Do MLLMs Understand Pointing? Benchmarking and Enhancing Referential Reasoning in Egocentric Vision

Anonymous ACL submission

Abstract

Egocentric AI agents, such as smart glasses, rely on pointing gestures to resolve referential ambiguities in natural language commands. However, despite advancements in Multimodal Large Language Models (MLLMs), current systems often fail to precisely ground the spatial semantics of pointing. Instead, they rely on spurious correlations with visual proximity or object saliency—a phenomenon we term "Referential Hallucination." To address this gap, we introduce EgoPoint-Bench, a comprehensive question-answering benchmark designed to evaluate and enhance multimodal pointing reasoning in egocentric views. Comprising over 11k high-fidelity simulated and real-world samples, the benchmark spans five evaluation dimensions and three levels of referential complexity. Extensive experiments demonstrate that while state-of-the-art proprietary and open-source models struggle with egocentric pointing, models fine-tuned on our synthetic data achieve significant performance gains and robust Sim-to-Real generalization. This work highlights the importance of spatially-aware supervision and offers a scalable path toward precise egocentric AI assistants. The code and samples are available at https://anonymous.4open.science/r/EgoPoint-BFBD/.

1 Introduction

Egocentric Vision AI agents, particularly intelligent assistants integrated into wearable devices such as smart glasses, are fundamentally reshaping the paradigms of Augmented Reality and Human-Computer Interaction (Li et al., 2025). By perceiving the physical world through the user's perspective, these systems aim to provide precise, context-aware Question Answering (QA) services. In such naturalistic interaction scenarios, users exhibit a strong preference for minimalistic spoken commands. These utterances often blend explicit object descriptions with highly ambiguous deictic expressions (e.g., "*How do I use this?*" or "*How is the stuff over there?*"). When retrieving information from complex visual scenes, relying solely on unimodal language is often insufficient to resolve such referential ambiguity. Conversely, pointing gestures—instinctual and high-frequency actions in human communication—have been empirically proven to significantly enhance referential clarity and reduce the requisite length of natural language instructions (Mane et al., 2024; Chen et al., 2021). Consequently, endowing multimodal models with the capability to precisely comprehend "egocentric pointing" is critical for egocentric AI agents.

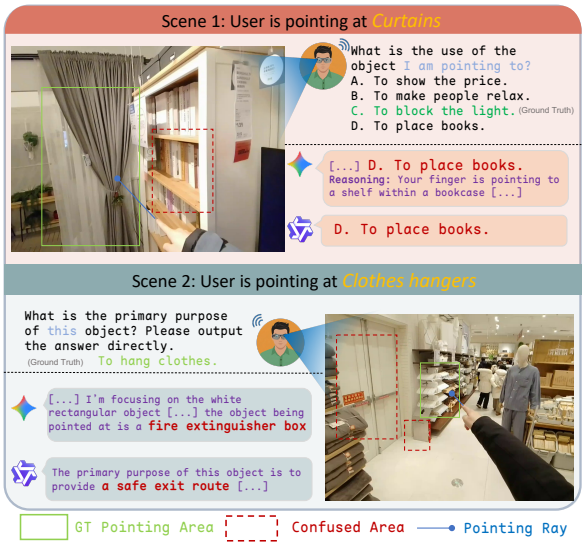

Figure 1: Spatial ambiguity in egocentric pointing. Two examples where current VLMs (e.g., Gemini 3, Qwen3-VL) fail to recognize the target spatially aligned with the pointing gesture. This highlights a critical gap in fine-grained 3D spatial reasoning. Note that neither bboxes nor rays were in the model inputs.

Despite the remarkable semantic understanding demonstrated by Multimodal Large Language Models (MLLMs) in general image captioning and QA tasks (OpenAI, 2024; Liu et al., 2023b), our investigation reveals a critical deficiency in spatial

reasoning when adapting current state-of-the-art models to egocentric pointing QA. Specifically, as depicted in Fig. 1, instead of tracing the precise geometric projection of the pointing finger, models frequently fixate on objects proximal to the hand or visually salient entities, leading to *referential hallucination*. This indicates that these models fail to grasp the intrinsic spatial mechanism of "pointing", relying instead on spurious correlations based on visual proximity.

A critical bottleneck is the scarcity of high-quality, unambiguous data aligned within the "Vision-Language-Space". While visual grounding is well-studied, benchmarks like RefCOCO (Kazemzadeh et al., 2014) and Visual Genome (Krishna et al., 2017) rely on third-person internet imagery, lacking the wide-angle nature of egocentric vision. Conversely, large egocentric datasets like Ego4D (Grauman et al., 2022) and EPIC-KITCHENS (Damen et al., 2022) prioritize action recognition or hand-object interactions (Liu et al., 2022), missing dense QA annotations that capture "pointing-object" geometry. Without this spatially-aware supervision, MLLMs fail to separate hand appearance from spatial pointing intent, hindering deictic referencing performance.

To address this challenge, we propose `EgoPoint-Bench`, a benchmark designed to systematically evaluate and enhance multi-modal spatial reasoning in egocentric views. To balance data scale with realism, our construction process involves two complementary phases: In the simulation phase, we introduce a physics-based synthesis pipeline leveraging ray-casting to generate noise-free pointing labels in 3D environments; in the real-world phase, we collect real-scenario data to validate practical applicability. For QA construction, we implemented a hybrid "machine-generation, human-verification" pipeline to ensure rigorous standards. Crucially, to capture interaction diversity and enable fine-grained assessment, we incorporated three referring language patterns ranging from explicit descriptions to implicit instructions, and structured the benchmark across five core capability dimensions. In total, the dataset comprises 10,567 high-fidelity simulation QA pairs and 1,162 real-world samples.

To evaluate generalization, we employed a hybrid test set combining held-out simulation data (in-domain) and real-world data (zero-shot cross-domain). We benchmarked open-source (e.g., Qwen3-VL) and proprietary models (e.g., GPT-5), followed by LoRA fine-tuning on simulation data. The fine-tuned models significantly outperformed baselines, including top proprietary ones. These results validate the efficacy of high-quality synthetic data and highlight the scarcity of egocentric pointing examples in current foundation models.

The main contributions of this paper are summarized as follows:

- We propose `EgoPoint-Bench`, a novel benchmark designed to evaluate multi-modal spatial reasoning in egocentric views. Our extensive benchmarking reveals that current state-of-the-art MLLMs significantly lack the capability to understand fine-grained pointing gestures in first-person scenarios.
- We develop a **physics-driven data generation pipeline** that ensures both geometric precision and linguistic diversity. By leveraging ray-casting in simulation and incorporating hierarchical referring patterns (from explicit descriptions to implicit instructions), we construct a high-quality dataset containing over 11k pairs across simulation and real-world domains.
- We demonstrate the efficacy of **sim-to-real generalization**. Models fine-tuned on our high-fidelity synthetic data significantly outperform strong proprietary models on real-world test sets, validating the potential of synthetic data in addressing data scarcity for egocentric interaction.

2 Related Work

To contextualize our contributions, we compare `EgoPoint-Bench` with representative benchmarks in visual grounding, embodied perception, and pointing-based interaction (see Table 1).

2.1 Third-Person Grounding

Foundational visual grounding benchmarks, ranging from 2D (Mao et al., 2016; Krishna et al., 2017) to 3D (Chen et al., 2020; Achlioptas et al., 2020) and robotic settings (Qi et al., 2020), rely predominantly on third-person views and explicit linguistic descriptions. Critically, they lack the *egocentric pointing signal* essential for intuitive HCI, often causing models to rely on semantic priors rather than geometric cues.

2.2 Egocentric Vision Perception

Large-scale datasets like Ego4D (Grauman et al., 2022) and EPIC-KITCHENS (Damen et al., 2018)

Table 1: Comparison with existing datasets. Unlike benchmarks that rely on third-person views or pure text, EgoPoint-Bench uniquely combines egocentric vision with natural 3D hand pointing. It supports diverse question types and multi-level linguistic granularity. **R**: Real-world data, **S**: Synthetic data.

Dataset	Egocentric	Scenes	Natural Pointing	Task	Annotation Granularity	Size
RefCOCOg (Mao et al., 2016)	✗	**R**	✗	Grounding	Image + BBox + Text	26k imgs
ScanRefer (Chen et al., 2020)	✗	**R**	✗	Grounding	3D BBox + Text	11k scenes
YouRefIt (Chen et al., 2021)	✗	**R**	✓	Grounding	BBox + Gesture + Text	3k clips
Ego4D (Grauman et al., 2022)	✓	**R**	✗	Forecasting	Activity Labels	3.6k hrs
Look & Point (Nguyen et al., 2024)	✓	**R**	✓	Grounding	Gaze/Point Vector	1.3k hrs
Ges3ViG (Mane et al., 2024)	✗	**S**	✗	Grounding	3D Grounding + Gesture	35k samples
EOC-Bench (Dang et al., 2025)	✓	**R**	✗	QA	Temporal/Cognitive QA	3.2k QAs
EgoPoint-Bench (Ours)	✓	**R+S**	✓	**QA**	**Image + Name + BBox + QA**	**11.7k QAs**

capture rich first-person activities. However, they focus primarily on passive observation, such as action recognition. They lack active interaction scenarios. Attempts to add language, like RefEgo (Kurita et al., 2023), still rely on pure text without gesture signals. Recent works like EOC-Bench (Dang et al., 2025) introduce open-ended QA to egocentric videos. Yet, they rely on artificial visual prompts, such as red boxes drawn on images. This reliance creates a domain gap for Augmented Reality (AR). In real AR scenarios, systems should interpret natural, unaugmented user gestures.

2.3 Pointing-based Interaction

To enable pointing-driven interaction, Ges3ViG (Mane et al., 2024) introduces 3D directional gestures through synthesized avatars; however, it focuses on object localization within 3D scenes rather than question-answering (QA) interaction and lacks validation on real-world datasets. While COSM2IC (Weerakoon et al., 2022) achieves deictic interaction using virtual environments, it is limited by a lack of diversity in both object categories and scene types. Furthermore, most existing datasets rely on exhaustive descriptive language to resolve target ambiguity, creating a significant gap between these benchmarks and real-life interaction scenarios. In contrast, EgoPoint-Bench integrates high-fidelity synthetic and real-world data. We shift linguistic inputs from explicit descriptions (e.g., "the object I point at") to implicit deictics (e.g., "this"), evaluating MLLMs' pointing comprehension across diverse semantic dimensions.

3 EgoPoint-Bench

3.1 Overview

As shown in Fig. 2, we propose EgoPoint-Bench, a multimodal question-answering benchmark focused on first-person pointing gestures. It is designed to quantitatively evaluate the understanding and reasoning capabilities of MLLMs regarding pointing gestures and referring language in egocentric visual perception. Given the scarcity of labeled data in this domain, we employ a dual-source data construction strategy combining simulation and real-world data. On one hand, we introduce the **Point-sim** fully automated simulation framework, which utilizes 42 hand models to generate 10,567 synthetic samples across 1,838 high-fidelity 3D scenes (sourced from Ai2-THOR (Kolve et al., 2017; Deitke et al., 2022), HSSD (Khanna et al., 2023), ReplicaCAD (Szot et al., 2021), and HM3D (Ramakrishnan et al., 2021)). On the other hand, to enhance the realistic diversity of the dataset, we collected 1,162 samples featuring natural pointing interactions in diverse real-world environments. Furthermore, the benchmark covers five core dimensions and includes three question types—multiple-choice, true/false, and open-ended questions—with established standard splits for training, validation, and testing.

3.2 Image Collection

3.2.1 Point-Sim Simulation Framework

To synthesize diverse and high-fidelity scene-object pairs, we utilized the Habitat-Sim 3.0 simulator (Puig et al., 2023) and integrated static environments sourced from the AI2-THOR, HSSD, ReplicaCAD, and HM3D datasets. Specifically, we acquired high-quality 3D arm-hand models from ArtStation (ArtStation, 2025) and leveraged the Blender package (Blender Online Community, 2018) to manipulate parameters—such as joint articulation and scaling—thereby introducing structural diversity into the generated pointing gestures. Furthermore, we applied textures representing 3 distinct skin tones and 7 clothing styles across both

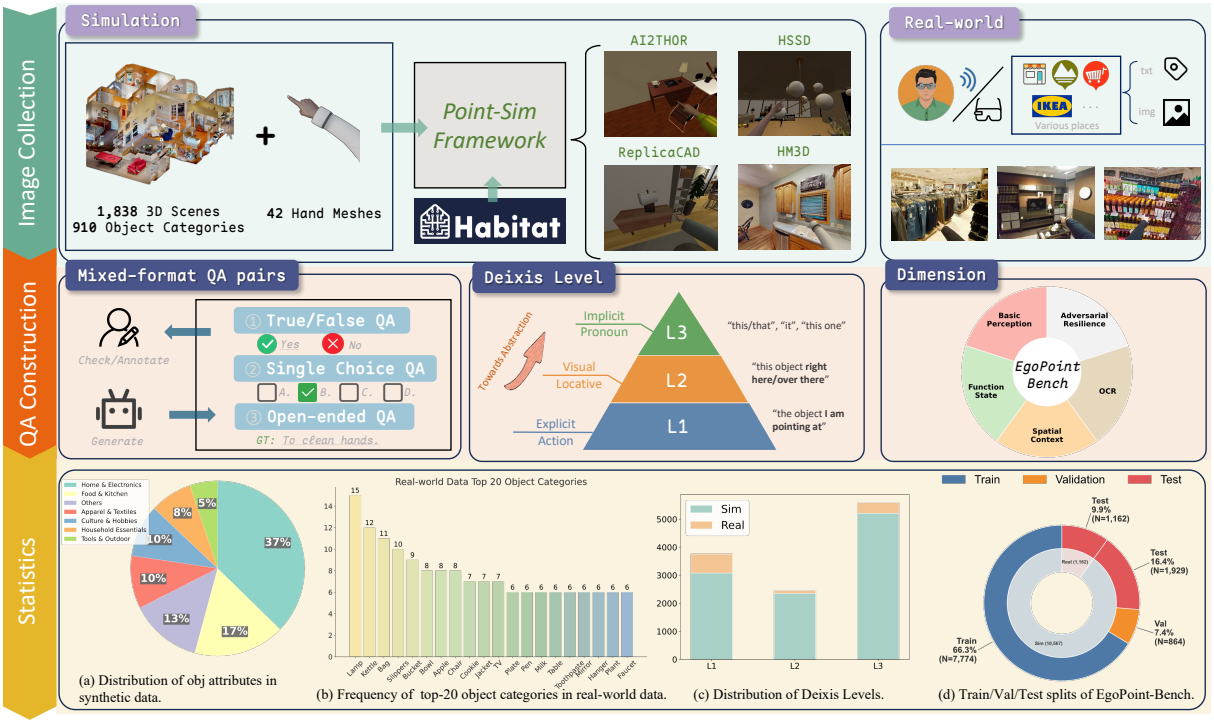

Figure 2: Overview of EgoPoint-Bench. Top: We construct the dataset using a scalable simulation pipeline (*Point-Sim*) alongside real-world collection to ensure visual diversity. Middle: The QA generation process spans five capability dimensions (Basic Perception, Function State, Spatial Context, OCR, and Adversarial Resilience) and incorporates a hierarchical deixis level taxonomy (L1: Explicit Action, L2: Visual Locative, L3: Implicit Pronoun), challenging models to resolve referential ambiguity based on finger-pointing gestures. Bottom: Detailed statistics showing object attributes, category frequency, and data distribution.

left and right hands, resulting in a total of 42 unique pointing models.

Simulation Initialization. To ensure domain robustness, we initialize the simulation with a diverse set of intrinsic and extrinsic parameters. To replicate the wide-angle optical characteristics of modern smart glasses, the camera's vertical field of view (FOV) is uniformly sampled from $[100°, 115°]$. The agent is modeled with an ocular height $h_{eye} \sim \mathcal{U}(1.45, 1.70)$ meters, equipped with a multi-modal sensor suite capturing aligned RGB, Depth, and Semantic observations. Hand dominance (left/right) is randomized to balance the dataset distribution.

Target-Oriented Spatial Arrangement. For a selected target object O centered at $P_{obj} \in R^3$, we compute the navigable manifold of the scene, represented as a Navigation Mesh (NavMesh) (Mononen, 2009). We sample a candidate agent position P_{agent} on this manifold within a constrained radius r_{search} (default ≤ 3.0m), conditioned on a minimum collision clearance of 0.4m. To mitigate scale ambiguity, the sampling distance is dynamically scaled based on the object's volumetric size; this prevents scenarios where the object is either imperceptible or encompasses the entire field of view.

Once P_{agent} is fixed, we orient the agent's camera to face the target. We construct the camera rotation matrix $R_{cam} \in SO(3)$ by aligning the optical axis with the forward vector $\mathbf{f} = (P_{obj} - P_{agent})/\|P_{obj} - P_{agent}\|$. The rotation is defined compactly as:

$$R_{cam} = \left[\frac{\mathbf{f} \times \mathbf{u}_w}{\|\mathbf{f} \times \mathbf{u}_w\|}, \quad \frac{(\mathbf{f} \times \mathbf{u}_w) \times \mathbf{f}}{\|\mathbf{f} \times \mathbf{u}_w\|}, \quad -\mathbf{f} \right]^\top \tag{1}$$

where $\mathbf{u}_w$ is the global up vector.

Kinematic Hand Alignment. We instantiate the hand model within the lower visual field of the camera. The core objective is to align the index finger's direction vector with the line of sight to the object. Let $\mathbf{u}_{rest}$ denote the normalized initial directional vector of the index finger and $\mathbf{u}_{target}$ be the normalized vector pointing from the hand to the object. We compute the minimal rotation R_{hand} via *Rodrigues' rotation formula*. The rotation is parame-

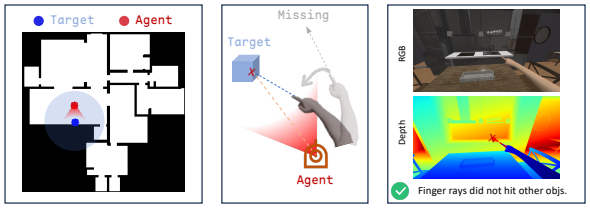

Figure 3: Point-sim Simulation Framework.

terized by the unit rotation axis $\mathbf{k} = \frac{\mathbf{u}_{rest} \times \mathbf{u}_{target}}{\|\mathbf{u}_{rest} \times \mathbf{u}_{target}\|}$ and angle $\theta = \arccos(\mathbf{u}_{rest} \cdot \mathbf{u}_{target})$:

$$R_{hand} = I + [\mathbf{k}]_\times \sin\theta + [\mathbf{k}]_\times^2 (1 - \cos\theta) \quad (2)$$

where $[\mathbf{k}]_\times$ denotes the skew-symmetric matrix of $\mathbf{k}$. Subsequently, to simulate realistic human pointing behavior, we apply small stochastic perturbations to the pitch and yaw of the computed camera orientation.

Validation and Data Format. We enforce a validity check by casting a ray from the index finger tip toward P_{obj}. An instance is discarded if the ray intersects with any obstacle before reaching the target. The pipeline explicitly exports a comprehensive data tuple $\mathcal{D} = \{I_{rgb}, I_{depth}, I_{sem}, \mathbf{b}_{obj}, P_{2D}, y_{id}\}$, containing the images, 2D bounding boxes, projected coordinates, and semantic identifiers. This pipeline is generalized to support any scene compatible with Habitat-Sim.

3.2.2 Real-world Data Collection

We recruited eight volunteers equipped with MLVision smart glasses (MLVision, 2025) to collect data on objects of interest in diverse real-world environments. The data collection scenarios spanned a broad spectrum of settings, including but not limited to indoor places like furniture stores, convenience stores, and apartments, as well as outdoor locations such as shopping malls, zoos, and streets. Participants were instructed to record a video whenever they encountered an object of interest, explicitly pointing at the target while verbally stating its name to serve as the ground truth and posing a relevant description or question. In total, 1,162 valid image frames were curated from the collected footage (see Appendix C.1 for details).

3.3 Capability Taxonomy

Inspired by canonical multimodal benchmarks like MMBench (Liu et al., 2024b) and MME (Fu et al., 2025a), we design a five-dimensional taxonomy to comprehensively evaluate MLLMs within first-person pointing interactions. This framework is structured to bridge the gap between low-level perception and high-level robust reasoning:

- **Basic Perception (BP):** Identifies fundamental attributes (category, color, texture) and visual distinctiveness for gesture alignment.
- **Function & State (FS):** Infers semantic properties (e.g., edibility, operability) and dynamic functional states.
- **Spatial Context (SC):** Perceives egocentric spatial relationships, including localization, scene compatibility, and reachability.
- **OCR:** Extracts textual info from targets, such as brand names, slogans, and instructions.
- **Adversarial Resilience (AR):** Maintains reliability against adversarial inputs like counterfactuals, fallacies, and void references.

3.4 QA Pair Construction

For comprehensive deictic evaluation, our dataset employs a hierarchical taxonomy and hybrid question format.

Hierarchical Deixis Taxonomy. We design three levels of deixis to cover the broadest possible semantic range of referential inquiries: L1 (Explicit Action) describes the gesture directly (e.g., "the object I am pointing at"); L2 (Visual Locative) implies spatial proximity (e.g., "that thing over there"); and L3 (Implicit Pronoun) relies purely on visual context (e.g., "this").

Task Formulation. To balance ecological validity with objective evaluation, we adopt diverse question formats. We incorporate Open-ended questions to reflect the natural, unrestricted nature of human inquiry. However, to ensure a fair, consistent, and automated testing benchmark, we also construct True/False and Single-Choice Questions. This hybrid composition retains the semantic complexity of realistic user intent while facilitating rigorous quantitative comparison.

Human-Machine Collaborative Data Curation. To ensure both diversity and scalability, we established a collaborative data generation pipeline. For the simulation subset, we leveraged a generative model to synthesize QA pairs, thereby mitigating the rigidity of fixed templates and expanding the dimensionality of potential questions (Liu et al., 2023b). To prevent model hallucinations—specifically the misidentification of pointed-at objects—we implemented a visual

Table 2: Main results on real-world and simulation testsets. We highlight the best Direct results in blue and the best LoRA results in orange. The Gain column shows the improvement of LoRA over Direct.

Model	Method	Simulation testset						Real-world testset						Overall	
		BP	FS	SC	OCR	AR	Mean	BP	FS	SC	OCR	AR	Mean	Avg.	Gain
Random	-	27.95	26.83	38.89	43.24	52.17	31.14	25.19	22.74	37.30	26.32	45.76	28.94	30.24	-
Human	-	91.86	97.14	100	93.33	100	95.80	96.24	98.04	96.39	95.65	89.09	96.00	95.90	-
Closed-source Models															
Gemini-3.0-pro	Direct	52.47	51.39	70.47	74.85	57.16	56.44	66.63	75.44	79.06	83.28	60.16	72.00	62.29	-
Gemini-3.0-flash	Direct	54.39	53.33	66.58	73.64	58.39	57.21	67.04	73.98	78.89	80.90	63.02	71.84	62.71	-
GPT-5-2-Instant	Direct	54.14	49.81	66.14	75.45	50.88	54.80	55.31	67.49	81.62	69.55	71.27	66.76	59.29	-
GPT-5-mini	Direct	59.96	58.22	67.65	68.79	36.09	57.66	52.81	66.73	67.32	66.27	52.38	60.57	58.75	-
Open-source Models (Direct vs. LoRA)															
Llava-1.5-7B	Direct	50.83	46.89	54.86	50.91	41.92	48.82	36.48	45.85	62.13	22.69	69.37	47.19	48.21	-
	LoRA	76.41	72.06	60.63	66.06	86.44	73.18	37.50	56.55	64.17	33.43	95.40	54.54	66.17	+17.96
Llava-Next-7B	Direct	47.42	45.42	55.92	53.33	46.59	48.17	31.68	51.75	60.09	39.40	56.19	46.44	47.52	-
	LoRA	80.39	80.86	79.56	72.42	86.13	80.93	40.10	66.32	71.23	40.90	90.63	59.64	72.93	**+25.41**
GLM-4.6V-Flash	Direct	56.16	50.81	66.14	61.52	36.17	53.29	48.32	59.77	67.32	72.84	43.49	56.42	54.47	-
	LoRA	77.16	73.28	82.01	80.00	64.21	74.86	53.88	60.70	66.55	67.16	72.70	61.26	69.74	+15.27
InternVL-3.5-2B	Direct	51.97	55.14	61.50	66.97	26.05	51.74	44.85	60.47	62.55	59.40	43.65	53.73	52.49	-
	LoRA	71.40	75.36	76.61	78.79	81.99	75.43	46.33	64.04	71.83	57.31	89.68	62.03	70.39	+17.90
InternVL-3.5-8B	Direct	52.86	52.50	63.51	66.36	35.63	52.62	50.05	60.88	63.32	68.96	50.79	57.09	54.30	-
	LoRA	74.60	77.81	82.76	78.79	86.21	78.86	50.56	69.88	74.47	63.88	90.00	66.13	74.07	+19.77
InternVL-3.5-14B	Direct	46.79	51.14	62.07	71.52	33.56	49.99	47.76	65.09	72.51	65.07	45.24	58.59	53.23	-
	LoRA	75.99	76.00	83.01	76.36	86.51	78.59	54.03	73.10	80.26	68.66	82.86	68.92	74.95	+21.72
Qwen3-VL-8B	Direct	57.55	54.00	70.34	77.58	52.11	58.29	47.81	58.42	74.55	68.96	53.17	58.14	58.23	-
	LoRA	81.31	80.92	80.56	84.24	82.91	81.36	60.36	72.28	81.96	71.94	88.57	71.96	77.83	+19.60
Qwen3-VL-32B	Direct	56.52	53.75	65.64	79.39	60.23	58.28	56.38	65.03	76.09	79.70	56.83	64.30	60.54	-
	LoRA	80.75	82.50	83.39	83.03	82.84	82.20	62.09	71.35	81.96	73.43	83.81	71.84	78.30	+17.76

prompting strategy (Yang et al., 2023): ground-truth bounding boxes were rendered directly onto the input images to explicitly guide the model's focus. Furthermore, ground-truth category labels and attributes were injected into text prompts to ensure context-aware responses. We validated the fidelity of this automated pipeline through a manual inspection of the test set, identifying and correcting a minimal 3% error rate. The real-world dataset followed a rigorous human-in-the-loop workflow. Annotators labeled the bounding boxes of target objects based on raw open-ended descriptions or questions. Additionally, they provided factual answers and underwent strict cross-verification.

3.5 Dataset Statistics

`EgoPoint-Bench` comprises 10,567 simulation and 1,162 real-world QA pairs, with an average question length of 9.81 words. The simulation subset is partitioned into 8,638 samples for training/validation (9:1 split) and 1,929 for testing, while the real-world data serves exclusively as a test set. To ensure rigorous evaluation, each `(scene, object)` tuple in the simulation data appears exactly once. The dataset covers 1,838 unique scenes and 910 object categories. Fig. 2 presents detailed statistics regarding (a) synthetic object attributes, (b) top-20 real-world object categories, (c) deixis levels, and (d) dataset splits.

4 Experiments

4.1 Experimental Setup

We conduct a comprehensive evaluation across a wide spectrum of MLLMs, spanning both proprietary and open-source architectures. For proprietary models, we test the latest iterations including Gemini-3.0 (Pro/Flash) (Team et al., 2025a) and the GPT-5 series (5.2-Instant/5-Mini) (OpenAI). For open-source models, we select representative baselines with varying scales: InternVL-3.5 (2/8/14B) (Wang et al., 2025), Qwen3-VL (8/32B) (Bai et al., 2025), LLaVA v1.5 (Liu et al., 2023a), LLaVA-NeXT (Liu et al., 2024a), and GLM-4.6v-Flash (Team et al., 2025b). To establish performance bounds, we incorporate a random baseline for choice-based tasks and report human performance evaluated on 1,000 samples (balanced between simulation and real-world data) by three volunteers. The evaluation operates under two settings: (1) **Zero-shot Inference**, where models directly predict answers from visual-textual inputs; and (2) **Instruction Tuning**, where we apply LoRA-based (Hu et al., 2022) parameter-efficient fine-tuning. Crucially, our training set consists exclusively of simulation data to assess sim-to-real generalization. Implementation details are provided in Appendix A.

Table 3: Detailed Breakdown by Question Type. Types: Single-Choice ($\mathcal{SCQ}$), True/False($\mathcal{TF}$), Open-Ended questions ($\mathcal{OQ}$). Dimensions: Basic Perception (BP), Function & Affordance (FS), Spatial Context (SC), OCR & Text (OCR), Adversarial Relation (AR). Blue indicates best Direct performance; Orange indicates best LoRA performance.

Model	Method	BP			FS			SC			OCR			AR		
		$\mathcal{SCQ}$	$\mathcal{TF}$	$\mathcal{OQ}$	$\mathcal{SCQ}$	$\mathcal{TF}$	$\mathcal{OQ}$	$\mathcal{SCQ}$	$\mathcal{TF}$	$\mathcal{OQ}$	$\mathcal{SCQ}$	$\mathcal{TF}$	$\mathcal{OQ}$	$\mathcal{SCQ}$	$\mathcal{TF}$	$\mathcal{OQ}$
Random	-	26.25	40.62	-	23.28	49.37	-	29.44	48.06	-	26.67	46.67	-	26.67	50.26	-
Closed-source Models																
Gemini-3.0-pro	Direct	60.39	50.00	33.23	61.70	64.56	35.14	80.95	74.27	60.34	95.56	76.67	67.59	53.33	67.69	48.02
Gemini-3.0-flash	Direct	61.44	59.38	33.87	61.81	70.89	37.84	79.22	69.90	60.51	91.11	70.00	70.34	60.00	69.74	49.04
GPT-5-2-Instant	Direct	55.87	56.25	36.45	57.80	62.03	32.79	76.19	69.90	70.77	73.33	80.00	67.93	66.67	73.85	38.76
GPT-5-mini	Direct	57.72	62.50	44.52	62.61	78.48	35.50	67.10	74.76	55.56	71.11	70.00	63.45	33.33	55.90	26.10
Open-source Models (Direct vs. LoRA)																
Llava-1.5-7B	Direct	44.83	56.25	40.65	48.74	45.57	30.09	60.61	62.14	45.30	26.67	80.00	22.07	13.33	75.38	27.01
	LoRA	60.16	53.12	68.06	71.79	40.51	48.83	67.53	63.11	49.74	44.44	90.00	32.76	73.33	98.97	80.11
Llava-Next-7B	Direct	40.77	56.25	35.81	49.20	54.43	28.83	58.87	61.65	48.38	44.44	83.33	28.62	46.67	64.10	34.12
	LoRA	63.41	78.12	62.58	79.01	77.22	53.15	79.65	83.50	55.73	51.11	93.33	41.72	86.67	95.38	79.10
GLM-4.6V-Flash	Direct	53.08	71.88	41.29	54.70	70.89	33.51	67.10	68.93	61.71	75.56	60.00	64.48	46.67	46.67	28.93
	LoRA	67.71	81.25	59.03	70.87	81.01	47.93	77.92	77.18	67.52	73.33	83.33	68.62	80.00	76.41	55.48
InternVL-3.5-2B	Direct	49.83	62.50	31.29	60.89	67.09	17.84	66.23	68.45	42.05	64.44	76.67	55.17	26.67	43.08	19.77
	LoRA	60.98	81.25	52.58	74.89	79.75	41.08	78.79	82.04	53.16	62.22	90.00	61.03	80.00	92.82	75.71
InternVL-3.5-8B	Direct	52.85	56.25	33.55	58.72	64.56	20.90	71.43	65.05	44.79	71.11	73.33	62.07	46.67	54.36	24.86
	LoRA	64.69	78.12	58.39	78.56	79.75	46.13	83.55	84.47	61.54	66.67	96.67	61.72	80.00	94.36	80.45
InternVL-3.5-14B	Direct	47.62	65.62	31.61	58.83	64.56	24.14	71.00	69.90	51.62	71.11	83.33	58.28	46.67	49.74	22.94
	LoRA	67.71	78.12	50.97	78.33	78.48	47.03	84.42	86.41	68.72	77.78	86.67	61.03	86.67	89.23	80.90
Qwen3-VL-8B	Direct	54.36	62.50	37.74	57.68	62.03	32.97	73.16	76.70	62.05	73.33	76.67	71.38	53.33	58.97	45.20
	LoRA	73.17	78.12	63.55	81.31	81.01	51.17	80.52	89.32	68.03	77.78	93.33	70.34	73.33	91.79	77.97
Qwen3-VL-32B	Direct	57.61	65.62	35.81	59.98	67.09	30.09	74.89	68.93	62.56	80.00	80.00	78.97	60.00	65.13	52.43
	LoRA	73.64	75.00	64.52	81.65	88.61	50.45	82.68	88.83	72.31	77.78	86.67	74.14	80.00	85.13	81.24

4.2 Evaluation Metrics

`EgoPoint-Bench` comprises three task types: True/False (TF), Single Choice Questions (SCQ), and Open-ended Questions (OQ). Following established protocols (Fu et al., 2025b; Li et al., 2024), we adopt exact matches for the TF and SCQ tasks. For the OQ task, evaluating open-ended responses remains challenging; therefore, we employ an LLM-as-a-Judge approach (Zheng et al., 2023). Specifically, GPT-4o (OpenAI, 2024) scores the model predictions against ground-truth answers on a scale of 0 to 1 (with an increment of 0.2). Further details can be found in Appendix A.4.

4.3 Main Results

Table 2 presents the performance of proprietary and open-source models across simulation and real-world test sets. We reported 3 key observations:

Off-the-shelf VLMs struggle with fine-grained egocentric deictic understanding. In the Direct inference setting, even the most advanced proprietary models (e.g., Gemini-3.0-pro, GPT-5-mini) and open-source models fail to achieve satisfactory performance, hovering around 60% accuracy overall. A significant gap remains compared to human performance (95.90%), particularly in tasks requiring precise spatial geometric reasoning (AR and BP metrics). This underscores that general-purpose pre-training is insufficient for comprehending complex "finger-pointing" semantics in egocentric views.

Simulation-based tuning yields significant gains. Fine-tuning with our generated simulation data via LoRA brings substantial improvements across all open-source models. As shown in the "Gain" column, we observe a consistent performance boost ranging from +15.27% to +25.41%. Notably, LLaVA-Next-7B achieves a remarkable 25.41% improvement, demonstrating that the visual-semantic alignment provided by our synthetic data effectively unlocks the models' potential for pointing-oriented VQA tasks.

Strong Sim-to-Real generalization. Crucially, the models trained on simulation data generalize exceptionally well to the Real-world testset. For instance, Qwen3-VL-8B improves its real-world mean accuracy from 58.14% to 71.96% after tuning on simulation data. This suggests that the geometric and semantic features of finger-pointing learned from our high-fidelity simulation environment are robust and transferrable, validating the efficacy of our data generation pipeline for real-world applications.

4.4 Detailed Analysis

Analysis Across Different Question Types. Table 3 dissects model performance across three answer formats ($\mathcal{SCQ}, \mathcal{TF}, \mathcal{OQ}$), revealing three critical insights: (1) **Generative bottleneck.** Direct

Table 4: Performance evaluation of representative MLLMs on Sim and Real test sets across three deixis levels (L1-L3). The best results are highlighted in **bold**.

Model	Method	Sim			Real		
		L1	L2	L3	L1	L2	L3
Gemini-3.0-pro	Direct	51.03	59.00	59.53	**72.57**	65.20	72.76
GPT-5-mini	Direct	58.22	59.82	56.02	59.32	54.40	64.38
InternVL-3.5-2B	Direct	52.51	53.83	49.96	56.25	48.20	50.73
	LoRA	74.71	74.60	76.47	59.03	61.40	67.50
Llava-1.5-7B	Direct	48.11	51.47	47.98	42.27	52.60	54.48
	LoRA	72.01	75.60	72.83	51.21	60.80	58.80
Qwen3-VL-32B	Direct	50.52	62.72	62.28	64.31	62.40	64.79
	LoRA	**83.77**	**81.63**	**81.20**	69.59	**71.80**	**75.83**

models exhibit a sharp performance drop in Open-Ended questions ($\mathcal{OQ}$) compared to discriminative formats ($\mathcal{SCQ}, \mathcal{TF}$), indicating that while pre-trained models can distinctively *recognize* correct references, they struggle to actively *formulate* precise spatial descriptions without specific tuning. (2) **Geometric alignment in Adversarial Relations.** The AR dimension, which requires distinguishing targets from spatial distractors, sees the most dramatic gains from LoRA (e.g., Llava-1.5-7B AR-$\mathcal{OQ}$ jumps from 27.01% to 80.11%). This proves that our dataset effectively teaches the specific "logic of pointing" absent in general pre-training. (3) **Spatial-semantic saturation.** While text-heavy tasks (OCR) show robust baseline performance, spatial tasks (BP, SC, AR) benefit disproportionately from fine-tuning, confirming that our method primarily enhances fine-grained spatial capabilities rather than basic visual recognition.

Impact of different deixis levels. Contrary to the intuition that explicit instructions should mitigate ambiguity, our results reveal that L1 (Explicit Action) does not consistently outperform L2 (Visual Locative) or L3 (Implicit Pronoun). For instance, in the Sim dataset, the Direct Qwen3-VL-32B model shows a significant drop in L1 (50.52%) compared to L2 (62.72%) and L3 (62.28%). This counter-intuitive finding underscores a critical deficiency in current MLLMs: even when explicitly prompted to attend to a pointing gesture, models struggle to grounded the spatial action, indicating a lack of genuine understanding of fine-grained geometric cues. Furthermore, L3 often achieves the highest accuracy in the Real dataset (e.g., 75.83% for Qwen3-VL-32B LoRA). This suggests that instead of resolving the specific deictic gesture, models may over-rely on object saliency or scene priors to infer the target.

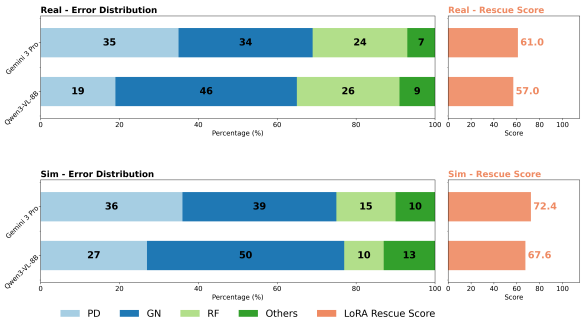

Figure 4: Distribution of error types and rescue scores.

4.5 Error Types

To probe the limitations of current VLMs in finger-pointing VQA, we conducted a manual analysis on 400 error cases generated by Qwen3-VL-8B and Gemini-3-Pro (balanced between simulated and real-world data). We classified errors into three primary categories: (1) **Proximal Distraction (PD)**, where the model fails to follow the pointing ray and instead grounds the answer to a distractor immediately adjacent to the finger; (2) **Gesture Neglect (GN)**, where the model ignores the gesture entirely, attending to visually salient or distant objects; and (3) **Reasoning Failure (RF)**, where the target is correctly localized, but the model fails in downstream reasoning. Fig. 4 (Left) illustrates the error distribution, revealing that PD and GN are the most prevalent failure modes. Fig. 4 (Right) demonstrates the efficacy of our approach by reporting the "Rescue Score"—defined as the percentage of these specific failure cases successfully corrected by our LoRA-finetuned Qwen3-8B. Our method achieves Rescue Scores ranging from 57.0% to 72.4% across datasets, confirming its capability to effectively recover from the spatial ambiguity and gesture perception issues inherent in the baselines. More examples are provided in Appendix B.2.

5 Conclusion

We introduced `EgoPoint-Bench` to evaluate and enhance MLLMs' understanding of egocentric finger-pointing gestures. Our evaluation reveals that while existing MLLMs struggle with this task, fine-tuning on high-quality synthetic data mitigates referential hallucinations, enabling robust real-world generalization. This work paves a scalable path toward precise egocentric AI assistants.

Limitations

While EgoPoint-Bench provides a benchmark for evaluating current egocentric multimodal fingerpointing understanding, it possesses two primary limitations: 1) Although fine-tuning with automatically synthesized simulation data has proven effective on real-world datasets, we observed that the performance gain on real-world data is smaller than that on simulated data. This suggests that real-world user pointing behaviors, along with environmental complexities such as arm backgrounds, are significantly more intricate and challenging than those in simulation. Simulated data struggles to sufficiently cover the behavioral characteristics of the real world. 2) To facilitate easier evaluation, current dataset questions and answers are relatively brief, which diverges from the complex, multi-turn dialogue patterns found in real-world interactions. We focus first on whether MLLMs can explicitly understand the fundamental meaning of "pointing," as our experimental results indicate that even this poses a significant challenge for current models. Mastering these basic comprehension skills is a vital prerequisite before addressing more difficult and complex multi-turn interaction tasks.

Ethical Statement

University ethics review board approves human-subjects research and they approved this project. In our real-world data collection environment, we have anonymized all human faces and any identifying information within the images by applying a blurring treatment. This ensures that no privacy leaks occur and that the dataset contains no harmful content. All datasets used in this work, including HM3D, AI2-THOR, ReplicaCAD, and HSSD, are properly cited and used strictly for non-commercial academic research purposes.

References

Panos Achlioptas and 1 others. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*.

ArtStation. 2025. ArtStation. https://www.artstation.com. Accessed: 2025-12-05.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Blender Online Community. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*.

Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. 2021. Yourefit: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1385–1395.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and 1 others. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130(1):33–55.

Dima Damen and 1 others. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*.

Ronghao Dang and 1 others. 2025. Ecbench: Can multimodal foundation models understand the egocentric world? *arXiv preprint arXiv:2501.05031*.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. 2022. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *NeurIPS*. Outstanding Paper Award.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025b. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Kristen Grauman and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. 2023. Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation. *arXiv preprint*.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73.

Shuhei Kurita and 1 others. 2023. Refego: Referring expression comprehension dataset from first-person perception of ego4d. In *ICCV*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Xiang Li, Heqian Qiu, Lanxiao Wang, Hanwen Zhang, Chenghao Qi, Linfeng Han, Huiyu Xiong, and Hongliang Li. 2025. Challenges and trends in egocentric vision: A survey. *arXiv preprint arXiv:2503.15275*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *37th Conference on Neural Information Processing Systems (NeurIPS)*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yunze Liu, Yun Liu, Che Jiang, K Alvarez, Su Yang, Yanwei Fu, and 1 others. 2022. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022.

Atharv Mahesh Mane and 1 others. 2024. Ges3vig: Incorporating pointing gestures into language-based 3d visual grounding. In *CVPR*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*.

MLVision. 2025. Mlvision official website. `https://mlvison.com/`. Accessed: 2026-01-05.

Mikko Mononen. 2009. Recast: Navigation-mesh construction toolkit for games. Https://github.com/recastnavigation/recastnavigation.

Tuan Nguyen and 1 others. 2024. Look & point: Egocentric-exocentric pointing-based reference resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

OpenAI. Gpt-5 system card.

OpenAI. 2024. Gpt-4o system card. Accessed: 2026-01-05.

Xavi Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Ruslan Partsey, Jimmy Yang, Ruta Desai, Alexander William Clegg, Michal Hlavac, Tiffany Min, Theo Gervet, Vladimir Vondrus, Vincent-Pierre Berges, John Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, and 4 others. 2023. Habitat 3.0: A co-habitat for humans, avatars and robots.

Yuankai Qi and 1 others. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. 2021. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, and 2 others. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025a. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Dulanga Weerakoon, Vigneshwaran Subbaraju, Tuan Tran, and Archan Misra. 2022. Cosm2ic: optimizing real-time multi-modal instruction comprehension. *IEEE Robotics and Automation Letters*, 7(4):10697–10704.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Experimental Setup

A.1 Model Configurations

Regarding the configurations of the mainstream MLLMs we evaluated: specifically, for the `Qwen3-VL` and `InternVL-3.5` series, we utilized their `Instruct` variants. Furthermore, for all open-source models, we set `Do Sample=False` during inference; and for all closed-source models, we set `Temperature=0.0` and `Top-P=1`. This implies that we employed deterministic decoding strategies (i.e., greedy search) to eliminate randomness during generation, thereby ensuring the reproducibility of the evaluation results and fairness in comparisons across different models.

A.2 Additional Implementation Details

To systematically evaluate the performance of Multi-modal Large Language Models (MLLMs) on `EgoPoint-Bench`, we utilized the official open-source implementations of each model. All evaluation experiments and instruction tuning processes were conducted on NVIDIA A100 GPUs. Our evaluation framework is built upon the Hugging Face Transformers library[1] and leverages the LLaMA-Factory framework (Zheng et al., 2024) for efficient fine-tuning.

To ensure fair comparison and reproducibility, we standardized training configurations across all models using LoRA ($r = 8$) applied to all linear layers. We utilized a global batch size of 64 (per-device batch size 8 with 8 accumulation steps), enabled bfloat16 precision, and trained for 3 epochs with a learning rate of 1×10^{-4} using a Cosine learning rate scheduler.

A.3 Curated Prompt Templates

The text data utilized for both zero-shot inference and LoRA fine-tuning remains consistent across all models, formatted as follows:

Prompt Templates

```
Single Choice
USER: {Question} \n {Options} \n Answer
directly using the letters of the options
given.

True/False
USER: {Question} \n Answer directly with
'True' or 'False'
```

[1] https://huggingface.co/docs/transformers

```
Open Ended
USER: {Question} \n Please output the answer directly.
```

A.4 Scoring Open-ended Question

We use the following carefully crafted prompts and to score each open-ended question:

Evaluation Prompt Template

```
Role: You are a helpful assistant evaluation judge.
Please evaluate the candidate answer against the reference answer based on the question.
Assign a score from 0 to 5.

Scoring Criteria:
0: Completely incorrect or irrelevant.
1: Contains some keywords but fails to answer the question logic.
2: Partially correct but misses key constraints.
3: Mostly correct, but contains minor hallucinations or ambiguity.
4: Correct meaning, but phrased awkwardly or includes unnecessary fluff.
5: Perfect match in meaning and accuracy.

Input:
Question: {question}
Reference Answer: {answer}
Candidate Answer: {model_output}

Output Format:
You MUST return a valid JSON object strictly adhering to the following structure:
{
   "score": <integer_0_to_5>,
                                "reason": "<short_explanation_string>"
}
```

B Additional Analysis

B.1 Detailed Dataset Statistics

Fig. 5 illustrates the top 50 most frequent object categories in the simulation dataset. These categories primarily encompass complex indoor scenes, where high spatial coupling and environmental complexity pose significant challenges for model understanding. Consequently, the dataset demonstrates high sample diversity and task difficulty.

Fig. 6 illustrates the word cloud of all questions within EgoPoint-Bench. The results reveal a prevalence of deictic expressions (e.g., this, pointing at, here, that), indicating a strong emphasis on both explicit pointing and ambiguous reference. This distribution aligns perfectly with the core design philosophy of EgoPoint-Bench: to evaluate the model's capability in referential understanding during egocentric multimodal interactions.

Table 5 provides a detailed breakdown of the data sources across the training, validation, and testing sets. Extensive samples were drawn from HM3D due to its high-fidelity rendering of real-world environments. Conversely, ReplicaCAD was sampled sparingly and utilized only for training and validation, given its limited variety of scenes and objects. Notably, real-world data was reserved exclusively for testing to evaluate zero-shot generalization. Furthermore, the average question length of 9.81 underscores the distinctive nature of deictic language in egocentric VQA tasks.

Table 5: Dataset Statistics and Split Details

Source	Subset	Train	Val	Test	Total	Avg. QA Len.
Sim	HM3D	3227	365	718	4310	10.12
	HSSD	1964	214	605	2783	8.68
	AI2-THOR	1982	220	606	2808	10.22
	ReplicaCAD	601	65	-	666	8.67
Real	-	-	-	1162	1162	11.02

Figs. 7 and 8 illustrate the distribution of question dimensions and types in the test set, respectively. The dataset primarily evaluates Basic Perception and Affordance, mirroring common queries in daily life regarding object attributes and functional utilities. To ensure objective benchmarking, the questions are predominantly binary and multiple-choice, while open-ended questions are included to better simulate real-world QA scenarios.

Furthermore, Fig. 9 shows a balanced distribution of question types in the training set, preventing the model from developing a preference bias toward specific answer labels.

B.2 Error Analysis

Figs. 10 and 11 illustrate three representative error types made by Gemini-3-Pro and Qwen3-VL-8B on real-world and simulation datasets, respectively (where Q denotes the question, A the model's response, and GT the ground-truth intent). The results indicate that these models are highly susceptible to interference from objects in close proximity to the hand or prominent objects in the background.

Fig. 12 presents two examples of random inquiries conducted in real-world environments. In

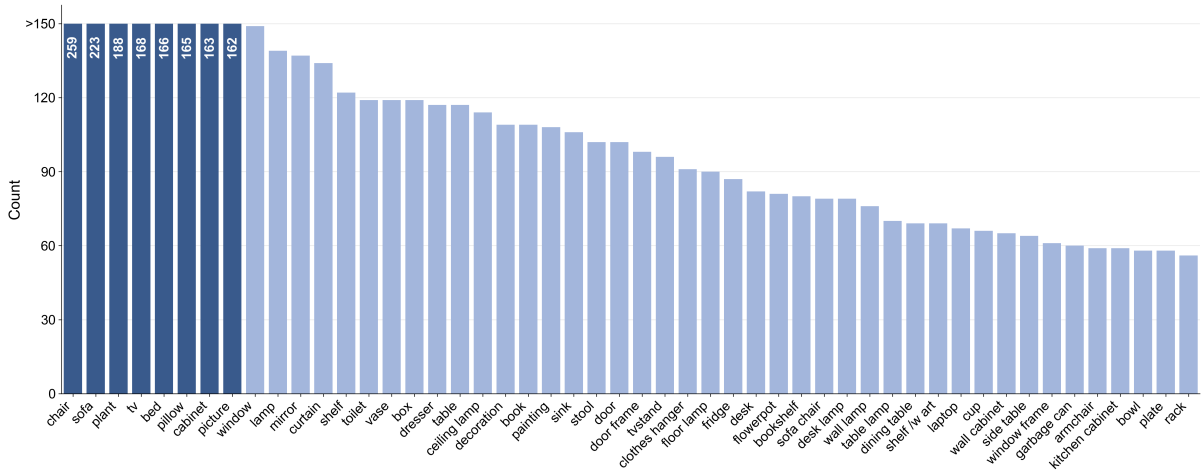

Figure 5: Frequency of top-50 object categories in simulation data.

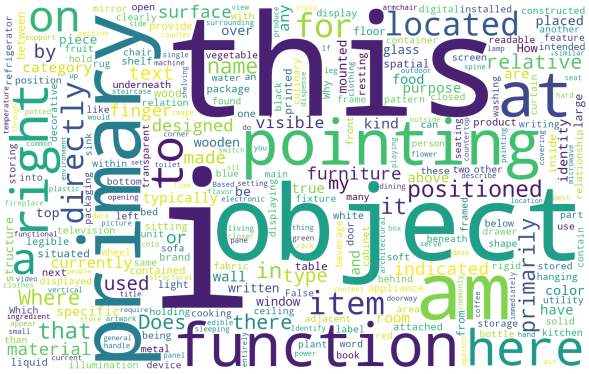

Figure 6: Worldcloud of questions in EgoPoint-Bench.

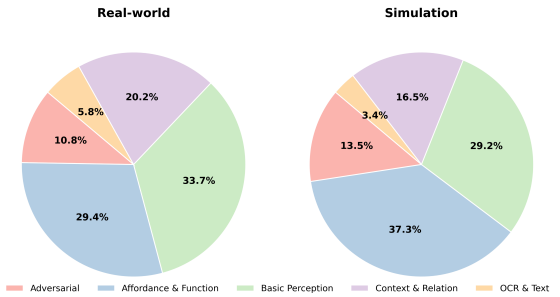

Figure 7: Distribution of 5 dimensions in EgoPoint-Bench testset.

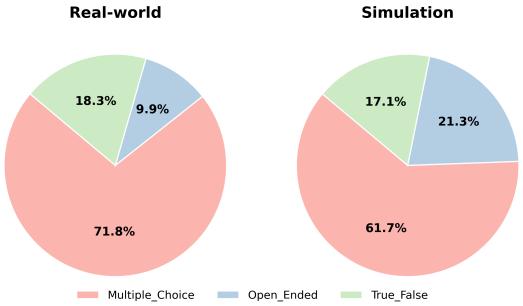

Figure 8: Distribution of 3 question types in EgoPoint-Bench testset.

the first example, both Gemini-3-Pro and Qwen3-VL-8B provide incorrect and inconsistent answers, highlighting their tendency to make arbitrary guesses in the background when the reference is unclear. In the second example, featuring a white and a brown jacket, the user points toward the white one; however, due to perspective effects, the finger region appears closer to the brown jacket in the image. Consequently, both base models consistently fail this task. In contrast, our Qwen3-VL-8B model, fine-tuned with LoRA on simulation data, is able to answer both questions with complete accuracy.

C Additional Information

C.1 Real-World Data Construction

To bridge the domain gap between simulation and reality, we constructed a high-quality real-world dataset focusing on egocentric pointing interactions.

C.1.1 Data Acquisition and Automated Pre-processing

Automated Alignment Pipeline. We designed a precision pipeline combining automated extraction with manual verification to achieve alignment across "Pointing Action – Target Object – Speech Description – Semantic QA."

- **Voice-Driven Keyframe Localization:** The process begins with speech recognition. We

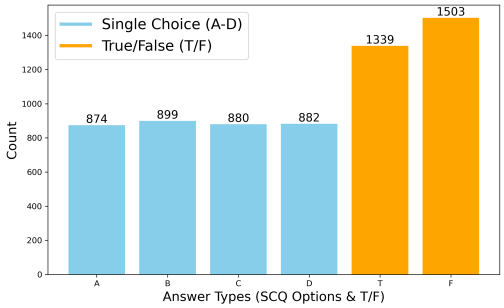

Figure 9: Option distribution of training set.

employed the industrial-grade open-source model **FunASR**[2] (`paraformer-zh`) to generate timestamped transcriptions.

  - We defined a specific trigger word (e.g., "Start") to mark the onset of a pointing action.
  - The system automatically detects the timestamp of this trigger and extracts the immediately following object noun as the candidate target.
  - This process defines a temporal window of interest for visual extraction.

- **Clarity-Aware Frame Selection:** To mitigate motion blur caused by head movements and device jitter, we implemented a **Multi-Metric Clarity Assessment** algorithm rather than random frame sampling. This algorithm fuses three complementary metrics:

  1. **Laplacian Variance:** Captures high-frequency components to detect general focus blur.
  2. **Frequency Domain Analysis:** Analyzes the spectral energy distribution to identify motion blur patterns.
  3. **Edge Density:** Evaluates the sharpness of structural edges within the frame.

  By normalizing and computing a weighted fusion of these metrics (with all weighting coefficients set to 1.0), we assign a comprehensive clarity score to every frame within the identified time window. The top-performing frames with the highest scores are selected as candidate representative images.

[2] https://github.com/modelscope/FunASR

C.1.2 Human-in-the-Loop Annotation

To ensure high quality, we employed a rigorous *Human-in-the-Loop* (HITL) pipeline. The process involves close collaboration between annotators and data collectors to guarantee that annotations faithfully reflect the original pointing intent.

Manual Annotation Workflow. Based on the candidate clear frames selected by the automated algorithm, human annotators perform the following steps:

1. **Frame Selection & Privacy Protection:** Manually select the frames that clearly contain the hand gesture from the top candidates. Any visible faces in the background are blurred to protect privacy.

2. **Transcription Verification:** Verify the correctness of the object name and description automatically transcribed by the ASR system.

3. **BBox Annotation:** Manually draw Bounding Boxes (BBox) around the pointed-at object. This step requires deep cooperation and communication with the original data collectors to ensure the annotated object and BBox strictly align with the user's original pointing intention, especially in cluttered scenes. Each collector and annotator was paid $15 per hour.

C.2 QA Generation

To synthesize QA pairs, Gemini-3-Pro is employed across our simulated and real-world datasets. We ensure the generation of high-fidelity labels by leveraging simulator-derived ground truth, specifically by superimposing red bounding boxes on the target objects. To further guide the model's reasoning, visual inputs are supplemented with exact object nomenclature and exhaustive descriptions. Regarding real-world samples, the original open-ended user queries are utilized as description for prompting. After manual validation, the refined prompt templates are formulated as follows:

Data Generation Specialist Prompt

```
SYSTEM_PROMPT
# Role
You are an expert Data Generation
Specialist for Vision-Language Models.
Your goal is to create ONE single,
high-quality Question-Answer pair for
an egocentric image based strictly on
the specific constraints provided by the
user.
```

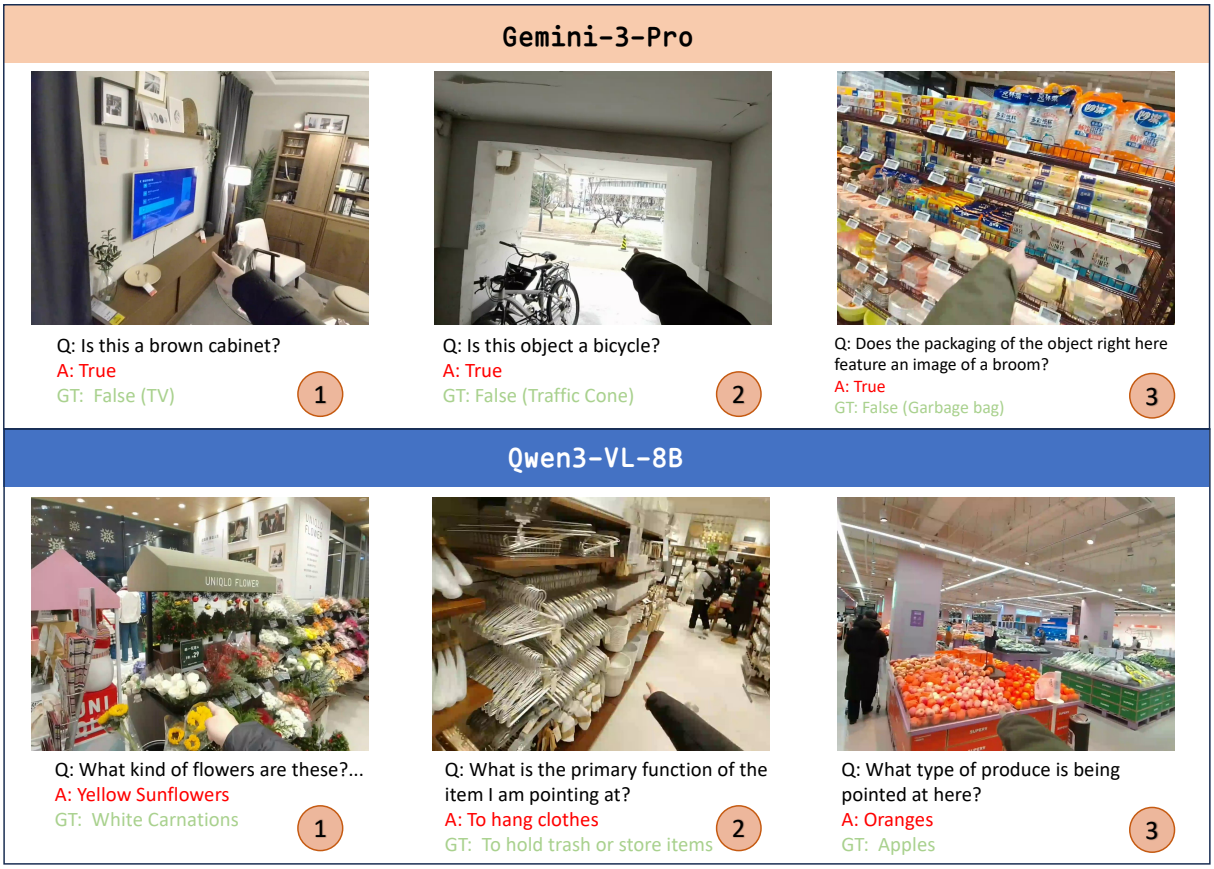

Figure 10: Error examples of three types in two methods from real-world data.

Context
You will be provided with:
1. The **Target Object** name (Ground Truth).
2. The **Target Object** description or question.
3. The specific **Dimension** (e.g., Affordance, Basic Perception).
4. The specific **Deixis Level** (how the object is referenced).
5. The specific **Question Type** (e.g., Multiple Choice).

Critical Constraint: The "Red Box" Rule
- The target object is highlighted with a red bounding box in your internal vision.
- **NEVER** mention "red box", "rectangle", "highlight", or "outline" in the text.
- Pretend the user is pointing at the object with their finger.

Guidelines for Quality
1. Anti-Cheating Option Generation (Crucial for Multiple Choice)
You must avoid "lazy" distractors. Follow this logic to generate options:
- **Correct Answer:** The ground truth label or attribute.
- **Distractor 1 (Scene Hard Negative):** An object that is **present elsewhere in the image** but NOT being pointed at.
- **Distractor 2 (Visual Hard Negative):** An object sharing similar **color, shape, or texture** with the target.
- **Distractor 3 (Contextual Hard Negative):** An object plausibly found in this specific environment, but definitely NOT the target.
- **Verification:** Ensure the correct answer is unique and unambiguous among options.

2. Zero-Leakage Question Formulation
- **The "Blindfold" Test:** If a human can guess the answer just by reading the question (without the image), the question is BAD.
- **Bad:** "What is this red round fruit?" (Reveals color, shape, category).
- **Good:** "What is the name of this object?" (Reveals nothing).

Figure 11: Error examples of three types in two methods from simulation data.

```
# Definitions of Constraints
## Deixis Levels (Reference Style)
- L1 (Explicit Action): “the object I am pointing at”, “what is indicated by my finger”.
- L2 (Visual Locative): “this object right here”, “that thing over there”.
- L3 (Implicit Pronoun): “this”, “it”, “this one”.

## Dimensions (Question Topic)
- Basic Perception: category, color, shape, material, counting.
- Affordance & Function: Edibility, operation method, state, safety, utility.
- Context & Relation: Spatial position, scene compatibility.
- OCR & Text: Reading text on the object.
- Adversarial: Asking about non-existent parts or false premises.

## Question Types
- True_False: Answer is “True” or “False”.
- Multiple_Choice: Provide 4 options (A/B/C/D).
- Open_Ended: Answer is a concise phrase.

# Output Format
Output ONLY a pure JSON object containing the single generated pair.
JSON Structure:
{
  “qa_pairs”: [
    {
      “question”: “string”,
      “options”: [“A. string”, “B. string”, “C. string”, “D. string”] OR null,
      “answer”: “string”,
      “dimension”: “string”,
      “deixis_level”: “string”,
      “type”: “string”,
      “rationale”: “string”
    }
  ]
}
```

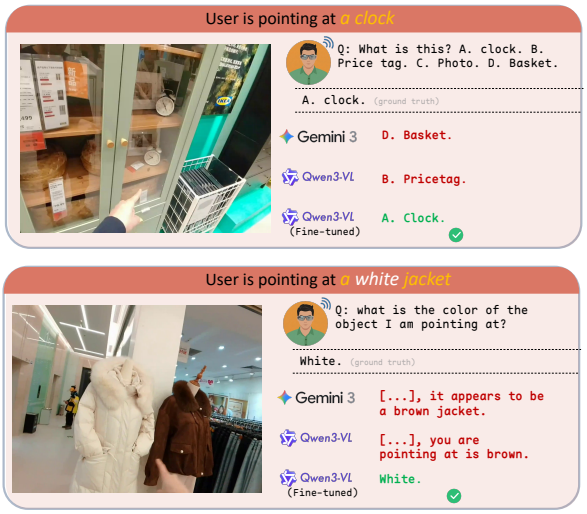

Figure 12: Comparison of model performance on real-world pointing tasks.

USER
I need you to generate a QA pair for the following object based on these strict requirements:

1. **Target Object (Ground Truth):** {{obj_name}}
2. **Desription or Question:** {{description}}
3. **Required Dimension:** {{dimension}}
4. **Required Deixis Level:** {{deixis_level}}
5. **Required Question Type:** {{q_type}}

Instruction: Generate a question that strictly fits the dimension above.
Use the specified deixis phrasing style.
Format the answer according to the question type.
Ensure no leakage of the object's name in the question.