced, and underspecified. A simple spoken command, such as “What about that one?”, is inherently ambiguous. Which object is “that one”? Is the object of interest clearly visible, or is it blurry or partially occluded in the camera feed? This ambiguity stems from a combination of sources: underspecified natural language, imperfect visual data from the wearable camera, and deictic gestures like pointing (Huang et al. 2016). As a result, even state-of-the-art AI assistants frequently misinterpret user intent, leading to frus-

trating and unproductive task failures (Gabriel et al. 2024; Weidemann and Rußwinkel 2021).

Current approaches, dominated by monolithic end-to-end Vision-Language Models (VLMs), fall short in robustly addressing this challenge (Huang et al. 2025). While large models like GPT-4o demonstrate strong general multimodal understanding, they function as opaque “black boxes.” When presented with ambiguous, mixed-signal inputs, they tend to “hallucinate” an interpretation or fail silently, as they lack a mechanism to proactively seek clarification. Requiring a single, massive model to concurrently manage linguistic interpretation, spatial reasoning, and visual quality assessment is inherently brittle. This approach is also computationally expensive, making it a poor fit for resource-constrained wearable devices.

To overcome these limitations, we propose a shift from monolithic reasoning to a structured, modular, and interactive approach. We introduce the **Plug-and-Play Clarifier**, a zero-shot, multimodal framework that resolves ambiguity in egocentric interactions. Instead of relying on a single model’s opaque reasoning, our framework decomposes the problem into discrete sub-tasks managed by an explicit, programmatic control loop. This architecture integrates three core modules: (1) a *text clarifier* with dialogue-driven reasoning that clarifies user intent through a structured, step-by-step conversation; (2) a *vision clarifier* that assesses visual input quality (e.g., framing, clarity) and provides real-time corrective feedback; and (3) a *cross-modal clarifier* with grounding mechanism that interprets 3D pointing gestures by casting a geometric ray into the scene to localize the referenced object. Our framework enhances existing foundation models without requiring fine-tuning. This modular approach proves more robust and efficient than end-to-end black-box models for complex real-world tasks. Our code and demos are available online¹. We summarize our contributions as follows: (1) We propose a zero-shot and plug-and-play framework that resolves multimodal intent ambiguity in egocentric interaction through problem decomposition and interactive clarification. (2) We show that our framework improves the intent clarification accuracy of small language models (4–8B) by 30% on textual tasks, making them competitive with much larger models. (3) On our newly introduced VRA-Ego benchmark, we demonstrate that our individual modules achieve significant improvements: the vision clarifier increases corrective guidance accuracy by over 20%, and the cross-modal clarifier with 3D pointing module improves semantic grounding accuracy by 5%, outperforming strong monolithic baselines. (4) We validate that a hybrid architecture combining the generative capabilities of LLMs with algorithmic modules is a more robust, efficient, and interpretable approach for building reliable egocentric AI.

Related Work

Intent Clarification in Dialogue Systems

The field of intent clarification has moved from traditional approaches, such as programmatic slot-filling and templated

questions (Zhang et al. 2025a,b; Li, Wu et al. 2025), towards end-to-end systems that use Large Language Models (LLMs) (Zhang et al. 2024a; Qian et al. 2024). To address the inconsistent performance of basic prompting methods like Chain-of-Thought (CoT), subsequent work has focused on incorporating more structure. For example, some methods require the LLM to first determine the type of ambiguity before generating a response (Tang, Soulier, and Guigue 2025). Others involve developing system-level frameworks that use specialized classifiers (Tanjim, Chen et al. 2025) or Bayesian inference (Wen et al. 2025) to guide the clarification process. Recent benchmarks (Li et al. 2025b) and datasets (Aliannejadi et al. 2021) support a move towards more structured and interpretable disambiguation, yielding specialized frameworks for enterprise (Murzaku et al. 2025) and knowledge-graph applications.

Despite these advances, existing work largely falls into two categories. The first is end-to-end LLM solutions, which can lack transparency and rely heavily on the model’s own reasoning capabilities (Nguyen et al. 2025; de Carvalho Souza, Souza, and Weigang 2025). The second is applications designed for text-only scenarios, such as travel planning (Zhang et al. 2024b; Wang, Ning et al. 2025) or e-commerce (Dammu, Alonso, and Poblete 2025). In contrast, we introduce a hybrid approach that uses a programmed, zero-shot external control loop to guide the clarification process. This architecture provides transparency and control by design, while reducing the reasoning load placed on the LLM. Furthermore, we are the first to apply such a framework to the challenging domain of multimodal, first-person interactions in the physical world, extending intent clarification beyond traditional text-based interfaces.

Egocentric Vision and Interaction

Egocentric Vision (EGV) provides a first-person perspective to infer user intent from tasks such as action recognition (Shiota et al. 2024), hand-object interaction (Xu et al. 2023, 2025), and scene understanding (Li et al. 2025a). This has led to large-scale projects like EgoLife and EgoM2P, which aim to build assistive agents from complex, real-world multimodal data (Tu et al. 2025). However, a key challenge in EGV is that its data is inherently noisy and ambiguous, unlike the curated data used in standard VQA or VLM benchmarks (Fan 2019; Perrett et al. 2025). Even state-of-the-art models like Gemini Pro struggle with it, revealing the limitations of current VLMs when processing uncurated, real-world egocentric video (Google DeepMind 2025).

Pointing gestures are a critical signal in Egocentric Vision (EGV) (Das 2021). While modern systems integrate gestures with language for 3D scene understanding (Mane et al. 2025), many earlier approaches were limited. For instance, some work relied only on the 2D screen position of a fingertip (Huang et al. 2015, 2016). This simplification ignores the 3D pointing vector, which is essential for determining what a person is referring to. As a result, the ambiguity inherent in the pointing action itself remained largely unaddressed.

Most existing research passively attempts to interpret noisy and ambiguous inputs, with a recent trend of proactive agents only just beginning to emerge (Lu et al. 2025b).

¹<https://github.com/YoungSeng/plug-and-play-clarifier>

In contrast, we propose a proactive interaction framework. Rather than analyzing potentially flawed data after the fact, our system actively identifies input ambiguity—such as an imprecise pointing gesture—and uses real-time multimodal feedback to guide the user toward providing a clearer, more reliable input. We argue this shift from passive processing to active guidance is a key step toward more robust and intuitive first-person interactive systems (Zhang et al. 2025c).

Multimodal Reasoning with Large Models

While state-of-the-art Vision-Language Models (VLMs) like GPT-4o (Hurst et al. 2024), Gemini 2.5 Pro (Google DeepMind 2025), and Grok-3 (xAI 2025) perform well on general multimodal tasks (Bai et al. 2025), they are known to be unreliable for tasks requiring precise spatial or geometric reasoning, often leading to hallucinations (Mouselinos, Michalewski et al. 2024; Feng, Denny et al. 2024; Huang et al. 2025; Ramachandran et al. 2025). One line of work addresses this by improving the monolithic models themselves, for instance, through specialized pre-training for high-resolution visuals (DeepSeek-AI et al. 2024), incorporating fine-grained object grounding (Bai et al. 2025), or using native multimodal pre-training methods (Zhu et al. 2025). In contrast, our work follows an alternative approach that builds hybrid systems. These systems decompose problems to combine the semantic understanding of LLMs with the precision of specialized algorithms (Wu et al. 2024; Sharma et al. 2025; Patil 2025).

Our work differs from prior efforts by implementing a hybrid design as an interactive, iterative clarification loop. Unlike approaches that attempt to generate all clarification questions in a single turn, our method refines its understanding through a step-by-step dialogue to resolve ambiguity. This iterative process is not only more efficient but, crucially, allows smaller, resource-constrained models to handle complex multimodal tasks that would otherwise be beyond their capabilities.

Methodology

Text-based Intent Clarification

We propose a zero-shot, **dialogue-driven reasoner** for resolving ambiguous text intents. Our approach extends Chain-of-Thought (CoT) prompting (Zou et al. 2024; Yin, Hwang et al. 2025; Lu et al. 2025a) by guiding a Large Language Model (LLM) to iteratively clarify a user’s goal through conversation. This process relies entirely on in-context learning and does not require any task-specific fine-tuning (Zhang et al. 2025a; Lu et al. 2025a).

Our method operates iteratively. In each turn t , given the initial user request U_0 and the conversation history H_t , the LLM first analyzes the user’s intent to identify known (K_t) and missing (M_t) pieces of information:

$$(K_t, M_t) = \text{Analyze}(U_0, H_t) \quad (1)$$

To maintain an efficient dialogue, the model then assigns a priority $p(m)$ (e.g., critical, important) to each missing item $m \in M_t$ and selects the highest-priority item $m_t^* =$

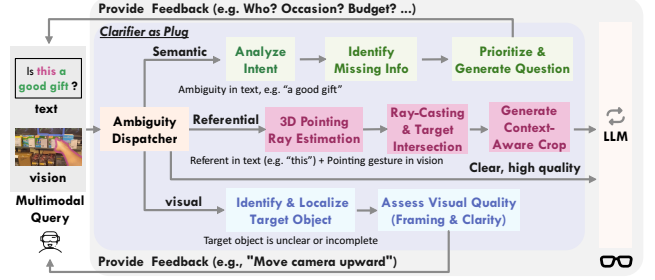


Figure 2: An overview of our clarification pipeline, a plug-in module for resolving ambiguous multimodal queries. The pipeline identifies and addresses three types of underspecification: (1) semantic ambiguity in language (e.g., “a good gift”) is clarified through dialogue; (2) visual ambiguity from unclear object views is handled by requesting a better view; and (3) referential ambiguity from pointing gestures (e.g., “this”) is improved by adaptive image cropping.

$\arg \max_{m \in M_t} p(m)$ to ask about next. The LLM generates a question Q_t based on m_t^* . The user’s answer, A_t , is added to the history ($H_{t+1} = H_t \cup \{(Q_t, A_t)\}$), and the process repeats. The loop terminates when no high-priority information is missing from M_t . Finally, the LLM uses the complete history H_{final} to generate a structured and actionable summary of the user’s intent. The entire process is driven by structured in-context prompts, requiring no updates to the LLM’s parameters (Kojima et al. 2022).

Vision-based Intent Clarification

To address the visual ambiguity inherent in first-person interactions, our framework proposes a **vision clarifier**. As shown in Figure 3, this module is designed to identify the user’s intended visual target and verify its image quality before downstream processing.

The process begins when an VLM parses the query U_v to extract the target object’s class label c in a zero-shot manner:

$$c = \text{ExtractEntity}(U_v) \quad (2)$$

This label c guides an open-set object detector (Liu et al. 2024) to find the object in the image frame I , producing a bounding box B . If the detector fails to find the object, the system prompts the user to point the camera at the target.

Once the object is localized, the system evaluates the quality of the region of interest (ROI) inside B . First, it checks for proper *framing*. The object’s relative area must be within a predefined range $[\tau_{\text{small}}, \tau_{\text{large}}]$, and its bounding box B must not be clipped by the image edges (margin δ_{edge}). Second, it evaluates *image clarity* using a score that combines two metrics: the variance of the Laplacian (C_{lap}) to detect focus blur (Bansal, Raj, and Choudhury 2016) and the FFT high-frequency energy ratio (C_{fft}) to detect motion blur (Shi, Xu, and Jia 2014). The final clarity score is a weighted sum of the normalized values:

$$\mathcal{S}_{\text{clarity}}(I_B) = w_{\text{lap}} \cdot \text{Norm}(C_{\text{lap}}(I_B)) + w_{\text{fft}} \cdot \text{Norm}(C_{\text{fft}}(I_B)) \quad (3)$$

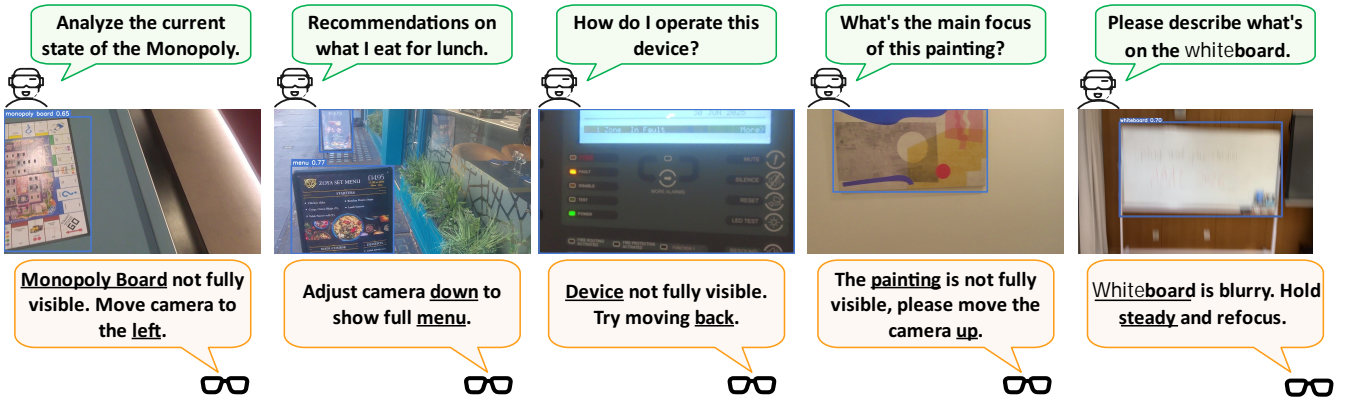


Figure 3: Overview of our vision-based clarification module. Given a user’s query about a physical object, the system first identifies the target class (e.g., “menu”) using an VLM. An open-set detector then localizes the object in the image frame. Subsequently, the visual quality is assessed for framing integrity and clarity. If issues like improper framing or blurriness are detected, the system provides real-time corrective feedback to the user, ensuring high-quality visual input before proceeding.

If the framing is poor or S_{clarity} is below the threshold τ_{blur} , the system gives the user specific, corrective feedback (e.g., “Move further away”, “Hold steady”). This feedback loop repeats until a well-framed, clear image is obtained, improving the reliability of subsequent vision-based tasks.

Cross-Modal Referential Clarification

To resolve referential ambiguity from pointing gestures (e.g., “that object”), our framework introduces a **cross-modal grounding mechanism** that converts the user’s pointing direction into a 3D ray to identify specific objects in the scene. This multi-stage pipeline, depicted in Figure 4, ensures robust grounding from a single egocentric image.

First, we estimate the 3D pointing ray (Nakamura et al. 2023). From a single image I , we concurrently generate a hand segmentation (Wang et al. 2025) mask M_{hand} (via pose estimation (Khanam et al. 2024; Afifi 2019)) and a dense depth map D (via monocular depth estimation (Yang et al. 2024b)). The pointing vector is derived from two keypoints on the contour of M_{hand} : the fingertip (p_{tip}^{2D}) and base (p_{base}^{2D}). To improve robustness against ambiguous contours, we refine the location of p_{tip}^{2D} by analyzing depth gradients along the finger’s axis, ensuring it lies on the foreground finger. These 2D points are then unprojected to 3D using the depth map D . The final pointing ray originates at p_{base}^{3D} with a normalized direction vector \vec{v} :

$$\vec{v} = \frac{p_{\text{tip}}^{3D} - p_{\text{base}}^{3D}}{\|p_{\text{tip}}^{3D} - p_{\text{base}}^{3D}\|} \quad (4)$$

Next, we perform target localization via ray-casting. We cast the ray into the 3D scene to find the intersection point $P_{\text{intersect}}$. This point is determined by finding the point p along the ray whose depth, $\text{depth}(p)$, most closely matches the value in the scene’s depth map D at the corresponding 2D projection of p , denoted as p_{xy} . This process is formulated as:

$$P_{\text{intersect}} = \arg \min_{p \in \text{Ray}(p_{\text{base}}^{3D}, \vec{v})} |\text{depth}(p) - D(p_{xy})| \quad (5)$$

subject to $|\text{depth}(p) - D(p_{xy})| \leq \tau_{\text{collision}}$.

Finally, we conduct object identification and context-aware cropping. Since simple point-based grounding is often brittle in cluttered scenes, our approach generates a depth-aware Region of Interest (ROI). We define a bounding box B_{target} centered on the 2D projection of $P_{\text{intersect}}$, with dimensions dynamically scaled by its depth. This adapts the ROI size to the object’s distance. To resolve the user’s query U_v while preserving deictic context, we perform a context-aware crop: we compute a consolidated bounding box, B_{context} , that minimally encloses both the target ROI B_{target} and the user’s hand. The resulting crop $I[B_{\text{context}}]$ is passed with the query U_v to the VLM, mitigating background noise while retaining the crucial gesture-object link. As validated by our ablation studies (see supplementary material), this method is more robust than direct grounding or simple visual overlays.

Experiments

Experimental Setup

Dataset. Our evaluation leverages established benchmarks for textual disambiguation and introduces a novel benchmark to address ambiguities unique to first-person vision. For textual tasks, we use IN3 (Qian et al. 2024) and CLAMBER (Zhang et al. 2024a), with the hierarchical attributes in IN3 being crucial for assessing our model’s prioritization mechanism. To address the lack of targeted evaluation for visuospatial ambiguity, we introduce VRA-Ego (Visual and Referential Ambiguity in Egocentric view), a new benchmark of 1000 samples captured with modern AR glasses (e.g., Ray-Ban Meta (2024), RayNeo X2/X3 Pro (2025)). VRA-Ego is composed of two purpose-built subsets:

- **Visual Ambiguity Set (500 images):** This subset features intentionally flawed visual data (e.g., blur, poor framing). The ground truth consists of the precise corrective guidance needed to resolve the issue.
- **Referential Ambiguity Set (500 samples):** This subset focuses on grounding deictic gestures, containing pointing

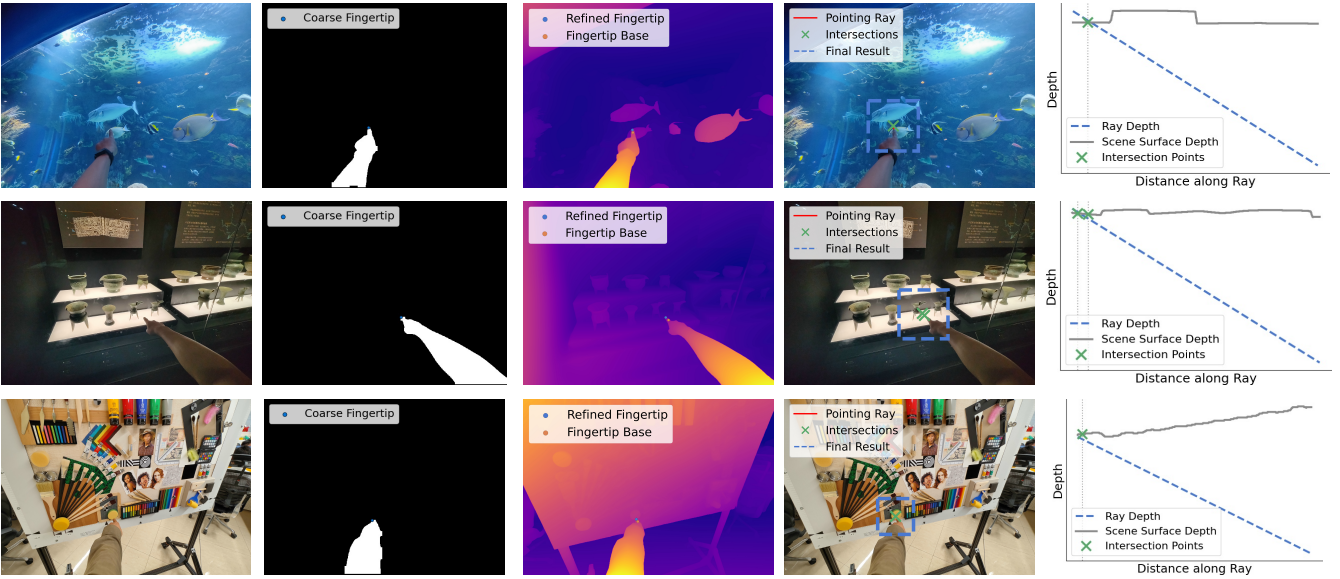


Figure 4: Our multi-stage pipeline for resolving cross-modal referential ambiguity. From a single image, we (1) estimate a 3D pointing ray from the user’s hand gesture, (2) cast this ray into the scene to find a 3D intersection point, and (3) identify the target object and generate a context-aware crop containing both the hand and the object, which is then passed to a VLM for final interpretation.

actions, ambiguous queries, and their corresponding annotated answers.

The entire dataset is meticulously curated for diversity across scenes, lighting, object distances, and hand-object configurations to ensure real-world robustness.

Evaluation Metrics. To evaluate our framework, we adopt a multi-faceted approach that assesses both component-level success and the semantic quality of the final output. For textual disambiguation, we first measure initial Vagueness Judgement Accuracy and dialogue efficiency through Average Conversation Rounds. The core performance is then captured by a Missing Details Recover Rate, calculated via a novel automated pipeline that simulates interaction to test the recovery of critical attributes. This principle of evaluating nuanced recovery extends to the vision module, where we assess not only the initial Target Identification Accuracy but also the quality of corrective feedback using a Strict and a Loose Recover Rate—the latter crediting partially correct suggestions (e.g., “left” for “top-left”). Ultimately, for cross-modal clarification, we evaluate pipeline applicability with Pointing Success Accuracy and measure the final output quality with a Semantic Answer Recover Rate. Here, to transcend the limitations of simple string matching and assess true comprehension, we employ a LLM judge to score the semantic similarity between the generated answer and the ground truth.

Comparison to Existing Methods

Choice of Baseline Model and Implementation Details. Our framework is designed as a zero-shot, external programmatic loop that enhances existing foundation models. To isolate and validate its benefits, we benchmark it against

strong monolithic baselines across three core clarification tasks. Our method is evaluated by augmenting a diverse set of foundation models (Qwen, Llama, GPT-4o, Gemini, etc.), creating enhanced versions we denote with a -Clarifier suffix. The experimental comparisons are as follows:

- **Textual Disambiguation:** We compare our iterative, multi-step clarification process against a standard monolithic prompt baseline. The baseline performs all reasoning steps (analysis, question generation) within a single, complex prompt, whereas our method guides the LLM through a sequence of targeted queries.
- **Visual Object Grounding:** The baseline VLM receives only the user query and image. In contrast, the VLM-Clarifier is augmented with our module that first invokes an open-set object detector and then performs a deterministic quality assessment on the resulting bounding box before proceeding.
- **Cross-Modal Referential Resolution:** The baseline VLM processes the raw image and query directly. The corresponding VLM-Clarifier is enhanced by our multi-stage pipeline that integrates depth estimation, hand segmentation, and 3D ray-casting to robustly interpret pointing gestures before feeding the identified target to the VLM.

For additional implementation details and baseline analyses (GPT-4 (OpenAI 2023), Mistral variants (Jiang et al. 2023; Qian et al. 2024)), please refer to the Appendix.

Quantitative Analysis. Our quantitative evaluation shows that our modular and programmatic framework achieves significant performance improvements on textual, visual, and cross-modal clarification tasks. These improvements are driven by the framework’s ability to either scaffold the rea-

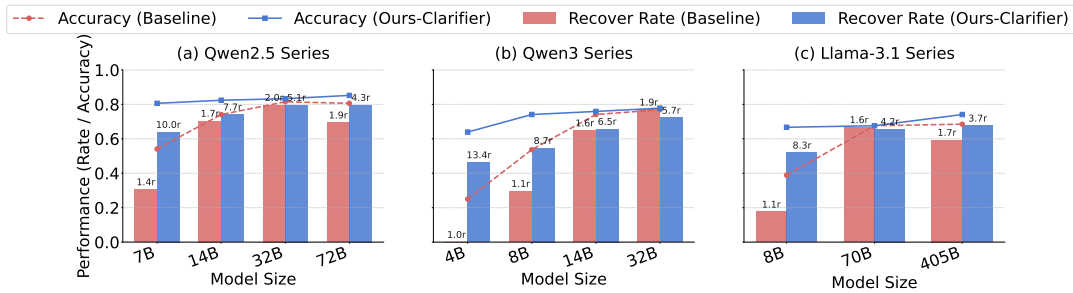


Figure 5: A comparative analysis of our proposed framework against the baseline across three open-source LLM families: (a) Qwen2.5, (b) Qwen3, and (c) Llama-3.1. The evaluation spans a range of model sizes to assess performance scalability. The primary performance metric, Recover Rate (representing the recovery of critical missing details), is shown using bar charts for both our method (blue) and the baseline (red). Additionally, the Accuracy (Baseline) is plotted as a dashed red line, while the Average Conversation Rounds (denoted by 'r' values) are annotated above each bar to measure dialogue efficiency.

Model	Parameter Size	Accuracy (%) \uparrow	
		w/o. Clarifier	w. Clarifier
Qwen2.5 (Yang et al. 2024a)	7B	24.4	53.0 (+ 28.6)
	14B	56.0	61.8 (+ 5.8)
	32B	56.0	66.0 (+ 10.0)
	72B	54.2	65.4 (+ 11.2)
LLama-3.1 (Dubey et al. 2024)	8B	25.9	52.9 (+ 27)
	70B	53.5	59.3 (+ 5.8)
	405B	54.9	60.1 (+ 5.2)

Table 1: Vagueness Judgement Accuracy on the CLAMBER benchmark. Our iterative Clarifier framework substantially boosts performance over single-prompt baselines. The framework yields an approximate 30% increase for smaller models like Llama-3.1-8B and Qwen2.5-7B, elevating them to be competitive with much larger counterparts. This demonstrates that our structured, step-by-step approach acts as a critical reasoning scaffold, enabling smaller models to handle complex clarification tasks reliably.

soning of smaller models or to incorporate deterministic algorithms for tasks where Vision-Language Models (VLMs) are known to perform poorly.

The benefit of our iterative framework is most pronounced for smaller language models. As shown in Figure 5, our method improves the Recover Rate for smaller models like Qwen3-4B and Qwen2.5-7B by 43% and 32% respectively. This is because our step-by-step process decomposes a complex task into manageable sub-problems, which often overwhelms the single-turn capabilities of these models. While still beneficial, the performance margin narrows for larger models (e.g., Llama-3.1-405B) that are already proficient with complex, monolithic prompts. This trend is corroborated on the CLAMBER dataset (Table 1), where our framework elevates the vagueness judgment accuracy of smaller models by nearly 30%, making them competitive with much larger counterparts.

Our framework also shows strong performance on visuospatial tasks that challenge the geometric reasoning capabilities of end-to-end VLMs. For example, baseline models can identify visual referents but fail to provide precise spa-

tial corrective feedback, achieving low Strict Recover Rates for directional guidance (31.5%–46.2%) (Table 2). To overcome this, our -Clarifier module performs deterministic geometric analysis on the object’s bounding box. This targeted approach improves the Strict Recover Rate by a large margin of 11.9% to 20.7%.

Similarly, monolithic VLMs struggle to interpret pointing gestures from raw images, with the top baseline reaching a Recover Rate of only 67.3% (Table 2). Our framework addresses this limitation by employing a two-step process: it first detects the pointing intent with high accuracy (87.2%–95.1%) before engaging a specialized 3D ray-casting pipeline. This explicit geometric modeling leads to consistent gains in task success, improving the Recover Rate by 3.1% to 6.6% for all evaluated VLMs. For both visuospatial tasks, this compositional strategy allows our framework to surpass the performance of purely learned systems, resulting in a more robust solution.

Qualitative Results. Our qualitative analysis substantiates the quantitative results, demonstrating the framework’s practical robustness and efficiency. The core of our approach is an iterative, step-by-step process that systematically reduces ambiguity across modalities.

In textual dialogues, this manifests as significantly lower inference latency. By using targeted, minimal prompts each turn rather than a single monolithic one, complex disambiguation is resolved efficiently. This structured methodology proves equally effective in vision-centric tasks (Figure 3). Our framework reliably analyzes complex scenes like chessboards or fine-print menus—scenarios where monolithic baselines often fail due to unresolved visual ambiguity.

The framework’s advantage is most pronounced in resolving cross-modal pointing in cluttered environments (Figure 4). While baselines are often distracted by salient but incorrect objects, our method robustly grounds the user’s deictic reference. It first reconstructs the 3D pointing ray, intersects it with the scene using the depth map, and then generates a context-aware crop that isolates the target while preserving the vital hand-object relationship. This focused input enables the VLM to succeed where it would otherwise fail. We note that the primary failure mode occurs when monocular

Model	(a) Vision-based Clarification			(b) Cross-Modal Pointing	
	Accuracy (%) ↑	Recover Rate (%) ↑		Accuracy (%) ↑	Recover Rate (%) ↑
		Strict	Loose		
gemini-2.5-pro (Google DeepMind 2025)	91.8	46.2	60.2	-	67.4
gemini-2.5-pro-Clarifier	95.4 (+ 3.6)	64.6 (+ 18.4)	75.8 (+ 15.6)	95.2	72.6 (+ 5.2)
GPT-4o (Hurst et al. 2024)	90.0	40.6	57.0	-	65.2
GPT-4o-Clarifier	92.4 (+ 2.4)	61.4 (+ 20.8)	73.6 (+ 16.6)	94.4	69.0 (+ 3.8)
Qwen2.5-VL (Bai et al. 2025)	86.6	35.4	53.2	-	57.8
Qwen2.5-VL-Clarifier	91.2 (+ 4.6)	47.4 (+ 12.0)	70.0 (+ 16.8)	92.4	64.4 (+ 6.6)
llava-v1.6 (Li et al. 2024)	84.8	36.6	41.6	-	53.6
llava-v1.6-Clarifier	89.8 (+ 5.0)	53.2 (+ 16.6)	52.8 (+ 11.2)	91.6	58.0 (+ 4.4)
InternVL 3.0 (Zhu et al. 2025)	83.2	35.6	59.6	-	51.6
InternVL 3.0-Clarifier	88.6 (+ 5.4)	50.2 (+ 14.6)	70.2 (+ 10.6)	93.6	57.8 (+ 6.2)
Llama 3.2 (Dubey et al. 2024)	82.6	35.6	51.6	-	50.2
Llama 3.2-Clarifier	85.4 (+ 2.8)	51.2 (+ 15.6)	64.6 (+ 13.0)	91.2	53.8 (+ 3.6)
MiniCPM-V (Yao et al. 2024)	81.6	31.6	48.8	-	47.0
MiniCPM-V-Clarifier	86.2 (+ 4.6)	43.6 (+ 12.0)	63.4 (+ 14.6)	88.2	52.0 (+ 5.0)
Molmo (Deitke et al. 2025)	80.0	37.6	48.6	-	48.4
Molmo-Clarifier	84.0 (+ 4.0)	52.8 (+ 15.2)	62.2 (+ 13.6)	87.2	51.6 (+ 3.2)

Table 2: Performance on Vision-based and Cross-Modal Pointing Clarification. The table compares baseline VLMs against the same models enhanced by our framework (suffixed with -Clarifier) on two egocentric tasks: (a) Evaluates corrective guidance for imperfect visual input. Here, Accuracy measures initial object identification, while Recover Rate (Strict/Loose) assesses the quality of the generated guidance. (b) Tests grounding of deictic queries via 3D pointing. Here, Accuracy (N/A for baselines) is our framework’s success in detecting pointing intent, and Recover Rate measures the final answer’s semantic accuracy.

depth estimation is inaccurate for thin or reflective surfaces, causing the ray to pass through the intended object. A detailed gallery of qualitative comparisons is in the Appendix.

Ablation Studies

We conducted ablation studies to validate our key design choices and assess the sensitivity of our modular framework.

- **Object Detector Robustness:** We evaluated the framework’s sensitivity to the choice of object detector. While substituting our default model with Florence-2 (Xiao et al. 2024) caused a small decrease in performance (1% Accuracy, 3% Recover Rate), other models like YOLOE (Wang et al. 2025) and YOLO-World (Cheng et al. 2024) showed slightly larger drops (2% Accuracy, 5% Recover Rate). Nevertheless, the performance with all alternative detectors remained substantially higher than the baseline without our clarifier. This demonstrates that our core contribution provides a consistent benefit, independent of the underlying detection model.
- **Importance of Specialized Fingertip Detection:** To measure the impact of our custom fingertip detector, we substituted it with the standard MediaPipe library (Lugaresi, Tang et al. 2019). This led to a significant decrease in Pointing Success Accuracy (15%). This is because MediaPipe is not robust to the challenging hand views—often occluded or low-resolution in egocentric data, which confirms the need for our tailored approach.
- **Effective VLM Input Grounding:** Finally, as detailed in the Appendix, our context-aware cropping strategy for grounding VLM input proved more effective than alternative methods like point-based segmentation or using

the full image (Guo, Wu et al. 2025), further justifying our specific design.

Collectively, these results highlight a key finding: while our framework provides a robust scaffold that is not overly sensitive to the choice of a general object detector, its peak performance critically relies on components specifically tailored to the challenges of egocentric vision, such as our specialized fingertip detector.

Conclusion

In this paper, we introduced a modular framework for resolving multimodal user intent ambiguity in egocentric vision. Our approach emulates a Chain-of-Thought process, decomposing complex and ambiguous queries into a sequence of simpler, verifiable sub-problems. These sub-problems are solved by specialized modules, including modules for textual analysis, visual quality assessment using a hybrid of a Vision-Language Model (VLM) and traditional algorithms, and 3D gesture grounding via ray-casting. A key result of our work is that this structured reasoning process significantly improves the performance of smaller language models (e.g., 7B models) on text intent disambiguation, making our approach practical for deployment on resource-constrained platforms, such as AR glasses. Our work shows that hybrid architectures, which combine the reasoning capabilities of large models with the precision of deterministic algorithms, present a promising direction for building more capable and reliable embodied AI. Future work will improve conversational efficiency and extend our approach to physically embodied agents that actively seek clarification.

Acknowledgments

This work was supported by the Shenzhen Science and Technology Program (Grant No. ZDSYS20220323112000-001).

References

- Affi, M. 2019. 11K Hands: Gender recognition and biometric identification using a large dataset of hand images. *Multim. Tools Appl.*, 78(15): 20835–20854.
- Aliannejadi, M.; et al. 2021. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, 4473–4484. ACL.
- Bai, S.; Chen, K.; Liu, X.; et al. 2025. Qwen2.5-VL Technical Report. *CoRR*, abs/2502.13923.
- Bansal, R.; Raj, G.; and Choudhury, T. 2016. Blur image detection using Laplacian operator and Open-CV. In *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, 63–67. IEEE.
- Cheng, T.; Song, L.; Ge, Y.; et al. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Seattle, WA, USA*, 16901–16911. IEEE.
- Dammu, P. P. S.; Alonso, O.; and Poblete, B. 2025. A Shopping Agent for Addressing Subjective Product Needs. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, 1032–1035. ACM.
- Das, S. S. 2021. A data-set and a method for pointing direction estimation from depth images for human-robot interaction and VR applications. In *IEEE International Conference on Robotics and Automation, ICRA*, 11485–11491. IEEE.
- de Carvalho Souza, M. E.; Souza, M. E. D. C.; and Weigang, L. 2025. Unveiling the Black Box: The Significance of XAI in Making LLMs Transparent. *Authorea Preprints*.
- DeepSeek-AI; Liu, A.; Feng, B.; et al. 2024. DeepSeek-V3 Technical Report. *CoRR*, abs/2412.19437.
- Deitke, M.; Clark, C.; Lee, S.; et al. 2025. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 91–104. IEEE/CVF.
- Dubey, A.; Jauhri, A.; Pandey, A.; et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.
- Fan, C. 2019. EgoVQA - An Egocentric Video Question Answering Benchmark Dataset. In *IEEE International Conference on Computer Vision Workshops*, 4359–4366. IEEE.
- Feng, T. H.; Denny, P.; et al. 2024. An Eye for an AI: Evaluating GPT-4o’s Visual Perception Skills and Geometric Reasoning Skills Using Computer Graphics Questions. In *SIGGRAPH Asia Educator’s Forum*, 5:1–5:8. ACM.
- Gabriel, I.; Manzini, A.; Keeling, G.; et al. 2024. The Ethics of Advanced AI Assistants. *CoRR*, abs/2404.16244.
- Google DeepMind. 2025. Gemini 2.5 Pro Preview Model Card. Technical report, Google. (preview release).
- Guo, D.; Wu, F.; et al. 2025. Seed1.5-VL Technical Report. *CoRR*, abs/2505.07062.
- Huang, K.; Qin, C.; Qiu, H.; et al. 2025. Why Vision Language Models Struggle with Visual Arithmetic? Towards Enhanced Chart and Geometry Understanding. In *Findings of the Association for Computational Linguistics, ACL, Vienna, Austria*, 4830–4843. ACL.
- Huang, Y.; Liu, X.; Jin, L.; and Zhang, X. 2015. DeepFinger: A Cascade Convolutional Neuron Network Approach to Finger Key Point Detection in Egocentric Vision with Mobile Camera. In *IEEE International Conference on Systems, Man, and Cybernetics*, 2944–2949. IEEE.
- Huang, Y.; Liu, X.; Zhang, X.; and Jin, L. 2016. A Pointing Gesture Based Egocentric Interaction System: Dataset, Approach and Application. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 370–377. IEEE.
- Hurst, A.; Lerer, A.; Goucher, A. P.; et al. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.
- Jiang, A. Q.; et al. 2023. Mistral 7B. *CoRR*, abs/2310.06825.
- Khanam, R.; et al. 2024. YOLOv11: An Overview of the Key Architectural Enhancements. *CoRR*, abs/2410.17725.
- Kojima, T.; Gu, S. S.; Reid, M.; et al. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in neural information processing systems*, volume 35, 22199–22213.
- Li, B.; et al. 2024. LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild. <https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/>.
- Li, X.; Wu, X.; et al. 2025. MDSD: Multi-Turn Diverse Synthetic Dialog Generation for Domain Specific Incomplete Requests Understanding. *SSRN Electronic Journal*, 1–9.
- Li, X.; et al. 2025a. Challenges and Trends in Egocentric Vision: A Survey. *CoRR*, abs/2503.15275.
- Li, Z.; Li, Y.; Xie, H.; and Qin, S. J. 2025b. CondAmbigQA: A Benchmark and Dataset for Conditional Ambiguous Question Answering. *CoRR*, abs/2502.01523.
- Liu, S.; Zeng, Z.; Ren, T.; et al. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *European conference on computer vision*, volume 15105 of *Lecture Notes*, 38–55. Springer.
- Lu, L.; Meng, C.; Ravenda, F.; et al. 2025a. Zero-Shot and Efficient Clarification Need Prediction in Conversational Search. In *European Conference on Information Retrieval*, volume 15572 of *Lecture Notes*, 389–404. Springer.
- Lu, Y.; Yang, S.; Qian, C.; et al. 2025b. Proactive Agent: Shifting LLM Agents from Reactive Responses to Active Assistance. In *The 13th International Conference on Learning Representations, ICLR, Singapore*. OpenReview.net.
- Lugaresi, C.; Tang, J.; et al. 2019. MediaPipe: A Framework for Building Perception Pipelines. *CoRR*, abs/1906.08172.
- Mane, A. M.; Weerakoon, D.; Subbaraju, V.; et al. 2025. Ges3ViG : Incorporating Pointing Gestures into Language-Based 3D Visual Grounding for Embodied Reference Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 9017–9026. CVF / IEEE.
- Meta, R.-B. 2024. <https://www.ray-ban.com/>.
- Mouselinos, S.; Michalewski, H.; et al. 2024. Beyond Lines and Circles: Unveiling the Geometric Reasoning Gap in

- Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*, 6192–6222. ACL.
- Murzaku, J.; et al. 2025. ECLAIR: Enhanced Clarification for Interactive Responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 28864–28870. AAAI Press.
- Nakamura, S.; Kawanishi, Y.; Nobuhara, S.; et al. 2023. DeePoint: Visual Pointing Recognition and Direction Estimation. In *IEEE/CVF International Conference on Computer Vision, ICCV, Paris, France*, 20520–20530. IEEE.
- Nguyen, G.; et al. 2025. Interpretable LLM-based Table Question Answering. *Trans. Mach. Learn. Res.*, 2025.
- OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.
- Patil, A. 2025. Advancing Reasoning in Large Language Models: Promising Methods and Approaches. *CoRR*, abs/2502.03671.
- Perrett, T.; Darkhalil, A.; Sinha, S.; et al. 2025. HD-EPIC: A Highly-Detailed Egocentric Video Dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Nashville, TN, USA*, 23901–23913. CVF / IEEE.
- Qian, C.; et al. 2024. Tell Me More! Towards Implicit User Intention Understanding of Language Model Driven Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1088–1113. ACL.
- Ramachandran, R.; Garjani, A.; Bachmann, R.; et al. 2025. How Well Does GPT-4o Understand Vision? Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks. *CoRR*, abs/2507.01955.
- RayNeo. 2025. <https://www.rayneo.com/>.
- Sharma, A.; et al. 2025. GeoCoder: Solving Geometry Problems by Generating Modular Code through Vision-Language Models. In *Findings of the Association for Computational Linguistics: NAACL*, 7340–7356. ACL.
- Shi, J.; Xu, L.; and Jia, J. 2014. Discriminative Blur Detection Features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2965–2972. IEEE.
- Shiota, T.; Takagi, M.; Kumagai, K.; et al. 2024. Egocentric Action Recognition by Capturing Hand-Object Contact and Object State. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV*, 6527–6537. IEEE.
- Tang, A.; Soulier, L.; and Guigue, V. 2025. Clarifying Ambiguities: on the Role of Ambiguity Types in Prompting Methods for Clarification Generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, 20–30. ACM.
- Tanjim, M. M.; Chen, X.; et al. 2025. Detecting Ambiguities to Guide Query Rewrite for Robust Conversations in Enterprise AI Assistants. *CoRR*, abs/2502.00537.
- Tu, Y.; Luo, H.; Chen, X.; et al. 2025. PlayerOne: Egocentric World Simulator. *CoRR*, abs/2506.09995.
- Wang, A.; Liu, L.; Chen, H.; et al. 2025. YOLOE: Real-Time Seeing Anything. *CoRR*, abs/2503.07465.
- Wang, J.; Ning, H.; et al. 2025. A Data Synthesis Method Driven by Large Language Models for Proactive Mining of Implicit User Intentions in Tourism. *CoRR*, abs/2505.11533.
- Weidemann, A.; and Rußwinkel, N. 2021. The role of frustration in human–robot interaction—what is needed for a successful collaboration? *Frontiers in psychology*, 12: 640186.
- Wen, L.; Xiong, G.; Mo, T.; et al. 2025. CLEAR-KGQA: Clarification-Enhanced Ambiguity Resolution for Knowledge Graph Question Answering. *CoRR*, abs/2504.09665.
- Wu, Z.; et al. 2024. Divide-or-Conquer? Which Part Should You Distill Your LLM? In *Findings of the Association for Computational Linguistics: EMNLP*, 2572–2585. ACL.
- xAI. 2025. Grok 3 Beta — The Age of Reasoning Agents. <https://x.ai/blog/grok-3>.
- Xiao, B.; Wu, H.; Xu, W.; et al. 2024. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Seattle, WA, USA*, 4818–4829. IEEE.
- Xu, B.; Wang, Z.; Du, Y.; et al. 2025. Do Egocentric Video-Language Models Truly Understand Hand-Object Interactions? In *The 13th International Conference on Learning Representations, ICLR, Singapore*. OpenReview.net.
- Xu, Y.; Li, Y.; Huang, Z.; et al. 2023. EgoPCA: A New Framework for Egocentric Hand-Object Interaction Understanding. In *IEEE/CVF International Conference on Computer Vision, ICCV, Paris, France*, 5250–5261. IEEE.
- Yang, A.; Yang, B.; Zhang, B.; et al. 2024a. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.
- Yang, L.; et al. 2024b. Depth Anything V2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.
- Yao, Y.; Yu, T.; Zhang, A.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *CoRR*, abs/2408.01800.
- Yin, Y.; Hwang, E.; et al. 2025. SWI: Speaking with Intent in Large Language Models. *CoRR*, abs/2503.21544.
- Zhang, G.; et al. 2025a. SummAct: Uncovering User Intentions Through Interactive Behaviour Summarisation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 265:1–265:17. ACM.
- Zhang, T.; Qin, P.; Deng, Y.; et al. 2024a. CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand*, 10746–10766. ACL.
- Zhang, X.; Deng, Y.; Ren, Z.; et al. 2024b. Ask-before-Plan: Proactive Language Agents for Real-World Planning. In *Findings of the Association for Computational Linguistics: EMNLP*, 10836–10863. ACL.
- Zhang, X.; Shen, Y.; Zheng, Z.; et al. 2025b. AskToAct: Enhancing LLMs Tool Use via Self-Correcting Clarification. *CoRR*, abs/2503.01940.
- Zhang, Y.; Dong, X. L.; Lin, Z.; et al. 2025c. Proactive Assistant Dialogue Generation from Streaming Egocentric Videos. *CoRR*, abs/2506.05904.
- Zhu, J.; Wang, W.; Chen, Z.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *CoRR*, abs/2504.10479.
- Zou, J.; Huang, J. X.; Ren, Z.; and Kanoulas, E. 2024. Learning to Ask: Conversational Product Search via Representation Learning. *CoRR*, abs/2411.14466.

Appendix

This document provides supplementary material for our AAAI 2026 submission, “Plug-and-Play Clarifier: A Zero-Shot Multimodal Framework for Egocentric Intent Disambiguation.” Here, we include extended methodological details, additional experimental results, in-depth analyses, and broader discussions that were omitted from the main paper due to space constraints. This material is intended to offer greater insight into our design choices, the robustness of our framework, and avenues for future research. For more details, please refer to the core code of our project.

Detailed Cross-Modal Referential Clarification

A fundamental challenge in egocentric interaction is the robust resolution of referential ambiguity, where a user’s verbal query containing deictic terms (e.g., “this”, “that”) must be precisely grounded to a specific physical object indicated by an accompanying pointing gesture. This necessitates effective visual grounding, which remains a non-trivial task, particularly in dynamic, unconstrained real-world environments.

Object Identification and Context-Aware Cropping for VLM Input As mentioned in the main paper, our pipeline accurately identifies the focal point of attention via 3D ray-casting, yielding a precise intersection point. Upon localizing this focus, the subsequent critical challenge is to appropriately delineate the visual entity at this location for effective processing by a Vision-Language Model. Early explorations into direct grounding methods, such as point-prompted segmentation (e.g., using SAM) or open-vocabulary object detection, proved brittle in egocentric settings. These approaches were particularly susceptible to inaccuracies stemming from cluttered scenes, minor pointing ray discrepancies, or the inherent challenges of novel object recognition.

Naively submitting the full egocentric image to the VLM introduces considerable visual noise and irrelevant contextual information, which can dilute the model’s focus. Conversely, a simplistic tight crop of only the putative object severs the crucial visual link between the user’s deictic gesture (e.g., the pointing hand) and the target object, thereby undermining the referential context. These limitations critically underscored the necessity for a more sophisticated input strategy, leading us to develop our context-aware cropping approach.

Ablation Study of VLM Input Grounding Strategies

As promised in the main paper, we conducted a comprehensive ablation study comparing our context-aware cropping against several alternative VLM input grounding strategies, visually summarized in Figure 6.

- **Full Image with Visual Cue** (Figure 6a and 6g): We provided the full egocentric image to the VLM, augmented with a visual cue (e.g., a green dot at the 2D projection, or a rendered bounding box (Guo, Wu et al. 2025)) to highlight the target. While performing marginally better than a completely naive full-image submission, this approach remained highly susceptible to distraction from

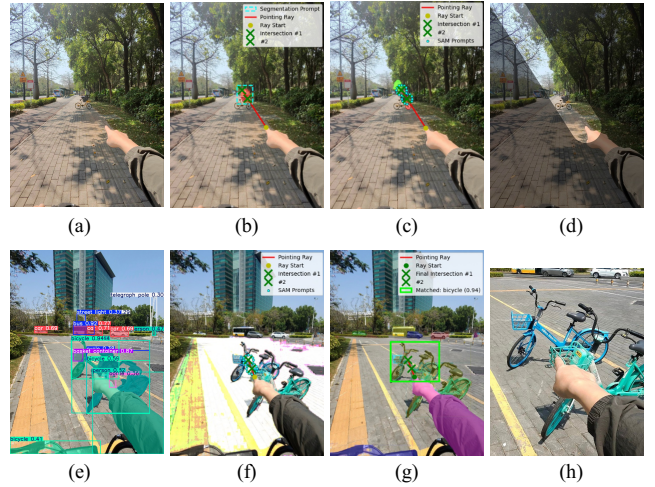


Figure 6: Various VLM input grounding strategies explored. (a) Original egocentric view. (b-f) Alternative rule-based and pre-processing strategies: (b) Intersection point connectivity graph, illustrating a graph-based segmentation approach. (c) Proximity-aware segmentation, where surrounding points are first gathered for mask generation. (d) Local attention derived from extended ray peripheral brightness, leveraging ambient light cues. (e) Recognition-first matching, where objects are identified before linking to the pointing gesture. (f) Segmentation-first matching, where scene segments are generated prior to association. (g) Bounding box rendered onto the original image. (h) **Our Adaptive Depth-Aware Contextual Cropping** method, which dynamically adjusts crop size based on object depth (further objects, larger crop) while ensuring the user’s hand/arm is consistently included, preserving critical deictic context.

other salient objects within the broad field of view, leading to a $\sim 4\%$ drop in Semantic Answer Recover Rate compared to our method. Furthermore, we critically observed that directly rendering bounding boxes (e.g., as depicted in Figure 6g) onto the full image, intended to explicitly demarcate the target, consistently yielded sub-optimal performance compared to actual cropping. Our hypothesis, supported by empirical observations, is that not all large language models (LLMs) within the VLM architecture possess the inherent capability to effectively process or interpret such graphical overlays within the image plane, potentially perceiving them as noise or distractive elements rather than meaningful cues.

- **Rule-Based Grounding Strategies** (Figures 6b-6f): Beyond simple point-prompted segmentation (Figure 6c), we explored a suite of rule-based grounding approaches:
 - **Intersection Connectivity Graph** (Figure 6b): This method involved inferring object boundaries based on the local connectivity and cluster density of pixels immediately surrounding.
 - **Proximity-Aware Segmentation** (Figure 6c): Utilizing the 2D projection as a seed for a segmentation model (e.g., SAM) to generate a mask, then cropping



Figure 7: Demonstration of our framework’s proactive visual clarification, a key differentiator from monolithic models. While baseline systems might fail or hallucinate when presented with flawed egocentric views, our method first diagnoses the input quality. It robustly identifies issues like occluded objects (*Monopoly Board*), distance-related ambiguity (*street sign*), and motion blur (*display screen*), providing users with explicit instructions for correction. By resolving visual ambiguity at the source, our framework establishes a reliable foundation for subsequent high-level reasoning.

the image to that mask. This was highly sensitive to the precision of the pointing ray and frequently failed on small or partially occluded objects, resulting in a $\sim 5\%$ performance drop.

- **Brightness-Weighted Local Context** (Figure 6d): This approach extended the pointing ray and leveraged the average brightness of the surrounding area as a heuristic for local attention, assuming the target object would exhibit a distinct brightness profile from its immediate background.
- **Recognition-First Matching** (Figure 6e): Here, a general object detection model was first employed to identify common objects across the scene, which were then post-hoc matched to the pointing ray.
- **Segmentation-First Matching** (Figure 6f): This involved generating generic segmentation masks for all discernible objects in the scene, followed by associating the pointing ray with the most plausible mask.

While offering varying degrees of operational simplicity, we collectively found that these purely rule-based and pre-processing strategies (Figures 6b-6f) were notably brittle and exhibited significant performance degra-

dation when encountering less common or novel objects, or when tasked with complex Question-Answering (QA) and reasoning queries that demanded a deeper, semantic understanding of object context beyond simple geometric heuristics. Their inherent reliance on predefined rules severely limited their generalizability and robustness in unconstrained egocentric scenarios.

- **Our method** (Figure 6h): Our proposed method adaptively determines the crop region based on the estimated depth of the target object. This ensures that objects further away are encompassed within a proportionally larger contextual window, providing sufficient surrounding information. Crucially, our strategy explicitly guarantees the consistent inclusion of the user’s hand and a portion of their arm within the crop. This dynamic and contextually rich cropping strategy preserves the critical deictic gesture-object relationship while providing the VLM with a focused, yet semantically rich, input. This approach consistently outperformed all alternative strategies, yielding a substantial 3-5% improvement in the final Semantic Answer Recover Rate over the next-best method. This robust empirical evidence unequivocally confirms that preserving the immediate hand-object in-

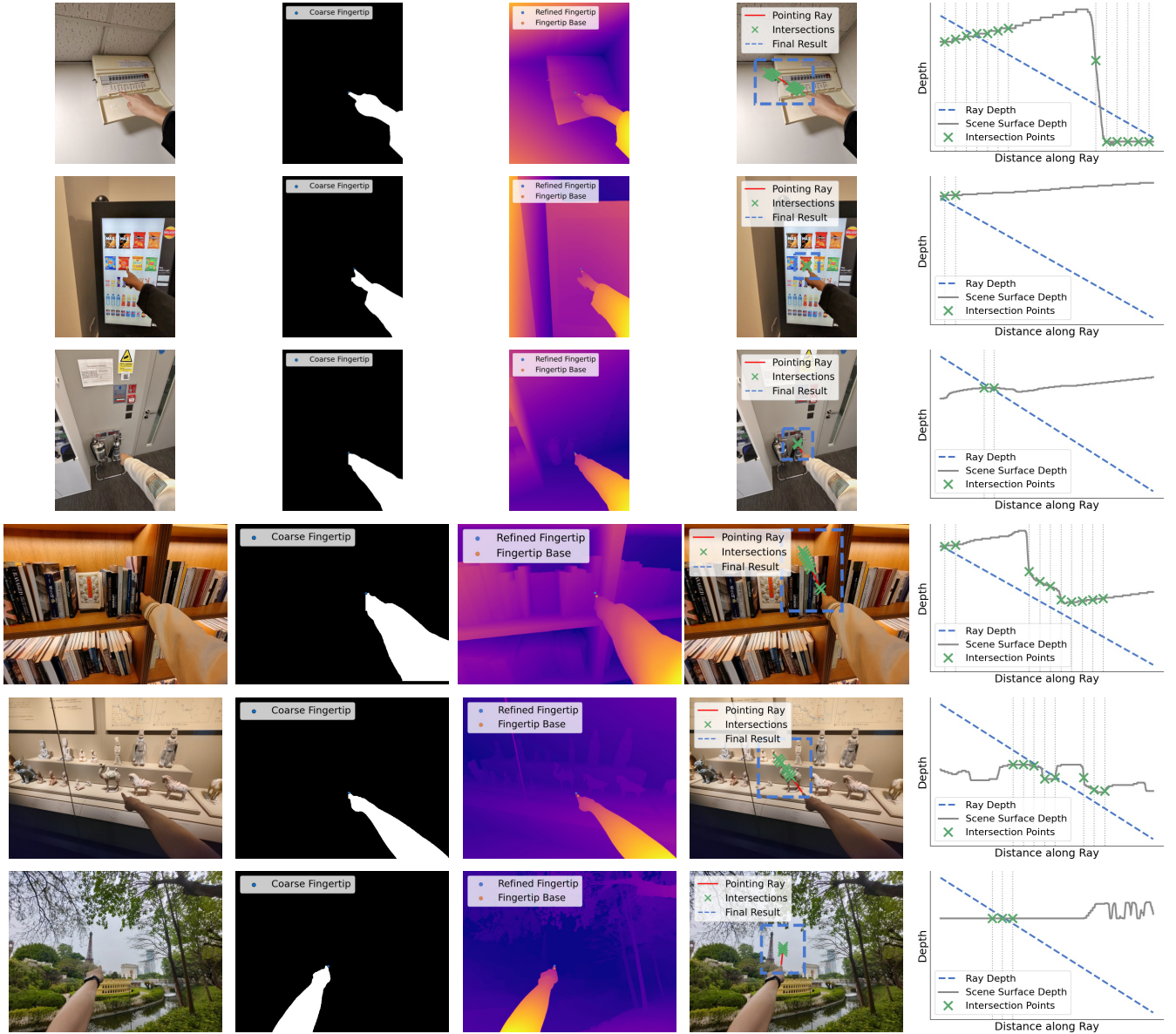


Figure 8: Visualization of our multi-stage pipeline for 3D pointing gesture interpretation across diverse egocentric scenes. Our method deterministically grounds deictic references through a sequence of geometric operations. From left to right: The process begins with the original egocentric view, followed by coarse hand segmentation to isolate the forearm. We then leverage a monocular depth estimate to refine the 3D locations of the fingertip and finger base. These points define a 3D pointing ray, which is cast into the scene. The final target is identified by analyzing the intersections between the ray and the scene’s surface geometry. The rightmost column plots the ray’s depth against the scene’s surface depth, illustrating how our algorithm robustly identifies the correct intersection point, even in the presence of complex, non-planar surfaces (e.g., bookshelves) or distant targets. This geometric approach provides a robust alternative to end-to-end models, which often struggle with such spatial reasoning.

teraction context is paramount for facilitating the VLM’s robust reasoning capabilities in egocentric referential resolution.

Extended Experimental Details

Novelty and Evaluation

Our core novelty lies in demonstrating that a hybrid, modular architecture—combining programmatic control with LLM reasoning—is a more robust and efficient paradigm for egocentric AI than monolithic models, especially for tasks

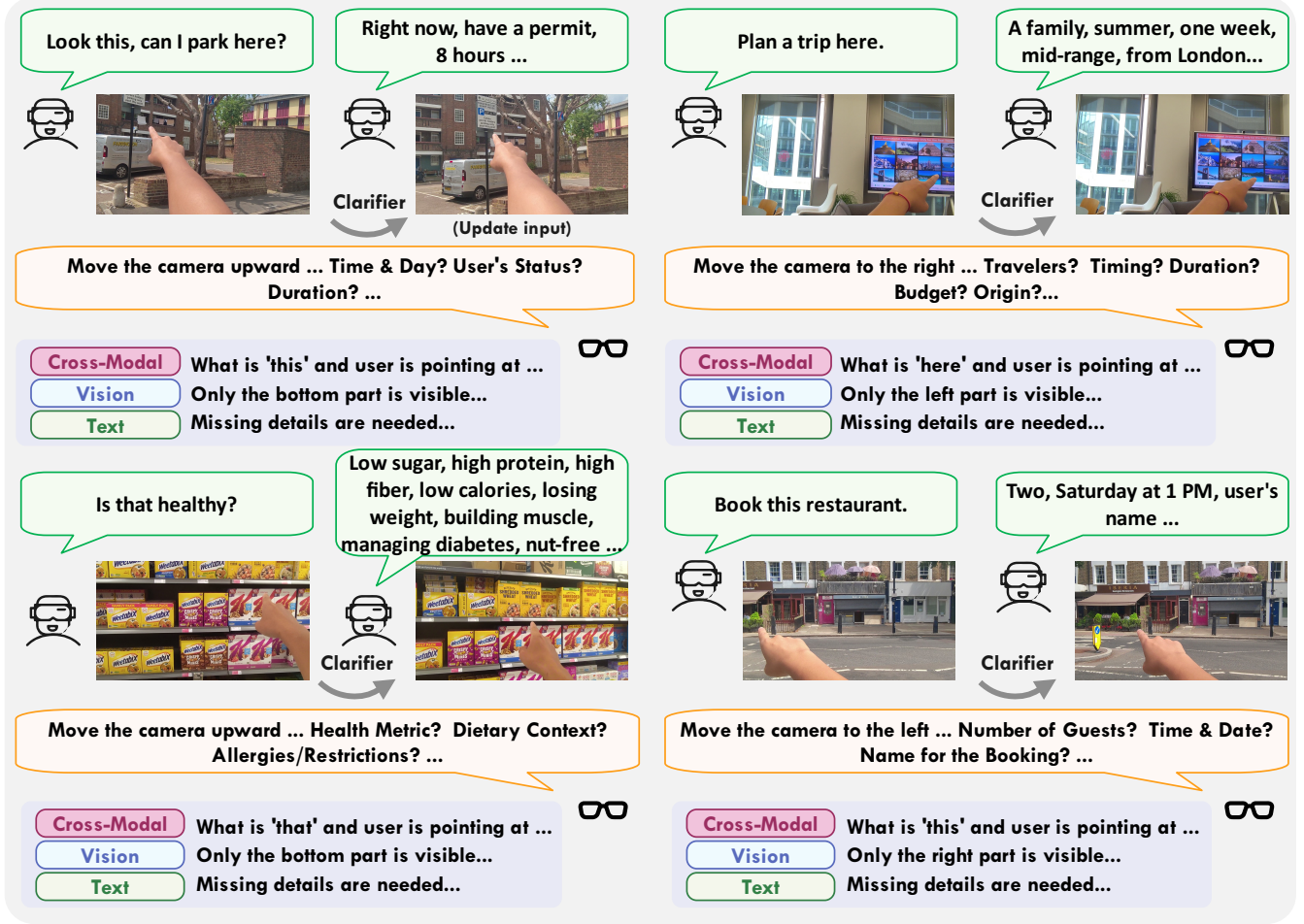


Figure 9: End-to-end examples of our framework resolving complex, co-occurring visual, textual, and cross-modal ambiguities in real-world egocentric scenarios. The area beneath the glasses icon illustrates the system’s internal thought process, where it diagnoses different types of ambiguity. This diagnosis recognizes when the text is unclear because details are missing, when the vision is incomplete because an object is partially occluded, and how a cross-modal reference like the word “this” is connected to the user’s pointing gesture. Based on this comprehensive analysis, the Clarifier generates a compound clarification request, prompting the user for both a physical camera adjustment and additional contextual information. This synergistic approach, which explicitly decomposes and addresses ambiguity across modalities, allows the framework to robustly handle initially underspecified tasks and transform them into solvable problems.

requiring spatial precision where VLMs are known to be brittle. We created VRA-Ego because existing datasets do not target the specific problem of interactive, first-person referential disambiguation. For instance, DeePoint (Nakamura et al. 2023) (third-person), or RefEgo (Kurita, Katsura, and Onami 2023) (textual), differing from the direct “what is this?” pointing gestures in real-world assistance. Our work addresses this critical gap. Regarding generalization (e.g. Ego4D (Grauman et al. 2025)), our framework is a plug-in module that only activates when ambiguity is detected, otherwise defaulting to the base VLM’s performance.

VRA-Ego Dataset Construction and Setup

Our ambiguity-oriented egocentric benchmark, VRA-Ego, was collected with current AI / AR glasses. The dataset cov-

ers pointing targets at distances from 0.2 m up to 10 m, so that the clarifier must handle both large, near-field objects and very small, far-field objects in the same pipeline. To make the cross-modal module robust to different user habits, we recorded both left- and right-hand pointing gestures, and we captured scenes in both portrait and landscape orientations. Each sample stores the RGB frame, the intended target object, and whether the raw capture was immediately answerable or required vision-based clarification. This information is later used by our automated evaluation pipeline to decide whether the controller should trigger the vision clarifier.

Automated Evaluation Pipeline for Textual Disambiguation

To comprehensively evaluate textual disambiguation and ensure scalable, reproducible assessment, we introduced a sophisticated automated evaluation pipeline. This three-stage process involves:

1. **An Interaction Simulator** to generate complete dialogue logs by interacting with the model under test.
2. **A Query Disentangler** to parse the agent’s generated questions into atomic informational units.
3. **A Semantic Matcher** that uses a powerful LLM judge (GPT-4o) to compare these atomic units against the ground-truth missing attributes from the benchmark, calculating the Recover Rate.

This pipeline allows for nuanced evaluation beyond simple accuracy, capturing the model’s ability to effectively extract specific, critical details.

Additional Results and Analysis

User Experience

User experience and efficiency are critical. Qualitatively, our framework makes the assistant feel significantly more intelligent and responsive, as shown in our supplementary demo video. And the ablations are straightforward, reflecting our method’s plug-and-play design.

Additional Baseline Results for Textual Clarification

In addition to the open-source model families presented in the main body of our work, we conducted a supplementary evaluation on the IN3 benchmark (Qian et al. 2024). This served to benchmark our -Clarifier framework against prominent proprietary models and other widely-used open-source variants. The results, detailed in Table 3, corroborate our central thesis: our programmatic, iterative approach provides a critical reasoning scaffold, substantially enhancing the instruction-following capabilities of LLMs, especially for models that are less powerful or not extensively instruction-tuned.

Our initial step was to validate our experimental setup by replicating the results reported in the original IN3 paper (denoted by ‘*’). Our baseline implementations for GPT-4 and Mistral-Interact yielded performance metrics largely consistent with the reported figures, confirming the fidelity of our reproduction. However, a significant discrepancy emerged with Mistral-7B-v0.2. The monolithic baseline prompt, employed in the original work, failed to elicit valid, format-compliant responses from the model in our environment, resulting in a complete failure to produce measurable outcomes (indicated by ‘-’). We attribute this to the model’s limited capacity to parse and execute complex, multi-part instructions embedded within a single prompt.

Furthermore, our investigation of the fine-tuned Mistral-Interact model revealed potential signs of overfitting within the original model’s training regime. While the model performs well on the IN3 test set, its instruction-following capability proved brittle; minor variations to the prompt or query

structure caused it to fail completely. This fragility prevented a robust evaluation of our -Clarifier on this specific model. We have contacted the original authors regarding this potential training artifact and details about baseline models performance but have not yet received a response.

In stark contrast, our framework demonstrates its efficacy most dramatically on models that struggle with monolithic prompting. For instance, Mistral-7B-v0.2, which failed outright under the baseline condition, becomes a competent performer when guided by our -Clarifier, achieving a 0.712 Accuracy. Even more strikingly, Mistral-7B-v0.3, whose baseline performance is near-random (0.213 Accuracy, 0 Recover Rate), is transformed into a highly effective system (0.806 Accuracy, 0.748 Recover Rate) with our framework. This substantial improvement—from complete failure to performance rivaling that of GPT-4—underscores the power of our structured, iterative decomposition in unlocking the latent capabilities of smaller language models.

Model	Accuracy	Recover Rate	Rounds
GPT-4*	0.824	0.752	2.69
GPT-4 (Baseline)	0.796	0.754	1.85
GPT-4-Clarifier	0.815	0.805	3.71
Mistral-7B-v0.2*	0.491	0.684	1.62
Mistral-7B-v0.2 (Baseline)	-	-	-
Mistral-7B-v0.2-Clarifier	0.712	0.358	2.47
Mistral-7B-v0.3 (Baseline)	0.213	0	1
Mistral-7B-v0.3-Clarifier	0.806	0.748	12.07

Table 3: Performance comparison for textual clarification on the IN3 benchmark. ‘*’ denotes results from the original paper (Qian et al. 2024). ‘-’ indicates the model failed to produce format-compliant outputs. Our -Clarifier framework not only enhances capable models like GPT-4 but also successfully scaffolds weaker models (e.g., Mistral-7B variants) that otherwise fail completely, elevating their performance to be competitive with much larger systems.

Runtime and Latency Reporting

To make the efficiency aspect of the clarifier explicit, we report the per-stage wall-clock latency on an RTX 4090 GPU. All numbers below are averaged over the full runs we used for the image-based (capture-quality) and cross-modal (pointing-based) clarifiers. We exclude one-time model loading and I/O bookkeeping. In practice, the geometric stack (depth, pose, hand, fusion, ray) remains below 400 ms per frame on this hardware, so the end-to-end latency is primarily determined by the choice and deployment of the LLM/VLM used for clarification.

Note on thresholds. All ambiguity-related thresholds (e.g., blur score, minimum crop size, depth-aware expansion ratio) used in the above runs follow the default configuration shipped with our implementation. We did not heavily tune them for the reported numbers; a dedicated sensitivity study can further tighten these timings and trigger conditions.

Stage	Avg. latency (ms)
Image I/O	11.26
Detection	185.52
Feedback generation	4.47
Visualization (optional)	6.06

Table 4: Per-stage latency for the image-based clarifier on RTX 4090. Detection is the dominant cost; feedback generation is negligible.

Stage	Avg. latency (ms)
Depth estimation	268.62
Pose detection	17.81
Hand segmentation	39.19
2D–3D pointing fusion	43.83
Ray intersection	5.50
LLM processing	5684.59

Table 5: Per-stage latency for the cross-modal clarifier on RTX 4090. The geometric part (depth, hand, ray) stays within a few hundred ms; the dominant cost is the LLM call. Development-time visualization was ≈ 3.7 s and is excluded here because it is not part of the inference path.

Broader Discussion and Future Work

Design Choices and Practicality

Our modular design is a deliberate choice for real-time egocentric assistants. Unlike systems using video, large models, or complex “thinking” agents (e.g., VideoAgent (Wang et al. 2024)), our approach prioritizes low latency. We argue that quick, iterative feedback more closely mimics natural human conversation, which is crucial for user experience. While gaze (e.g., in MICA (Sarch et al. 2025)) could enhance clarity, its high cost, power consumption, and weight make it impractical for current consumer devices. Our framework provides a more readily deployable solution.

Alternative Interaction Paradigms

Our work introduces a modular, plug-and-play framework designed to proactively resolve multimodal intent ambiguity in egocentric interactions. While the quantitative results demonstrate significant performance gains, it is equally important to discuss our approach in the context of alternative design philosophies and the broader vision for human-AI collaboration.

One might argue that simpler, more direct solutions could address the ambiguities we tackle. For instance, visual ambiguity caused by poor framing or blurriness could ostensibly be resolved by providing the user with a real-time camera preview, much like one might adjust their phone during a video call. Similarly, referential ambiguity could be eliminated if the user provided a more explicit linguistic description (e.g., “the second bottle from the left”), and underspecified textual intent could be clarified by requiring the user to input their complete request via a text box.

However, these solutions, while viable in constrained scenarios, fundamentally conflict with the vision of a seamless, conversational AI assistant that defines the next era of human-computer interaction. The ideal interaction paradigm is not a series of rigid, user-driven corrections but a fluid, natural dialogue. Imagine a phone conversation where nearly every statement required a pedantic clarification; the interaction would become intolerably cumbersome. Natural human dialogue is built on context and inference, and our goal is to imbue AI systems with this capability. Requiring explicit descriptions or manual camera adjustments imposes a cognitive load that disrupts this natural flow.

Furthermore, these manual-correction alternatives face significant practical and systemic drawbacks, especially for wearable devices like AR glasses.

1. **Limited Generalizability:** User-dependent clarification strategies do not readily generalize to autonomous agents, such as robots, which must interpret and act upon ambiguous instructions without constant human intervention. An effective disambiguation model must be part of the agent’s core intelligence.
2. **Power Consumption:** Constant camera previews or high-bandwidth video processing for user feedback are highly power-intensive, a critical bottleneck for all-day wearable devices.
3. **User Experience and Hardware Limitations:** Real-time visual feedback can induce visual fatigue. Critically, most commercially available AI glasses are equipped with fixed-focus lenses and do not support autofocus, rendering any manual, user-driven adjustments for visual clarity largely impractical from the outset. Adding such capabilities would, in turn, increase the device’s weight and complexity.

Our Plug-and-Play Clarifier is designed precisely to navigate these challenges. By building a system that proactively identifies and resolves ambiguity through an intelligent, multimodal dialogue, our framework offloads the burden of clarification from the user to the agent. It paves the way for the kind of fluid, context-aware assistance that future egocentric AI promises, where interaction feels as natural as speaking with another person. This makes our approach not just an improvement, but a foundational necessity for the future of embodied, conversational AI.

Extended Limitations and Future Work

The main paper briefly touched on limitations. Here, we expand on those points and provide more context on future directions, drawing from our initial drafts.

Our approach introduces certain trade-offs. The iterative nature of our dialogue framework, while improving accuracy, inherently increases the number of conversational turns, which could affect user experience. Furthermore, as a pluggable module, the framework’s overall efficacy is intrinsically tied to the speed and capability of the underlying foundation models. Our work also simplifies the multifaceted nature of real-world intent ambiguity. True ambiguity often stems from deep contextual dependencies or polysemy, where a query like “How should I move?” could refer

to a chess move, a fashion catwalk, or a travel route, depending on a rich context that our current model does not fully leverage.

These limitations illuminate promising avenues for future research. To enhance efficiency, our framework could be augmented with proactive query rewriting techniques (Chen et al. 2025) to clarify user intent in a single step. To improve robustness, the system could incorporate mechanisms to detect and correct misleading or contradictory information within the user’s query itself (Li et al. 2025), rather than only addressing what is missing. Ultimately, our framework provides a foundational stepping stone towards more sophisticated embodied agents. Future work can extend our principles to physical robots that must not only understand ambiguous commands but also actively navigate their environment to seek clarification and execute tasks (Ramrakhya et al. 2025), moving us closer to creating truly intelligent agents capable of navigating the full spectrum of human communicative intent.

Appendix References

- [A70] Chen, J.; Wang, B.; Jiang, Z.; and Nakashima, Y. 2025. Putting People in LLMs’ Shoes: Generating Better Answers via Question Rewriter. In *AAAI, Sponsored by the Association for the Advancement of Artificial Intelligence, Philadelphia, PA, USA*, 23577–23585. AAAI Press.
- [A71] Grauman, K.; Westbury, A.; Byrne, E.; et al. 2025. Ego4D: Around the World in 3,600 Hours of Egocentric Video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(11): 9468–9509.
- [A16] Guo, D.; Wu, F.; et al. 2025. Seed1.5-VL Technical Report. *CoRR*, abs/2505.07062.
- [A73] Kurita, S.; Katsura, N.; and Onami, E. 2023. RefEgo: Referring Expression Comprehension Dataset from First-Person Perception of Ego4D. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 15168–15178. IEEE.
- [A74] Li, G.; Liu, W.; Wu, Y.; et al. 2025. From Misleading Queries to Accurate Answers: A Three-Stage Fine-Tuning Method for LLMs. *CoRR*, abs/2504.11277.
- [A36] Nakamura, S.; Kawanishi, Y.; Nobuhara, S.; et al. 2023. DeePoint: Visual Pointing Recognition and Direction Estimation. In *IEEE/CVF International Conference on Computer Vision, ICCV, Paris, France, 20520–20530*. IEEE.
- [A41] Qian, C.; et al. 2024. Tell Me More! Towards Implicit User Intention Understanding of Language Model Driven Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1088–1113. ACL.
- [A77] Ramrakhya, R.; Chang, M.; Puig, X.; et al. 2025. Grounding Multimodal LLMs to Embodied Agents that Ask for Help with Reinforcement Learning. *CoRR*, abs/2504.00907.
- [A78] Sarch, G. H.; Kumaravel, B. T.; Ravi, S.; et al. 2025. Grounding Task Assistance with Multimodal Cues from a Single Demonstration. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 12807–12833. Association for Computational Linguistics.
- [A79] Wang, X.; Zhang, Y.; Zohar, O.; and Yeung-Levy, S. 2024. VideoAgent: Long-Form Video Understanding with Large Language Model as Agent. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXX*, volume 15138 of *Lecture Notes in Computer Science*, 58–76. Springer.