Beyond Language: Grounding Referring Expressions with Hand Pointing in Egocentric Vision

Ling Li[1], Bowen Liu[2], Zinuo Zhan[3], Peng Jie[3], Jianhui Zhong[2], Kenglun Chang[4], and Zhidong Deng[1,*]

[1] Tsinghua University, Beijing, China
liling25@mails.tsinghua.edu.cn
[2] Dalian University of Technology, Dalian, China
[3] Northwestern Polytechnical University, Xi'an, China
[4] Apple, China
(*) Corresponding Author

Abstract. Traditional Visual Grounding (VG) predominantly relies on textual descriptions to localize objects, a paradigm that inherently struggles with linguistic ambiguity and often ignores non-verbal deictic cues prevalent in real-world interactions. In natural egocentric engagements, hand-pointing combined with speech forms the most intuitive referring mechanism. To bridge this gap, we introduce EgoPoint-Ground, the first large-scale multimodal dataset dedicated to egocentric deictic visual grounding. Comprising over **15k** interactive samples in complex scenes, the dataset provides rich, multi-grained annotations including hand-target bounding box pairs and dense semantic captions. We establish a comprehensive benchmark for hand-pointing referring expression resolution, evaluating a wide spectrum of mainstream Multimodal Large Language Models (MLLMs) and state-of-the-art VG architectures. Furthermore, we propose SV-CoT, a novel baseline framework that reformulates grounding as a structured inference process, synergizing gestural and linguistic cues through a Visual Chain-of-Thought paradigm. Extensive experiments demonstrate that SV-CoT achieves an **11.7%** absolute improvement over existing methods, effectively mitigating semantic ambiguity and advancing the capability of agents to comprehend multimodal physical intents. Code: https://github.com/lingli1724/Egopoint

Keywords: Visual Grounding · Hand-Pointing Grounding · Dataset

1 Introduction

Visual Grounding (VG) [4, 11, 17, 61, 81, 87, 89] aims to localize target objects in images based on natural language queries [39, 54, 80], serving as a fundamental capability for natural human-AI interaction [21,36,46,48,57,63,65,74,84]. Existing VG methods primarily rely on static third-person perspective images and purely textual descriptions for object localization [10, 14, 34, 40, 56, 62, 76, 85, 87]. However,

Fig. 1: Our dataset features images captured from an egocentric perspective where a hand is pointing towards a target object. The dataset includes four core annotations: Hand Annotation, Target Object Annotation, Image Caption, and Visual QA Pairs. While conventional Referring Expression Comprehension (REC) tasks only support language-assisted localization, our proposed task aims to address referring disambiguation by utilizing physical hand information as opposed to traditional linguistic assistance for target grounding.

this paradigm faces fundamental challenges in real-world interactive scenarios: language inherently exhibits ambiguity [25, 42], while humans heavily depend on non-linguistic cues (e.g., gestures) for referential disambiguation in daily communication [7, 20, 22, 28, 29, 38, 69]. For instance, when a user says "pass me that one" models struggle to identify the intended referent without gestural context. Figure 1 illustrates the significant difference between our proposed method and current approaches on the referring question-answering task [69, 82].

In physical interactions, the synergy between hand gestures and language constitutes the most natural and efficient referring mechanism [9, 32, 47, 60, 69]. A typical scenario involves a user wearing AR glasses [8, 43, 58] pointing at an object and asking "What is this?"—the finger direction directly resolves linguistic ambiguity [16,37] and precisely guides attention [43,77] to the target [16,31,32,67]. Despite the ubiquity of such multimodal referring in human interaction, existing VG research has largely overlooked gestural cues in first-person perspectives [2, 12, 45]. Mainstream datasets focus on third-person images with descriptive text, lacking modeling of gesture-language coordination in egocentric interactions [12, 14, 19, 24, 35, 44, 49, 56, 67, 68]. Accordingly, existing benchmark tasks have not incorporated gestures into evaluation protocols, limiting model applicability in embodied AI [23, 82] and human-robot collaboration scenarios [26, 33].

To bridge this gap, we introduce EgoPoint-Ground—the first large-scale multimodal dataset for first-person hand-pointing referring tasks. The dataset comprises over 15,000 high-quality samples captured from natural interactive

moments in complex scenes. Each sample is annotated with hand bounding boxes, target object bounding boxes, image captions, object captions, category labels, and referring question-answer pairs, providing fine-grained supervision signals for multimodal referring expression resolution. Based on this dataset, we establish a benchmark for hand-pointing referring expression resolution and conduct systematic evaluations of mainstream Multimodal Large Language Models (MLLMs) [1, 3, 6, 15, 30, 52, 72, 75, 83, 88] and visual grounding models. Experimental results demonstrate that existing methods suffer significant performance degradation when gestural cues are absent, quantitatively validating the critical role of hand gestures in resolving linguistic ambiguity.

Furthermore, we propose a novel baseline framework, SV-CoT, which explicitly synergizes gestural and linguistic cues through a Visual Chain-of-Thought reasoning paradigm. The model iteratively reasons about the spatial relationship between finger direction and candidate objects, effectively mitigating ambiguity introduced by purely linguistic descriptions. Experiments on EgoPoint-Ground show that our method achieves a 11.7% performance improvement over language-only baselines, fully demonstrating the necessity and effectiveness of multimodal fusion in real-world interactive scenarios.

Our main contributions are summarized as follows: **(1) Dataset:** We present EgoPoint-Ground, the first large-scale multimodal dataset for first-person hand-pointing referring, containing 15,000+ samples with multi-dimensional annotations including hand poses, object locations, and referring expressions. **(2) Benchmark:** We establish an evaluation benchmark for hand-pointing referring expression resolution and systematically analyze the limitations of existing state-of-the-art models on this task. **(3) Model:** We design a partial-Reasoning-based Visual Chain-of-Thought (SV-COT) baseline model, which fuses gestural and linguistic information via a Visual Chain-of-Thought mechanism, significantly improving referring localization accuracy. **(4) Open-source:** We release the dataset, benchmark code, and model implementation to advance research in multimodal referring and embodied AI.

(a) Distribution of Top 20 Samples by Count **(b)** Word Cloud of Sample Categories **(c)** Word Cloud of Image Caption

Fig. 2: Statistical overview and distribution analysis of the Ego-Point Ground dataset. The dataset encompasses a broad spectrum of everyday objects, accompanied by semantically rich and lexically diverse textual annotations.

Table 1: Feature comparison of EgoPoint versus representative reference datasets.

Dataset	Lang.	Gest.	Embo.	Ego/Exo	Pose	VQA	Source	No. images
PointAt [68]	✗	✓	✓	Ego	✗	✗	Lab	220
ReferAt [67]	✗	✓	✓	Exo	✗	✗	Lab	242
IPO [70]	✗	✓	✓	Exo	✗	✗	Lab	278
IMHF [71]	✗	✗	✓	Ego	✗	✗	Lab	1,716
RefIt [41]	✓	✗	✗	Exo	✗	✗	CLEF	19,894
YourRefIt [12]	✓	✓	✓	Exo	✗	✗	Crowd-Sourced	497,348
RefCOCO [86]	✓	✗	✗	Exo	✗	✗	MSCOCO	19,994
RefCOCO+ [59]	✓	✗	✗	Exo	✗	✗	MSCOCO	19,992
RefCOCOg [62]	✓	✗	✗	Exo	✗	✗	MSCOCO	26,711
Flickr30k ent. [64]	✓	✗	✗	Exo	✗	✗	Flickr30K	31,783
GuessWhat? [18]	✓	✗	✗	Exo	✗	✗	MSCOCO	66,537
Cops-Ref [13]	✗	✓	✗	Exo	✗	✗	COCO/Flickr	75,299
CLEVR-Ref+ [55]	✗	✗	✗	Exo	✗	✗	CLEVR	99,992
Ours	✓	✓	✓	**Ego**	✓	✓	**Mixed/Hybrid**	**15,338**

Note: Ego: Egocentric; Exo: Exocentric; Lang.: Language; Gest.: Gesture; Pose: Pose annotations; VQA: Visual Question Answering; Embo.: Embodied. *Source* indicates the data origin: Lab (expert capture), public datasets (e.g., MSCOCO, CLEVR), or Crowd-sourced (AMT). Mixed/Hybrid denotes a combined collection methodology.

2 EgoPoint-Ground Dataset

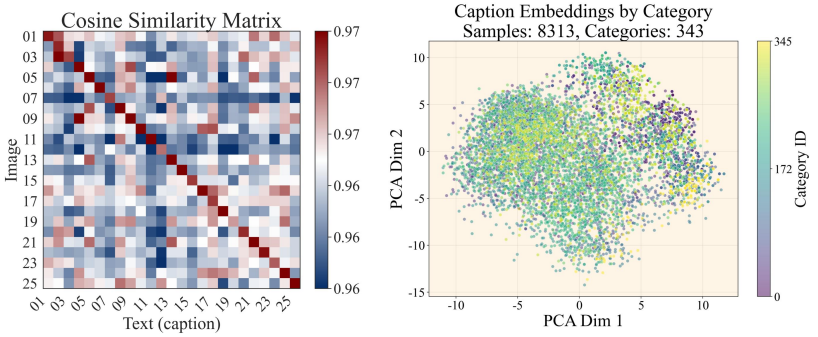
The EgoPoint-Ground Dataset represents the first high-fidelity and high-complexity egocentric benchmark designed explicitly for hand referential grounding. It features diverse scenarios critical for egocentric interaction, particularly simulating the challenge of referential Question Answering (QA) [5] within smart-glasses settings [73]. The images are meticulously annotated with precise pixel-level joint labels, encompassing a comprehensive set of visual elements: target object bounding boxes, detailed hand poses information, and corresponding natural language descriptions. Table 1 details the comparison between our dataset and existing reference understanding datasets.

Figure 1 provides an overview of the dataset's structure and content. The following sections detail the composition, creation pipeline, and statistical analysis of the EgoPoint-Ground Dataset. More details are in the Appendix.

2.1 Dataset Analysis

Our dataset comprises 15,338 images spanning more than 300 object categories and over 20 everyday scenarios, such as indoor, outdoor, street, and kitchen environments, as well as animal- and food-related scenes. Each image contains an average of 2.8 candidate objects, with 63.7% of samples featuring ≥ 2 objects of the same category and 42.5% of targets exhibiting moderate or severe occlusion. The dataset includes both positive and negative samples, with 2,171 negative samples (14.15%) that further increase localization difficulty.

We adopt a two-stage splitting strategy. First, under the mixed split, all data sources are combined and randomly partitioned into training, validation, and test sets at a $7:1:2$ ratio. Second, to simulate real-world deployment, we employ a domain-adaptive split: real-world data captured by smart glasses is reserved



(a) Similarity Matrix of Visual-Textual Semantic Alignment

(b) Semantic Distribution of Categories (Scatter Plot)

Fig. 3: Analysis of Cross-Modal Quality and Semantic Diversity. Figure (a) shows the visual-textual semantic alignment heatmap. The high diagonal similarity confirms strong semantic consistency between images and descriptions. Figure (b) displays the category semantic distribution scatter plot. Image description embeddings are visualized in 2D using PCA (color-coded by category). The points are uniformly distributed without significant clustering, demonstrating the diversity of the dataset's semantic space.

exclusively for testing, while synthetic and generative data are used for training. Specifically, real-captured, synthetically composed, and LLM-generated samples number 6,035, 5,976, and 3,327, accounting for 39.35%, 38.96%, and 21.69% of the dataset, respectively.

Category distribution reveals that person dominates at 47.1%, reflecting the inherent nature of egocentric perception—interactions between the wearer and other individuals constitute the most frequent and fundamental scenario. The remaining 19 major categories are relatively balanced (each ranging from 2.2% to 4.3%), as shown in Figure 2.

Data collection involved 17 participants with diverse demographic attributes (gender, skin tone, age) across varied environments and time periods, ensuring comprehensive representation in both subject and environmental diversity.

Cross-Modal Consistency Analysis We assess the cross-modal consistency between images and captions by randomly sampling N_I images and N_C captions and calculating the cosine similarity between their CLIP-generated [66] visual (v_i) and textual (t_j) embeddings:

$$\text{sim}_{ij} = \frac{v_i \cdot t_j}{|v_i||t_j|}. \tag{1}$$

As illustrated in Figure 3(a), the image-text similarity heatmap exhibits a pronounced diagonal dominance. This robust alignment confirms the precise semantic correspondence of the image-text pairs within our dataset, thereby validating its high annotation quality and consistency.

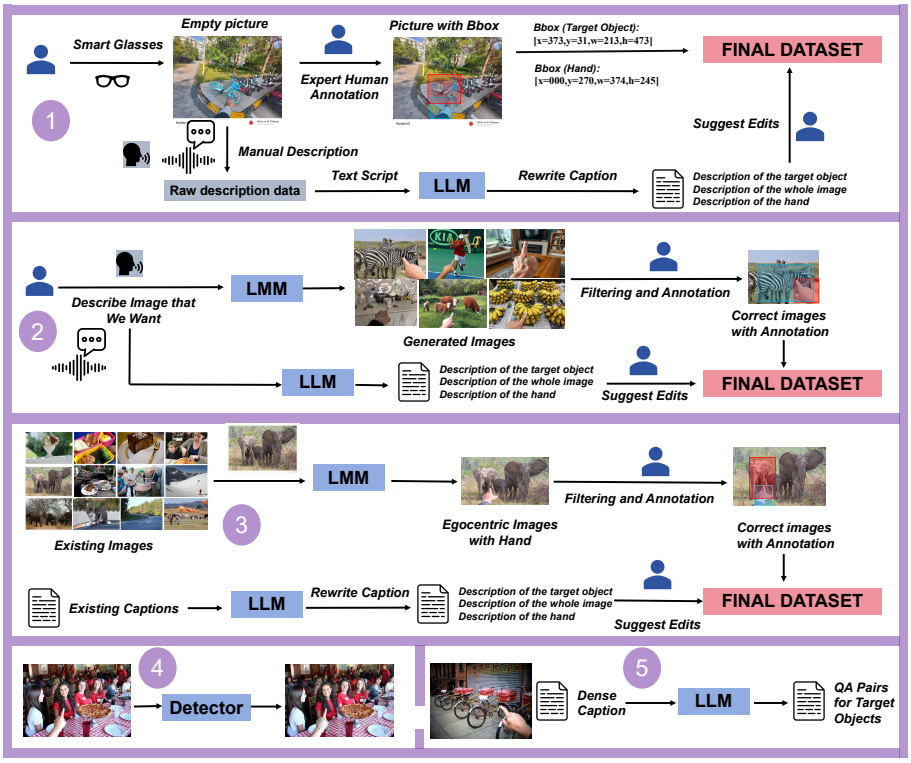


Fig. 4: Overview of the Ego-Point Ground Dataset Collection and Multi-Stage Annotation Pipeline. Stages 1–3 illustrate the raw data acquisition and fundamental annotation workflow. Stage 4 details the extraction of 2D keypoint pose information. Subsequently, Stage 5 focuses on generating high-quality Visual Question-Answering (VQA) pairs for the target objects.

Caption Semantic Diversity Lexical Diversity: Figure 2(c) shows that the dataset emphasizes interactive elements. High-frequency terms include common entities (e.g., "object", "person"), spatial indicators (e.g., "frame", "environment"), and dynamic interaction words (e.g., "pointing", "hand") paired with spatial terms (e.g., "right", "left"). This highlights the dataset's focus on context and dynamic egocentric behaviors.

Semantic Uniformity: To assess semantic diversity, we employ t-SNE [51] to project caption vector representations $t_i \in \mathbb{R}^d$ into a 2D scatter plot, as shown in Figure 3(b). The results demonstrate significant spatial dispersion for most category captions, confirming high semantic diversity. A few categories exhibit tighter clustering, indicating higher linguistic consistency. The dataset's descriptive richness is supported by its broad spatial and categorical diversity.

2.2 Data Collection Pipeline

The EgoPoint-Ground Dataset uses a Hybrid Data Generation Paradigm to blend real-world diversity with the efficiency of synthetic data, as shown in Figure 4. It

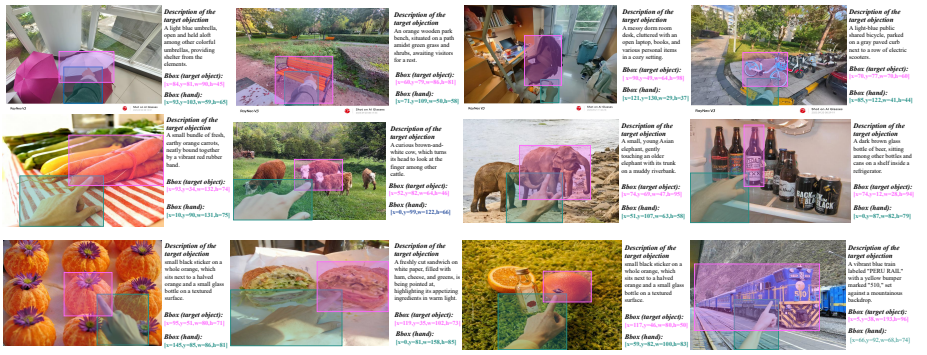


Fig. 5: Image Examples of Ego-Point Ground Dataset. We showcase example images and partial annotations from the dataset, including bounding boxes for the hand and the target object, as well as the corresponding caption for the target object. The first row specifically illustrates data collected from our real-world captures.

combines three main data sources, which are processed and annotated using a dedicated LMM-Aided Synthesis and Editing Pipeline [27, 83]. A partial example of the dataset is shown in Figure 5.

Real-World Data The first source ensures high fidelity by capturing real-world scenarios, We demonstrate this process in Stage 1 of Figure 4. Researchers wearing smart glasses recorded egocentric images of hands pointing at objects in various environments (e.g., schools, supermarkets, homes). Data collectors provided detailed descriptions via voice, which were transcribed and initially drafted by an LLM. Human annotators then reviewed and corrected the LLM-generated text, and manually annotated pixel-accurate bounding boxes for the target object and hand region. This process produced the core caption components: image description, object referent, and hand description.

Synthetic Generation The second component utilizes LMMs for directed image synthesis to efficiently generate challenging referential scenarios that are difficult to capture in the real world, such as complex occlusions or specific lighting conditions. The data collection procedure for this section is shown in Stage 2 of Figure 4. Researchers manually filtered the synthetic Egocentric images to remove unqualified samples. For captioning, participants provided voice descriptions, which were transcribed by an LLM and then subjected to rigorous human review and correction. Target bboxes were manually labeled for precise localization.

Image Editing To efficiently scale the dataset and facilitate domain migration, we employed LMM-aided image editing, as shown in Figure 4, stage 3. We selected third-person perspective images from existing datasets [50, 86] and guided a Multimodal Large Model (LMM) with Egocentric-specific prompts to edit them into egocentric scenes featuring hand pointing. The edited images underwent stringent human screening and secondary annotation to correct LMM-induced pointing inaccuracies. An LLM concurrently structured the captions into three parts: image description, object referent, and hand description.

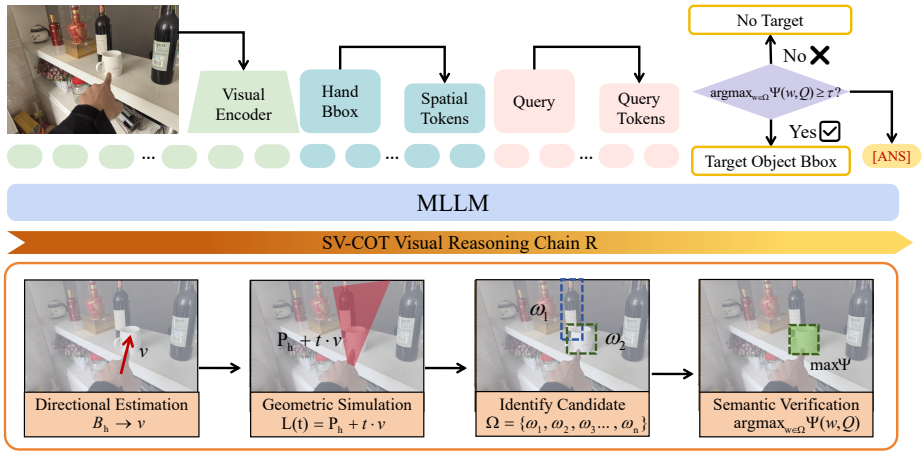


Fig. 6: Architectural overview of the SV-CoT framework. The egocentric hand bounding box $\mathcal{B}^{2D}_{hand}$ is aligned into discrete spatial anchors $\mathcal{T}_{pos}$. Zero-shot grounding is reformulated as a latent reasoning chain $\mathcal{R}$ that sequentially infers: (1) a deictic directional primitive $\boldsymbol{v}$; (2) a virtual ray $\mathcal{L}(t)$ to spatially prune distractors into a candidate set Ω; and (3) a multi-modal semantic verification score Ψ to either resolve the target $BBox_{\mathcal{O}}$ or explicitly reject it via threshold τ.

Auxiliary Annotation and Quality Control To enrich the dataset's dimensions and potential applications, we generated auxiliary annotations (Stage 4 and Stage 5 in Figure 4). An advanced pose estimation model was used to process hand regions in the images, extracting keypoints for Hand Pose annotations, which were refined through human review. Simultaneously, we leveraged the fine-grained captions and an LLM to construct challenging referential QA Pairs.

2.3 Ethical Considerations

The EgoPoint-Ground dataset was collected following strict ethical and privacy guidelines. Participants gave written consent, understanding the data's use and their privacy rights. All images and metadata were de-identified to protect Personally Identifiable Information.

Access is limited to certified researchers through a formal licensing system, with clear usage rules and penalties for violations. We ensure full compliance with ethical standards and prioritize participant privacy.

3 Task Setting

In this section, we establish a comprehensive benchmark for Egocentric Deictic Visual Grounding. By introducing high-quality annotations alongside a robust baseline framework, we provide a solid groundwork to catalyze future research in this novel domain. Leveraging the newly constructed EgoPoint-Ground dataset, we formally define the following evaluation tasks:

Table 2: Quantitative results of state-of-the-art models on the EDG task, evaluated across both Hybrid (mixed synthetic and captured) and Real-World (naturally captured) datasets. Shaded rows categorize the models into distinct performance tiers, while the Human Study establishes the theoretical upper bound.

Model	EDG — Hybrid			EDG — Real-World		
	$\mathcal{P}$@0.3 ↑	$\mathcal{P}$@0.5 ↑	$\mathcal{P}$@0.7 ↑	$\mathcal{P}$@0.3 ↑	$\mathcal{P}$@0.5 ↑	$\mathcal{P}$@0.7 ↑
Qwen3-VL-4B [83]	0.743	0.702	0.622	0.747	0.699	0.616
Qwen2.5-VL-3B [75]	0.731	0.671	0.592	0.738	0.632	0.488
Qwen3-VL-8B [83]	0.726	0.676	0.598	0.789	0.701	0.593
Qwen2.5-VL-7B [75]	0.721	0.653	0.563	0.776	0.688	0.533
LLaVA-OneVision-4B [3]	0.679	0.475	0.145	0.654	0.431	0.129
InternVL-3.5-8B [76]	0.554	0.413	0.270	0.592	0.427	0.214
GroundingGPT [49]	0.447	0.358	0.219	0.354	0.259	0.127
DeepSeek-VL2 [78]	0.344	0.293	0.259	0.364	0.296	0.245
LLaVA-v1.5-13B [53]	0.338	0.198	0.079	0.274	0.128	0.042
Ferret-7B [85]	0.282	0.221	0.181	0.258	0.203	0.160
Florence2-Large [79]	0.241	0.188	0.149	0.234	0.193	0.153
Florence2-Base [79]	0.149	0.110	0.080	0.141	0.121	0.075
GPT-5.4-mini	0.625	0.429	0.204	0.578	0.267	0.100
gemini-3-flash-preview	0.681	0.548	0.344	0.727	0.591	0.364
doubao-seed-2.0-lite	0.718	0.688	0.641	0.875	0.847	0.819
qwen3.6-flash	0.793	0.725	0.624	0.822	0.745	0.667
SV-CoT [Qwen3-vl] (Ours)	**0.795**	**0.756**	**0.718**	**0.820**	**0.818**	**0.727**
Human Study	**0.979**	**0.914**	**0.846**	**0.994**	**0.918**	**0.874**

3.1 Egocentric Deictic Visual Grounding (EDG)

Motivation: In real-world embodied interactions, deictic gestures and natural language exhibit strong synergistic complementarity. While physical pointing provides an immediate spatial prior, linguistic descriptions supply essential semantic attributes. Because neither modality is independently sufficient to resolve complex spatial or semantic ambiguities, we formulate the Egocentric Deictic Grounding (EDG) task. This benchmark challenges models to seamlessly integrate static hand-pointing cues with linguistic context to achieve robust cross-modal localization.

Task Definition: Formally, the EDG task requires the precise spatial localization of a target object $\mathcal{O}$ through the fusion of visual and linguistic information. The input consists of an egocentric image $\mathcal{I}$, the 2D bounding box of the pointing hand $\mathcal{B}^{2D}_{hand}$, and an underspecified linguistic referent $\mathcal{D}_{\mathcal{O}}$ (e.g., "this object" or "that boy"). Notably, $\mathcal{D}_{\mathcal{O}}$ identifies the target's category or intent but lacks

Table 3: Performance of state-of-the-art models on the D-VQA task across **Hybrid** (mixed synthetic/captured) and **Real-World** (exclusively captured) datasets. The "Human Study" denotes the evaluation results from human participants.

Model	D-VQA (Hybrid)			D-VQA (Real-World)		
	Acc.↑	**M-F1↑**	**Rec.↑**	**Acc.↑**	**M-F1↑**	**Rec.↑**
LLaVA-OneVision-8B [3]	0.531	0.519	0.548	0.544	0.525	0.526
LLaVA-OneVision-4B [3]	0.510	0.535	0.547	0.521	0.476	0.487
Qwen3-VL-8B [83]	0.494	0.484	0.490	0.517	0.510	0.503
Qwen3-VL-4B [83]	0.471	0.488	0.495	0.456	0.507	0.489
Qwen2.5-VL-7B [75]	0.457	0.427	0.445	0.444	0.428	0.411
Qwen2.5-VL-3B [75]	0.456	0.461	0.470	0.441	0.527	0.523
DeepSeek-VL2-Small [78]	0.413	0.459	0.455	0.457	0.523	0.506
DeepSeek-VL2 [78]	0.412	0.461	0.473	0.393	0.397	0.390
InternVL-3.5-4B [76]	0.445	0.461	0.445	0.433	0.368	0.367
InternVL-3.5-8B [76]	0.316	0.392	0.381	0.286	0.398	0.380
GroundingGPT [49]	0.334	0.363	0.364	0.392	0.352	0.350
LLaVA-v1.5-13B [53]	0.267	0.249	0.245	0.147	0.150	0.137
Human Study	**0.811**	**0.631**	**0.653**	**0.877**	**0.756**	**0.753**

unique spatial discriminability. Input: $(\mathcal{I}, \mathcal{B}^{2D}_{hand}, \mathcal{D}_{\mathcal{O}})$; Output: The 2D bounding box of the target object, $BBox_{\mathcal{O}}$.

3.2 Deictic Visual Question Answering (D-VQA)

Motivation: The ultimate goal of advanced embodied agents is to achieve meaningful, situated interaction and manipulation. Following successful localization, high-level reasoning about the target object in the environment is often required to determine subsequent actions and intent. The D-VQA task aims to evaluate a model's ability to perform high-level Situated Reasoning on the target object after successfully parsing the hand pointing intent and localization, thereby supporting decision-making and action planning.

Task Definition: Given the image $\mathcal{I}$, the hand's 2D bbox $\mathcal{B}^{2D}_{hand}$, and a specific question about the referred object $\mathcal{Q}$ (eg. "What is this?", "What is the boy wearing? "), the model must provide the answer. Input: $(\mathcal{I}, \mathcal{B}^{2D}_{hand}, \mathcal{Q})$; Output: The answer to the question, $\mathcal{A}$.

3.3 Language-Only Grounding(D-REC)

Motivation: The Language-Only Grounding task follows the paradigm of traditional Referring Expression Comprehension (REC) but is specifically tailored for egocentric deictic scenarios. The motivation is twofold: first, to quantify the model's independent capability to interpret referential linguistic cues (e.g., "this", "this boy") within a first-person visual context; and second, to serve as a rigorous

ablation baseline for evaluating the performance gain of multi-modal fusion (as in EDG) over a purely linguistic-driven approach on the same dataset.
Task Definition: Input: $(\mathcal{I}, \mathcal{D}_{\mathcal{O}})$; Output: $BBox_{\mathcal{O}}$.

3.4 Physical Pointing-Only Grounding (POG)

Motivation: This task aims to quantify the independent referencing capability of the hand-pointing pose. By providing only the 2D hand bbox, we evaluate the model's ability to infer spatial pointing intent purely from the static hand configuration. This forms another EDG ablation baseline.
Task Definition: Input: $(\mathcal{I}, \mathcal{B}^{2D}_{hand})$; Output: $BBox_{\mathcal{O}}$.

4 Experiment

4.1 Baseline Method

To bridge the granularity gap between continuous visual coordinates and the discrete linguistic vocabulary of MLLMs, we adopt a coordinate-to-token alignment protocol. Given an egocentric image $\mathcal{I}$ and the referrer's hand bounding box $\mathcal{B}^{2D}_{hand} = [x, y, w, h]$, we apply a spatial quantization function $\phi(\cdot)$ to map each coordinate into N discrete spatial bins (where $N = 1000$):

$$\mathcal{T}_{pos} = \left\{ \langle \text{bin}_c \rangle \mid c \in \mathcal{B}^{2D}_{hand}, \phi(c) = \left\lfloor \frac{c}{L} \cdot (N-1) \right\rfloor \right\} \tag{2}$$

These quantized tokens $\mathcal{T}_{pos}$ serve as **Spatial Anchors**, grounding the LLM's attention to the specific ego-gestural region within the multimodal input stream.

SV-CoT: Spatial-Reasoning-based Visual Chain-of-Thought Instead of treating grounding as a direct regression task, we reformulate it as a structured inference process. We present **SV-CoT**, a zero-shot inference framework instantiated on Qwen3-VL [83]. As depicted in Figure 6, this baseline systematically decomposes the complex deictic intent into a sequential latent reasoning chain $\mathcal{R}$. Formally, the target distribution $P(BBox_{\mathcal{O}} \mid \mathcal{I}, \mathcal{B}^{2D}_{hand}, \mathcal{Q})$ is modeled by marginalizing over a sequence of geometric rationales:

$$P(BBox_{\mathcal{O}} \mid \mathcal{X}_{in}) = \sum_{\mathcal{R}} P(BBox_{\mathcal{O}} \mid \mathcal{R}, \mathcal{X}_{in}) P(\mathcal{R} \mid \mathcal{X}_{in}) \tag{3}$$

where $\mathcal{X}_{in} = (\mathcal{I}, \mathcal{B}^{2D}_{hand}, \mathcal{Q})$ denotes the multimodal input context. The reasoning variable $\mathcal{R} = \{r_{dir}, r_{path}, r_{ref}\}$ is instantiated as a causal chain of linguistic tokens, mimicking human-like spatial deduction:

1. Deictic Intent Parsing (r_{dir}): The model first performs an *anatomical-to-spatial* projection. By analyzing the orientation of the hand within the anchor $\mathcal{B}^{2D}_{hand}$, the MLLM instantiates a directional primitive $\boldsymbol{v} \in \mathbb{R}^2$, effectively transforming local gesture posture into a global pointing prior.

2. Geometric Trajectory Simulation (r_{path}): This primitive defines a virtual ray $\mathcal{L}(t) = \mathcal{P}_{hand} + t \cdot \boldsymbol{v}$, where $\mathcal{P}_{hand}$ denotes the spatial centroid of $\mathcal{B}^{2D}_{hand}$. The model executes a **spatial pruning** operation, identifying a candidate set $\Omega = \{\omega_1, \ldots, \omega_n\}$ of visual regions that satisfy the intersection constraint $\omega_i \cap \mathcal{L}(t) \neq \emptyset$. This effectively filters spatial distractors from cluttered egocentric scenes.

3. Spatio-semantic Verification (r_{ref}): The final phase involves a multi-modal alignment check. The model resolves the target $BBox_{\mathcal{O}}$ by evaluating the semantic consistency Ψ between the linguistic query $\mathcal{Q}$ and the pruned candidates Ω:

$$BBox_{\mathcal{O}} = \begin{cases} \arg\max_{\omega \in \Omega} \Psi(\omega, \mathcal{Q}) & \text{if } \max_{\omega \in \Omega} \Psi(\omega, \mathcal{Q}) \geq \tau \\ \varnothing & \text{otherwise} \end{cases} \tag{4}$$

where τ is a confidence threshold for **explicit rejection**. By parameterizing the zero-shot inference as $\mathcal{Y} = [\mathcal{R}, BBox_{\mathcal{O}}]$, SV-CoT enforces a structural constraint that mitigates spatial hallucination and ensures referential consistency.

4.2 Evaluation Metrics

We evaluate the proposed benchmarks using the following metrics:
Localization Metrics: For EDG, D-REC, and POG tasks, we adopt Precision@τ as the core metric, which measures the proportion of samples where the Intersection over Union (IoU) between the predicted and ground-truth bounding boxes exceeds threshold τ. We report results at $\tau \in \{0.3, 0.5, 0.7\}$ to reflect varying localization precision.
Classification Metrics: For the D-VQA task, we employ Accuracy (Acc.), Macro F1-score, and Macro Recall to assess the correctness of answers. Macro-averaging is specifically used to ensure a fair evaluation across the class-imbalanced distribution of our dataset.

4.3 Experimental Results

Results on Egocentric Deictic Grounding We conduct a comprehensive evaluation of state-of-the-art MLLMs and VG models on the EDG task across both hybrid and real-world scenarios (see Table 2). The results reveal a distinct performance hierarchy: the Qwen3 and Qwen2.5 series lead the field due to their superior vision-language alignment (e.g., Qwen3-VL-8B achieves $\mathcal{P}$@0.5 of 0.701 on real-world), whereas foundational models like Florence2 and LLaVA-v1.5 struggle, with accuracies consistently below 0.34. Notably, while baseline models exhibit a sharp performance decline at higher precision thresholds ($\mathcal{P}$@0.7), highlighting the significant challenge of fine-grained gesture-to-object localization, our method demonstrates remarkable resilience. Specifically, a substantial gap of over 26% persists between the top-performing baseline and the **Human Study** upper bound ($\mathcal{P}$@0.5 $\geq$ 0.914), suggesting that general-purpose MLLMs have yet to master complex "hand-object" spatial logic. Our proposed baseline, **SV-CoT**, explicitly guides spatial reasoning via a Visual Chain-of-Thought, achieving an impressive absolute improvement of 11.7% (0.818 vs. 0.701) in real-world and 5.4% (0.756 vs. 0.702) in hybrid scenarios at $\mathcal{P}$@0.5 over the strongest open-source

Table 4: Performance comparison on conventional REC and dedicated POG tasks. Absolute scores are reported with relative gains/losses compared to the baseline (Inputs are provided without either spatial or linguistic cues) in parentheses.

Model	D-REC			POG		
	$\mathcal{P}$@0.3 ↑	$\mathcal{P}$@0.5 ↑	$\mathcal{P}$@0.7 ↑	$\mathcal{P}$@0.3 ↑	$\mathcal{P}$@0.5 ↑	$\mathcal{P}$@0.7 ↑
Qwen3-VL-4B [83]	0.737 (+0.087)	0.684 (+0.118)	0.605 (+0.123)	0.671 (+0.021)	0.612 (+0.046)	0.532 (+0.050)
Qwen3-VL-8B [83]	0.701 (+0.111)	0.648 (+0.087)	0.561 (+0.077)	0.653 (+0.063)	0.595 (+0.034)	0.514 (+0.030)
Qwen2.5-VL-7B [75]	0.703 (+0.190)	0.638 (+0.160)	0.541 (+0.125)	0.670 (+0.157)	0.611 (+0.143)	0.515 (+0.099)
Qwen2.5-VL-3B [75]	0.665 (+0.169)	0.592 (+0.167)	0.491 (+0.139)	0.575 (+0.079)	0.501 (+0.076)	0.407 (+0.055)
LLaVA-OneVision-4B [3]	0.586 (+0.194)	0.347 (+0.186)	0.088 (+0.071)	0.525 (+0.133)	0.292 (+0.131)	0.065 (+0.048)
LLaVA-OneVision-8B [3]	0.505 (+0.167)	0.258 (+0.110)	0.049 (+0.027)	0.403 (+0.065)	0.181 (+0.033)	0.025 (+0.003)
InternVL-3.5-4B [76]	0.535 (+0.149)	0.402 (+0.135)	0.272 (+0.092)	0.416 (+0.030)	0.305 (+0.038)	0.203 (+0.023)
InternVL-3.5-8B [76]	0.456 (+0.066)	0.325 (+0.045)	0.210 (+0.006)	0.497 (+0.107)	0.311 (+0.031)	0.178 (-0.026)
DeepSeek-VL2-Small [78]	0.443 (-0.001)	0.393 (-0.005)	0.341 (+0.023)	0.375 (-0.069)	0.318 (-0.080)	0.274 (-0.044)
Ferret-7B [85]	0.428 (+0.029)	0.350 (+0.015)	0.261 (-0.001)	0.413 (+0.014)	0.353 (+0.018)	0.281 (+0.019)
GroundingGPT [49]	0.378 (-0.149)	0.261 (-0.169)	0.129 (-0.162)	0.536 (+0.009)	0.443 (+0.013)	0.305 (+0.014)
SV-CoT [Qwen3-vl] (Ours)	**0.814**	**0.750**	**0.694**	**0.694**	**0.639**	**0.556**
Human Study	**0.952**	**0.926**	**0.819**	**0.956**	**0.916**	**0.806**

baselines. Furthermore, SV-CoT effectively maintains a high accuracy of over 0.71 even at the strict $\mathcal{P}$@0.7 threshold, significantly mitigating semantic ambiguity in deictic references. Interestingly, most baseline models perform slightly better in real-world than in hybrid scenarios. This discrepancy arises because the hybrid set serves as a rigorous stress test, incorporating a higher density of synthetic noise, extreme viewpoint augmentations, and occlusions that surpass the distributional complexity of natural captures.

Results on Deictic Visual Question Answering The D-VQA task requires models to establish deep semantic correlations between hand-pointing gestures and specific linguistic queries within complex visual scenes. As shown in Table 3, LLaVA-OneVision and the Qwen3-VL series demonstrate competitive per-

formance, with LLaVA-OneVision-8B achieving the highest accuracy of 0.544 in real-world scenarios. In contrast, foundational or smaller-scale models (e.g., LLaVA-v1.5) struggle with such gesture-based visual reasoning.

Despite progress by top-tier models, a performance gap of approximately 30% remains relative to the **Human Study** upper bound (Acc. > 0.81). This gap underscores the inherent difficulty for general-purpose MLLMs in precisely mapping spatial coordinates (Hand Bbox) to semantic question-answering, a challenge we define as deictic alignment failure.

Ablation Studies: Linguistic vs. Geometric Cues D-REC and POG evaluate referential understanding using isolated linguistic or physical cues (Table 4). Compared to the multimodal EDG task, all models show significant performance drops, confirming the severe referential ambiguity of unimodal inputs. While Qwen3 maintains a robust $\mathcal{P}$@0.5 of 0.684 in D-REC, POG results expose a critical deficiency in geometric understanding. Models generally perform worse on POG, collapsing at high-precision thresholds (e.g., LLaVA-OneVision-8B drops to 0.025 at $\mathcal{P}$@0.7). This indicates that current MLLMs fail to map the spatial trajectory from "fingertip" to "target" without linguistic aid, revealing a massive gap in physical intent comprehension compared to human performance.

Our baseline, **SV-CoT**, effectively mitigates these restrictions via a Visual Chain-of-Thought. It achieves 0.639 at $\mathcal{P}$@0.5 in POG (an absolute gain of 2.7% over the strongest baseline) and 0.750 in D-REC (a 6.6% gain). This demonstrates that explicit gesture-object spatial modeling effectively overcomes the limitations of isolated geometric or linguistic perception.

5 Conclusion

In this work, we formalized the egocentric deictic visual grounding task, directly addressing the inherent inadequacies of text-only queries in resolving physical interaction ambiguities. Our primary contribution is EgoPoint-Ground, a pioneering large-scale multimodal dataset comprising over 15,000 densely annotated samples that capture hand-gesture and linguistic co-reference. Upon this foundation, we established a rigorous benchmark evaluating a wide spectrum of state-of-the-art MLLMs and visual grounding architectures. Furthermore, we introduced SV-CoT, a robust baseline framework that synergizes gestural primitives and linguistic context through a Visual Chain-of-Thought reasoning mechanism. We will open-source our dataset and models to advance research in embodied AI, egocentric vision, and multimodal interaction.

Limitations and Future Work While EgoPoint-Ground provides a robust baseline, it currently focuses on canonical index-finger pointing in static images, limiting the exploration of complex gestures and continuous spatiotemporal trajectories. Furthermore, extreme cases involving severe occlusion or highly distal targets remain challenging. To capture dynamic pointing intents, future work will expand this paradigm into the video domain.

Acknowledgments This work was supported in part by Beijing Municipal Science & Technology Commission No. Z251100004625090, by the National Science Foundation of China (NSFC) under Grant No. 62176134, and by a grant from the Assisted Medical Consultation Project Based on DeepSeek.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 Technical Report. ArXiv Preprint ArXiv:2303.08774 (2023)
2. Ahuja, C.: Communication Beyond Words: Grounding Visual Body Motion with Language. Ph.D. thesis, Carnegie Mellon University (2022)
3. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Wu, C., et al.: Llava-onevision-1.5: Fully Open Framework for Democratized Multimodal Training. ArXiv Preprint ArXiv:2509.23661 (2025)
4. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3674–3683 (2018)
5. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)
6. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen Technical Report. ArXiv Preprint ArXiv:2309.16609 (2023)
7. Bansal, S., Arora, C., Jawahar, C.: My View is the Best View: Procedure Learning from Egocentric Videos. In: Proceedings of the European Conference on Computer Vision. pp. 657–675 (2022)
8. Bhattacharyya, P., Mitton, J., Page, R., Morgan, O., Menzies, B., Homewood, G., Jacobs, K., Baesso, P., Trickett, D., Mair, C., et al.: Helios: An extremely low power event-based gesture recognition for always-on smart eyewear. In: European Conference on Computer Vision. pp. 168–184. Springer (2024)
9. Chandel, H., Vatta, S.: Occlusion Detection and Handling: A Review. International Journal of Computer Applications **120** (2015)
10. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. ArXiv Preprint ArXiv:2306.15195 (2023)
11. Chen, W., Chen, L., Wu, Y.: An Efficient and Effective Transformer Decoder-Based Framework for Multi-Task Visual Grounding. In: European Conference on Computer Vision. pp. 125–141. Springer (2024)
12. Chen, Y., Li, Q., Kong, D., Kei, Y.L., Zhu, S.C., Gao, T., Zhu, Y., Huang, S.: YouRefIt: Embodied Reference Understanding With Language and Gesture. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1385–1395 (2021)
13. Chen, Z., Wang, P., Ma, L., Wong, K.Y.K., Wu, Q.: Cops-Ref: A New Dataset and Task on Compositional Referring Expression Comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10086–10095 (2020)
14. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: YOLO-World: Real-Time Open-Vocabulary Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16901–16911 (2024)
15. Chuang, Y.S., Li, Y., Wang, D., Yeh, C.F., Lyu, K., Raghavendra, R., Glass, J., Huang, L., Weston, J., Zettlemoyer, L., et al.: Meta CLIP 2: A Worldwide Scaling Recipe. ArXiv Preprint ArXiv:2507.22062 (2025)

16. Constantin, S., Eyiokur, F.I., Yaman, D., Bärmann, L., Waibel, A.: Interactive Multimodal Robot Dialog Using Pointing Gesture Recognition. In: European conference on computer vision. pp. 640–657. Springer (2022)
17. Dai, M., Yang, L., Xu, Y., Feng, Z., Yang, W.: SimVG: A Simple Framework for Visual Grounding with Decoupled Multi-modal Fusion. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 121670–121698 (2024)
18. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: GuessWhat?! Visual Object Discovery Through Multi-Modal Dialogue. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5503–5512 (2017)
19. Deichler, A., O'Regan, J., Beskow, J.: MM-Conv: A Multi-modal Conversational Dataset for Virtual Humans. arXiv preprint arXiv:2410.00253 (2024)
20. Delmas, G., Weinzaepfel, P., Moreno-Noguer, F., Rogez, G.: PoseEmbroider: Towards a 3D, Visual, Semantic-aware Human Pose Representation. In: European Conference on Computer Vision. pp. 55–73. Springer (2024)
21. Duan, J., Yu, S., Tan, H.L., Zhu, H., Tan, C.: A Survey of Embodied AI: From Simulators to Research Tasks. IEEE Transactions on Emerging Topics in Computational Intelligence **6**, 230–244 (2022)
22. Efron, R.: What is Perception? In: Proceedings of the Boston Colloquium for the Philosophy of Science 1966/1968. pp. 137–173 (1969)
23. Fan, Z., Ohkawa, T., Yang, L., Lin, N., Zhou, Z., Zhou, S., Liang, J., Gao, Z., Zhang, X., Zhang, X., et al.: Benchmarks and Challenges in Pose Estimation for Egocentric Hand Interactions with Objects. In: European Conference on Computer Vision. pp. 428–448. Springer (2024)
24. Farkhondeh, A., Tafasca, S., Odobez, J.M.: ChildPlay-Hand: A Dataset of Hand Manipulations in the Wild. In: European Conference on Computer Vision. pp. 101–118. Springer (2024)
25. Fortuny, J., Payrató, L.: Ambiguity in Linguistics. Studia Linguistica **78**(1), 1–7 (2024)
26. Foteinos, K., Angelidis, G., Psiris, A., Argyriou, V., Sarigiannidis, P., Papadopoulos, G.T.: FR-GESTURE: An RGBD Dataset For Gesture-based Human-Robot Interaction In First Responder Operations. arXiv preprint arXiv:2602.17573 (2026)
27. Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., et al.: Seedream 3.0 Technical Report. ArXiv Preprint ArXiv:2504.11346 (2025)
28. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4D: Around the World in 3,000 Hours of Egocentric Video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
29. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
30. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl Technical Report. ArXiv Preprint ArXiv:2505.07062 (2025)
31. Guo, L., Lu, Z., Yao, L.: Human-Machine Interaction Sensing Technology Based on Hand Gesture Recognition: A Review. IEEE Transactions on Human-Machine Systems **51**, 300–309 (2021)
32. Hashi, A.O., Hashim, S.Z.M., Asamah, A.B.: A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024. IEEE Access **12**, 143599–143626 (2024)

33. Hong Da, N.L., Miura, J., Hayashi, K.: Enhancing Human-Robot Collaborative Search Through Efficient Space Sharing with On-Demand Bidirectional Interaction. In: European Conference on Computer Vision. pp. 20–35. Springer (2024)
34. Huang, Z., Satoh, S.: LoA-Trans: Enhancing Visual Grounding by Location-Aware Transformers. In: European Conference on Computer Vision. pp. 405–421. Springer (2024)
35. Hummel, T., Karthik, S., Georgescu, M.I., Akata, Z.: EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval. In: European Conference on Computer Vision. pp. 1–17. Springer (2024)
36. Jiang, H., Lu, Z.: Visual Grounding for Object-Level Generalization in Reinforcement Learning. In: European Conference on Computer Vision. pp. 55–72. Springer (2024)
37. Jirak, D., Biertimpel, D., Kerzel, M., Wermter, S.: Solving Visual Object Ambiguities when Pointing: An Unsupervised Learning Approach. Neural Computing and Applications **33**(7), 2297–2319 (2021)
38. Ju, Y., Hu, K., Zhang, G., Zhang, G., Jiang, M., Xu, H.: Robo-ABC: Affordance Generalization Beyond Categories via Semantic Correspondence for Robot Manipulation. In: European Conference on Computer Vision. pp. 222–239. Springer (2024)
39. Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your Large Vision-Language Model Only Needs A Few Attention Heads For Visual Grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9339–9350 (2025)
40. Kang, W., Liu, G., Shah, M., Yan, Y.: SegVG: Transferring Object Bounding Box to Segmentation for Visual Grounding. In: European Conference on Computer Vision. pp. 57–75. Springer (2024)
41. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to Objects in Photographs of Natural Scenes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 787–798 (2014)
42. Kusnanti, E.A., Sierra, E., Putra, G.G.S., Cahyadi, E.S., Haq, A., Purwitasari, D.: Indonesian Lexical Ambiguity in Machine Translation: A Literature Review. In: 2024 International Conference on Information Technology Research and Innovation (ICITRI). pp. 59–64. IEEE (2024)
43. Lee, J., Wang, J., Brown, E., Chu, L., Rodriguez, S.S., Froehlich, J.E.: GazePointAR: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. pp. 1–20 (2024)
44. Li, H., Wang, Y., Hao, W., Zhang, P., Wang, D., Lu, H.: RAGTrack: Language-aware RGBT Tracking with Retrieval-Augmented Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28179–28189 (2026)
45. Li, H., Wang, Y., Hu, X., Hao, W., Zhang, P., Wang, D., Lu, H.: CADTrack: Learning Contextual Aggregation with Deformable Alignment for Robust RGBT Tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6109–6117 (2026)
46. Li, L., Chen, C., Wang, Y., Lyu, J., Chang, K., Chen, Y., Deng, Z.: From sparse to dense: Spatio-temporal fusion for multi-view 3d human pose estimation with densewarper. arXiv preprint arXiv:2605.14525 (2026)
47. Li, L., Gu, R., Wang, C., Xing, J., Yu, X., Zhang, X.P.: Multi-view 3d human pose estimation with weakly synchronized images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4833–4841 (2025)

48. Li, L., Yang, W., Yu, X., Xing, J., Zhang, X.P.: Translating motion to notation: Hand labanotation for intuitive and comprehensive hand movement documentation. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 4092–4100 (2024)
49. Li, Z., Xu, Q., Zhang, D., Song, H., Cai, Y., Qi, Q., Zhou, R., Pan, J., Li, Z., Tu, V., et al.: GroundingGPT: Language Enhanced Multi-modal Grounding Model. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. pp. 6657–6678 (2024)
50. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
51. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., Kluger, Y.: Efficient Algorithms for t-distributed Stochastic Neighborhood Embedding. ArXiv Preprint ArXiv:1712.09005 (2017)
52. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: DeepSeek-V3 Technical Report. ArXiv Preprint ArXiv:2412.19437 (2024)
53. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
54. Liu, J., Chen, Q., Wang, Z., Tang, Y., Zhang, Y., Yan, C., Wang, D., Li, X., Zhao, B.: AerialVG: A Challenging Benchmark for Aerial Visual Grounding by Exploring Positional Relations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5177–5187 (2025)
55. Liu, R., Liu, C., Bai, Y., Yuille, A.L.: CLEVR-Ref+: Diagnosing Visual Reasoning With Referring Expressions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4185–4194 (2019)
56. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In: Proceedings of the European Conference on Computer Vision. pp. 38–55 (2024)
57. Liu, Y., Chen, W., Bai, Y., Liang, X., Li, G., Gao, W., Lin, L.: Aligning Cyber Space With Physical World: A Comprehensive Survey on Embodied AI. IEEE/ASME Transactions on Mechatronics pp. 1–22 (2025)
58. Magay, A., Tripathi, D., Hao, Y., Fang, Y.: A Light and Smart Wearable Platform with Multimodal Foundation Model for Enhanced Spatial Reasoning in People with Blindness and Low Vision. In: European Conference on Computer Vision. pp. 323–339. Springer (2024)
59. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and Comprehension of Unambiguous Object Descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11–20 (2016)
60. Memmi, M.Y., Slama, R., Berretti, S.: Hand Gesture Recognition Using Dual Graph Hierarchical Edges Representation and Graph Transformer Network. In: European Conference on Computer Vision. pp. 53–68. Springer (2024)
61. Mo, M., Wang, J., Wang, L., Chen, H., Gu, C., Leng, J., Gao, X.: NexusAD: Exploring the Nexus for Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving. In: ECCV 2024 Workshop on Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving (2024)
62. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling Context between Objects for Referring Expression Understanding. In: Proceedings of the European Conference on Computer Vision. pp. 792–807 (2016)

63. Pfeifer, R., Iida, F.: Embodied Artificial Intelligence: Trends and Challenges. Lecture Notes in Computer Science pp. 1–26 (2004)
64. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2641–2649 (2015)
65. Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L., Ma, K.: ShapeLLM: Universal 3D Object Understanding for Embodied Interaction. In: European Conference on Computer Vision. pp. 214–238. Springer (2024)
66. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the International Conference on Machine Learning. pp. 8748–8763 (2021)
67. Schauerte, B., Fink, G.A.: Focusing Computational Visual Attention in Multi-Modal Human-Robot Interaction. In: Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction. pp. 1–8 (2010)
68. Schauerte, B., Richarz, J., Fink, G.A.: Saliency-Based Identification and Recognition of Pointed-At Objects. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4638–4643 (2010)
69. Shi, C., Yang, S.: Spatial and Visual Perspective-Taking via View Rotation and Relation Reasoning for Embodied Reference Understanding. In: European Conference on Computer Vision. pp. 201–218. Springer (2022)
70. Shukla, D., Erkent, O., Piater, J.: Probabilistic Detection of Pointing Directions for Human-Robot Interaction. In: Proceedings of the 2015 International Conference on Digital Image Computing: Techniques and Applications. pp. 1–8 (2015)
71. Shukla, D., Erkent, Ö., Piater, J.: A Multi-View Hand Gesture RGB-D Dataset for Human-Robot Interaction Scenarios. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication. pp. 1084–1091 (2016)
72. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805 (2023)
73. Waisberg, E., Ong, J., Masalkhi, M., Zaman, N., Sarker, P., Lee, A.G., Tavakkoli, A.: Meta Smart Glasses—Large Language Models and the Future for Assistive Glasses for Individuals with Vision Impairments. Eye **38**, 1036–1038 (2024)
74. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A.: Voyager: An Open-Ended Embodied Agent with Large Language Models. ArXiv Preprint ArXiv:2305.16291 (2023)
75. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. ArXiv Preprint ArXiv:2409.12191 (2024)
76. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. ArXiv Preprint ArXiv:2508.18265 (2025)
77. Watanabe, K.: Reflexive Attentional Shift Caused by Indexical Pointing Gesture. Journal of Vision **2**(7), 435–435 (2002)
78. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv preprint arXiv:2412.10302 (2024)

79. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4818–4829 (2024)
80. Xiao, L., Yang, X., Lan, X., Wang, Y., Xu, C.: Towards Visual Grounding: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
81. Xiao, L., Yang, X., Peng, F., Wang, Y., Xu, C.: OneRef: Unified One-tower Expression Grounding and Segmentation with Mask Referring Modeling. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 139854–139885 (2024)
82. Xie, P., Chen, S., Dai, Y., Hu, D., Yang, K., Wang, G.: Target-Oriented Object Grasping via Multimodal Human Guidance. In: European Conference on Computer Vision. pp. 305–322. Springer (2024)
83. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 Technical Report. ArXiv Preprint ArXiv:2505.09388 (2025)
84. Yang, J., Ding, R., Brown, E., Qi, X., Xie, S.: V-IRL: Grounding Virtual Intelligence in Real Life. In: European conference on computer vision. pp. 36–55. Springer (2024)
85. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and Ground Anything Anywhere at Any Granularity. ArXiv Preprint ArXiv:2310.07704 (2023)
86. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling Context in Referring Expressions. In: Proceedings of the European Conference on Computer Vision. pp. 69–85 (2016)
87. Zhang, H., Niu, Y., Chang, S.F.: Grounding Referring Expressions in Images by Variational Context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4158–4166 (2018)
88. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Leizhang, Li, C., et al.: LLaVA-Grounding: Grounded Visual Chat with Large Multimodal Models. In: Proceedings of the European Conference on Computer Vision. pp. 19–35 (2024)
89. Zhuang, B., Wu, Q., Shen, C., Reid, I., Van Den Hengel, A.: Parallel Attention: A Unified Framework for Visual Object Discovery Through Dialogs and Queries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4252–4261 (2018)