Thinking from the Robot's View: The CoT-HRC Benchmark for Human Intent Reasoning in Embodied Collaboration

Chenxi Deng, Jianming Liu, Chao Xu, and Shaofei Chen*

College of Intelligence Science and Technology, National University of Defense Technology, Changsha, China
{dengchenxi20, liujianming, xuchao_21, chenshaofei01}@nudt.edu.cn

Abstract. Effective Human-Robot Collaboration (HRC) requires robots to accurately infer human intents under partial observations. However, existing benchmarks mostly rely on idealized views, overlooking the spatiotemporal incompleteness such as low-angle truncation and dynamic occlusions inherent in real-world quadruped robot perception. Furthermore, current evaluations lack diagnostic transparency, failing to distinguish whether Vision-Language Models (VLMs) perform reliable compensatory reasoning or merely hallucinate based on statistical biases when visual features are degraded. To bridge these dual gaps, we introduce CoT-HRC, a large-scale benchmark built in Habitat 3.0 that couples simulated robot-centric views with explicit hierarchical Chain-of-Thought (CoT) annotations. We propose a diagnostic protocol featuring the Step-wise Consistency Assessment (SCA_R) to penalize spurious accuracy by enforcing intermediate logical consistency, and the Conditional Reasoning Accuracy (CRA) to explicitly decouple models' underlying reasoning capabilities from visual perception failures. Extensive experiments across state-of-the-art VLMs reveal that while current models inherently possess strong physical commonsense, their reasoning chains are severely disrupted by the dynamic geometric constraints of egocentric views. Our fine-grained ablations further highlight critical bottlenecks in temporal aggregation and the compensatory role of semantic priors, establishing CoT-HRC as a vital stepping stone for robust embodied intent inference.The full benchmark is publicly available at https://github.com/Thus-cx/CoT-HRC.

Keywords: Human Intent Inference · HRC · Robot Egocentric View · VLM

1 Introduction

Embodied AI is shifting from isolated physical interaction toward complex social collaboration. Safe and efficient Human-Robot Collaboration (HRC) requires robots to infer intents under partial observability [36]. Recently, Vision-Language

* Corresponding author.

Models (VLMs) have emerged as a crucial pathway for such embodied reasoning tasks [25].

However, transitioning to real-world deployment reveals two significant gaps. First is the **visual perception gap**. In decentralized HRC, robots interleave task execution with brief observations to infer intents. While current evaluations assume unobstructed views [7, 12], real-world quadruped robots face low-angle, highly dynamic egocentric perspectives. This causes severe spatiotemporal incompleteness such as torso occlusions and target truncation [1], breaking traditional fine-grained perception paradigms [34].

Second is the **semantic reasoning gap**. Current benchmarks predominantly use end-to-end black-box schemes. Given fragmented visuals, VLMs easily hallucinate via language priors [10] or exploit statistical biases [22]. Lacking intermediate reasoning chains, these benchmarks lack diagnostic transparency, obscuring the true mechanisms behind model successes or failures [13].

To bridge these gaps, we propose **CoT-HRC**, a large-scale benchmark for human intent inference from constrained robot perspectives. Unlike datasets assuming ideal observations, we integrate realistic egocentric views with explicit logical attribution. This transparency provides a novel platform to investigate VLM abductive reasoning under incomplete spatiotemporal information.

Regarding implementation, we simulate the low-angle, occluded views of quadruped robots in Habitat to construct a partial observability dataset. Concurrently, we leverage privileged simulator data to automatically generate hierarchical CoT annotations, progressing from spatial anchoring to the final intent. For diagnostic evaluation, we replace opaque end-to-end testing with a Step-by-Step Reasoning QA protocol. By introducing metrics like Reasoning Consistency (SCA_R) and Conditional Reasoning Accuracy (CRA), we penalize spurious statistical biases, effectively decoupling visual perception failures from cognitive reasoning deficiencies.

The main contributions of this paper are summarized as follows:

(1) **Dataset Construction:** We build a large-scale HRC dataset in Habitat 3.0 that explicitly simulates the low-angle view and dynamic spatiotemporal incompleteness inherent to quadruped robots, addressing the realistic visual perception gap.

(2) **Benchmark & Metrics:** We introduce a step-by-step reasoning QA protocol based on hierarchical CoT annotations. We propose the Reasoning Consistency (SCA_R) and Conditional Reasoning Accuracy (CRA) to quantitatively decouple visual perception failures from logical reasoning capabilities.

(3) **Experimental Insights:** Extensive evaluations reveal that while VLMs possess strong physical commonsense, their performance severely degrades under egocentric views primarily due to visual truncation rather than reasoning deficits. Ablations further highlight critical bottlenecks in temporal aggregation and the importance of semantic priors.

2 Related Work

Visual Human Intent Inference. Current methods generally fall into two paradigms. Bottom-up approaches focus on fine-grained visual features, such as poses, interaction bounding boxes, or contact points [8, 14, 18, 30]. Top-down approaches leverage the commonsense priors of LLMs/VLMs for semantic-level abductive reasoning over long-horizon trajectories [3, 15, 24, 27, 39]. However, both fail under the highly dynamic, low-angle egocentric views of real-world robots. Bottom-up methods degrade sharply without clear hand-object boundaries [1,30], while top-down VLMs are prone to visual hallucinations or exploiting statistical biases when foundational visual cues are occluded [13, 27].

Benchmarks for Embodied AI. While traditional datasets like Charades [33] and ActivityNet [4] lack interactive embodied perspectives, simulation platforms [11,17,29,31,32] have spurred embodied benchmarks like BEHAVIOR-1K. Multi-view datasets like LEMMA [16] and Ego-Exo4D [12] capture collaboration but rely on ideal, unobstructed human-worn cameras. Although HARPER [1] and PARTNR [5] introduce quadruped perspectives or multi-agent planning, they lack the diagnostic intermediate reasoning labels required to evaluate VLM cognitive consistency under partial observability.

VLMs in Embodied AI. Recent VLMs [2, 21, 23, 37] and embodied models [9,28] show immense potential by aligning visual features with LLM semantic spaces. However, their strong performance heavily relies on idealized pre-training data such as HowTo100M [26] featuring clear targets. When confronted with the dynamic occlusions of robot-centric views, these models frequently fabricate visual elements rather than performing rigorous reasoning [13, 15, 22]. CoT-HRC addresses this by enforcing a reasoning constraint chain to strictly evaluate logical consistency and strip away false accuracy.

3 Problem Formulation

In the aforementioned long-horizon collaborative task scenarios, robots and humans typically execute their respective sub-tasks asynchronously. The robot does not observe the human continuously throughout; instead, after its own task completes, it intermittently observes the human to determine the next action. Assuming the human executes sub-task $\mathcal{L}$, the execution time of this task on the timeline is $[t_{start}, t_{end}]$. The moment the robot begins observation is defined as t_{obs}. Due to asynchronous collaboration, the placement of t_{obs} is random. Assuming the visual input obtained by the robot is an Observation Window with a duration of $\Delta t (0 < \Delta t < t_{end} - t_{obs})$, denoted as $\mathcal{W} = [t_{obs}, t_{obs} + \Delta t]$. Since the duration of $\mathcal{W}$ is usually less than the execution time of the sub-task, the robot sometimes misses the initial state or key interaction frames of the human's actions, which presents sparse partial observability.

To accurately evaluate the capabilities of VLMs in realistic HRC, this section first formalizes the task hierarchy in collaborative scenarios, the practical observation constraints of robots, and the process of step-by-step human intent inference.

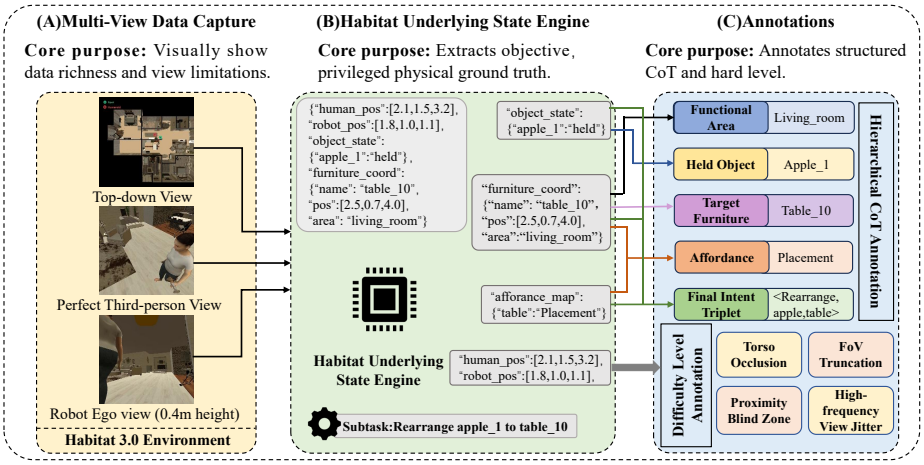


Fig. 1: Overview of the CoT-HRC benchmark dataset pipeline. **(A)** Multi-view data capture comparing ideal top-down and third-person views with the highly constrained robot egocentric view (*Ego*). **(B)** The underlying state engine in Habitat 3.0 used to extract objective, privileged physical ground truth during sub-task execution. **(C)** Our automated annotation framework, generating explicit hierarchical CoT labels and quantitative difficulty levels based on dynamic physical constraints.

3.1 Hierarchical Task and Intent Modeling

In complex HRC scenarios, tasks usually exhibit long-sequence characteristics. We define the overall collaborative goal as the global collaborative task $\mathcal{G}$. $\mathcal{G}$ typically consists of a sub-task sequence $S = s_1, s_2, ..., s_n$. A sub-task $s_i (i = 1, 2, ..., n)$ serves as an independent behavioral unit constituting the global collaborative task. We generally model sub-tasks as state-change or location-transfer processes. State changes involve variations in state values, such as cleanliness (is_clean) or fullness (is_filled). Location transfers generally encompass the *Search/Collect-Transit-Drop_off/Use* steps. For example, "storing the apple in the refrigerator" constitutes a complete sub-task. In the context of this paper, human intent $\mathcal{I}$ strictly refers to the sub-task the human is currently executing. The core objective of the robot's inference of human intent is to deduce the currently executing sub-task s_k based on the observation video. Each sub-task consists of an underlying atomic action sequence $A = a_1, a_2, ..., a_m$, where the atomic action $a_i (i = 1, 2, ..., m)$ represents low-level physical actions like navigate, pick up and so on.

3.2 Modeling Robot Egocentric Observation

Unlike traditional video understanding tasks that possess complete and clear third-person or first-person perspectives, realistic quadruped robots face both temporal and spatial observation constraints.

Local Observation in the Temporal Dimension. In the aforementioned long-horizon HRC scenarios, humans and robots usually execute their respective sub-tasks asynchronously. The robot's behavior abstracts into two alternating phases: the Execution Phase and the Observation Phase. The robot does not continuously observe the human. Instead, after finishing its own task, it intermittently observes the human to decide its next action. Suppose the human executes sub-task s_k, and the task's execution interval on the timeline is $[t_{start}, t_{end}]$. The robot begins observing at time t_{obs} with an observation duration of Δt, resulting in the robot's observation video clip $\mathcal{W}$ being $[t_{obs}, t_{obs} + \Delta t]$. Due to the uncertainty of the observation start time t_{obs} and the observation duration Δt, the video clip $\mathcal{W}$ often misses the initial state or key frames of the human's action.

Local Observation in the Spatial Dimension. The realistic robot egocentric view of quadruped robots features a significant low-angle view, and the robot's turning actions when tracking and observing the human also suffer from delays or spatial restrictions. Therefore, when the robot observes from a close distance or turns slowly, the human torso, furniture, and other elements create severe dynamic physical occlusion, causing critical interaction actions or objects to disappear in $\mathcal{W}$. More severely, the Humanoid completely moves out of the frame in certain frames.

3.3 Modeling Step-by-Step Intent Inference

Under the aforementioned Observation Window $\mathcal{W}$ with incomplete spatiotemporal information, traditional end-to-end intent inference models $f(\mathcal{W}) \rightarrow \mathcal{I}$ easily fail. To evaluate whether the model reasons based on actual observations, we introduce hierarchical CoT as an evaluation method and diagnostic probe. Under this evaluation setup, given a restricted observation $\mathcal{W}$, the tested model must output the following intermediate reasoning chain states to verify the consistency of its reasoning process: 1) Spatial Region Recognition S_{loc}: The final region where the human stops; 2) Held Object Status Recognition S_{obj}: Whether the human holds an object; 3) Target Furniture Speculation S_{fur}: The furniture the human most likely interacts with; 4) Furniture's Affordance S_{aff}: The semantic attributes of that furniture; 5) Final Intent $\mathcal{I}$: The final abductive reasoning combining the aforementioned cues.

By introducing the aforementioned hierarchical CoT, we provide a formal foundation for designing quantitative evaluation metrics later.

4 The CoT-HRC Benchmark

We introduce CoT-HRC, a large-scale embodied collaboration dataset designed to bridge the visual perception gap and the semantic reasoning gap. To ensure the physical authenticity of the data and the rigorous logic of the tasks, we build a controllable procedural generation pipeline based on the Habitat 3.0 simulation

platform [29]. Unlike previous datasets that only focus on third-person perspectives or single action labels, CoT-HRC uniquely integrates realistic quadruped robot perception characteristics and procedural CoT reasoning. By simulating the low-angle view and dynamic noise of robots in realistic HRC, we recreate the perception challenges of the realistic robot egocentric view and utilize the simulator's underlying information to generate explicit logical attribution paths, providing a reliable benchmark to evaluate models' reasoning abilities under partial observability.

4.1 Simulation Environment and Task Scenarios

Simulation Platform We choose Habitat 3.0 [29] as the underlying simulation platform. Compared to other simulators and earlier versions of Habitat [31], Habitat 3.0 [29] features Humanoids with realistic kinematic properties, enabling us to generate realistic continuous human action streams, such as walking and turning, rather than simple rigid-body displacements. For scenes, we utilize high-fidelity indoor scene assets to ensure the complexity and authenticity of the visual background.

Task Scenario Setup. To simulate realistic daily HRC, we adopt the Rearrange task from the PARTNR [27] framework. Given that complex long-horizon activities inherently decompose into sequential state-changes and object transfers, Rearrange serves as a fundamental atomic primitive for evaluating embodied collaboration. This setup features two independent agents: a Humanoid and a Spot robot. To mirror decentralized collaboration, both agents execute sub-tasks in parallel without real-time state synchronization. Consequently, Spot's behavior alternates between an Execution Phase and an Observation Phase. After completing its own sub-task, the robot switches to observation, autonomously searching for and tracking the human. Through these brief visual observations, Spot attempts to infer the human's current intent to plan its next action. The videos in our dataset are collected exclusively during this Observation Phase.

4.2 Multi-View Configuration

We synchronously sample task execution videos from three perspectives as shown in Fig. 2: the Spot egocentric view ($\mathcal{V}_{robot}$), the tracking third-person view ($\mathcal{V}_{third}$), and the global top-down view ($\mathcal{V}_{god}$). Specifically, $\mathcal{V}_{third}$ uses the Humanoid's *third_rgb_sensor* to perfectly observe human-object interactions. Meanwhile, $\mathcal{V}_{god}$ utilizes a 15m-high overhead camera to capture room layouts and agent trajectories.

As the core of CoT-HRC, $\mathcal{V}_{robot}$ authentically replicates a quadruped robot's spatial constraints. We utilize Spot's 0.4m-high jaw camera, which naturally shakes during movement. To enrich the diversity of perceptual perspectives, we also introduce alternative robot configurations at sensor heights of 1.5 meters and 0.8 meters, which further substantiates the intrinsic necessity and challenging nature of our primary low-angle viewpoint, as elaborated alongside comprehensive

Fig. 2: Visual examples of the CoT-HRC benchmark across diverse simulated indoor environments. For each task scenario, we synchronously capture video streams from three perspectives: the constrained 0.4m robot egocentric view (Top row), the unobstructed third-person tracking view (Middle row), and the global top-down view (Bottom row). Compared to the ideal perspectives, the egocentric view inherently suffers from severe field-of-view truncation and dynamic torso occlusions, highlighting the realistic perception challenges.

experimental evaluations in Appendix J. Furthermore, we replace perfect hard-coded tracking with a dynamic algorithm featuring delays, Field-of-View (FoV) truncation, and close-proximity blockages. This objectively reproduces the jitter and spatial offsets inherent to realistic egocentric perception.

To simulate temporal incompleteness, we extract Observation Windows $\mathcal{W}$ from the full sub-task video. A sub-task spans four timestamps: start t_0, pick-up t_1, target arrival t_2, and completion t_3. These define the *Search* $[t_0, t_1]$, *Transit* $[t_1, t_2]$, and *Drop_ off/Use* $[t_2, t_3]$ phases.

This benchmark primarily focuses on the occlusion-prone *Transit* phase. We design two Observation Window strategies: **1) Accumulative Window**: To evaluate reasoning as information progressively increases, we fix the start point at t_1 and truncate the video using a proportion ρ. The window is defined as: $\mathcal{W}_{acc}(\rho) = [t_1, t_1 + \rho \cdot L]$, where $L = t_2 - t_1$. In this benchmark, we extract slices for $\rho \in \{0.3, 0.6, 0.9\}$. **2) Fixed-length Phase Window**: To analyze intent exposure across different action stages while keeping observation duration constant, we maintain a window length of $l = 0.3L$ and slide it along the timeline to extract four key phases: **Pre-transit** (spanning the picking action context, $\mathcal{W}_{pre} = [\max(t_0, t_1 - 0.15 \cdot L), \max(t_0, t_1 - 0.15 \cdot L) + 0.3L]$); **Mid-transit 1** (first half of the transit phase, $\mathcal{W}_{mid1} = [t_1 + 0.30 \cdot L, t_1 + 0.60 \cdot L]$); **Mid-transit 2** (second half of the transit phase before substantive interaction, $\mathcal{W}_{mid2} =$

$[t_1 + 0.60 \cdot L, t_1 + 0.90 \cdot L]$); and **Post-transit** (adjoining the final interaction execution, $\mathcal{W}_{post} = [t_3 - 0.30 \cdot L, t_3]$).

4.3 Difficulty Quantification and Mapping

Beyond intent ground truth, evaluating visual models requires clarifying the physical observation difficulty of current video frames. This benchmark quantitatively models partial observability based on the relative positional relationship between the robot and the human in three-dimensional space.

Given a video slice with a time window of N frames, $\mathbf{p}_t^R, \mathbf{p}_t^H \in \mathbb{R}^2$ represent the 2D projection coordinates of the robot and human on the ground plane at time t, with distance $d_t = \|\mathbf{p}_t^H - \mathbf{p}_t^R\|_2$. Let $\mathbf{v}_t^R, \mathbf{v}_t^H \in \mathbb{R}^2$ be their respective forward vectors. We define instantaneous physical index functions across four dimensions to capture the truncation and loss of visual features:

Torso Occlusion $\mathbb{I}_{occ,t}$: Measures whether the human turns their back to the robot, causing the torso to occlude front-facing interaction objects or targets. Based on human anatomical operational spatial distribution, the occlusion judgment triggers when the dot product of their relative orientation and position is less than a specific view frustum threshold:

$$\mathbb{I}_{occ,t} = \mathbb{I}\left(\mathbf{v}_t^H \cdot \frac{\mathbf{p}_t^R - \mathbf{p}_t^H}{d_t} < -0.3\right) \tag{1}$$

Here, the condition less than -0.3 corresponds to a dynamic physical occlusion sector of approximately 145° behind the human.

FoV Truncation $\mathbb{I}_{vl,t}$: Measures whether the robot maintains the human within its core receptive field. Based on a 90° horizontal field of view of the camera, the boundary deviation threshold is set to $\cos(45°) \approx 0.707$:

$$\mathbb{I}_{vl,t} = \mathbb{I}\left(\mathbf{v}_t^R \cdot \frac{\mathbf{p}_t^H - \mathbf{p}_t^R}{d_t} \leq 0.707\right) \tag{2}$$

An angle deviation greater than 45 degrees between the human and robot means the human falls outside the robot's field of view.

Proximity Blind Zone $\mathbb{I}_{prox,t}$: Measures frame spillover caused by the quadruped robot's low-angle view. Based on geometric derivations of the 0.4m camera height and vertical field of view, when the distance is too close, parts of the Humanoid exceed the frame edges:

$$\mathbb{I}_{prox,t} = \mathbb{I}(d_t < 0.75\text{m}) \tag{3}$$

High-frequency View Jitter $\mathbb{I}_{shake,t}$: Measures instantaneous frame blur caused by the robot's kinematic tracking, defined as an instantaneous angular velocity $\omega_t^R > 0.8$ rad/s.

To obtain the overall degradation level of macroscopic video clips, we perform discrete integration of the above instantaneous index functions over time to calculate the temporal proportions of each condition, denoted as $\mathcal{R} = \{R_{occ}, R_{vl},$

Table 1: Comparison of CoT-HRC with mainstream datasets. Unlike previous benchmarks relying on ideal perspectives or lacking diagnostic labels, CoT-HRC uniquely features realistic robot-centric constraints and hierarchical CoT annotations.

Dataset	Environment	Viewpoint	HRC Focus	Dynamic Occlusion	Annotation Type
Charades [33]	Real	Third-person	×	Low	Action Class
BEHAVIOR-1K [20]	Sim	Ideal Ego / Global	×	Low	Task State
Ego-Exo4D [12]	Real	Human-Ego & Third	✓	Low	Action / Goal
PARTNR [27]	Sim	Global / Ego	✓	Low	High-level Plan
HARPER [1]	Real	Robot-Ego	✓	High	Action / Pose
CoT-HRC (Ours)	**Sim**	**Robot-Ego (0.4m)**	✓	**High**	**Hierarchical CoT**

$R_{prox}, R_{shake}\}$. Finally, we project the continuous degradation rate matrix $\mathcal{R}$ into three discrete difficulty levels:

$$\mathcal{D} = \begin{cases} Hard, & \text{if } R_{occ} > 0.6 \vee R_{vl} > 0.5 \vee R_{prox} > 0.6 \\ Easy, & \text{if } R_{occ} < 0.25 \wedge R_{vl} < 0.15 \wedge R_{prox} < 0.2 \wedge R_{shake} < 0.05 \\ Medium, & \text{otherwise} \end{cases} \tag{4}$$

Here, *Easy*, *Medium*, and *Hard* respectively represent the physical observation difficulty of the video.

4.4 Dataset Statistics and Feature Analysis

Built on Habitat 3.0 [29], CoT-HRC contains approximately 40 hours of continuous HRC videos covering diverse household tasks. By segmenting these videos via underlying state machines, we extract over 3,000 independent sub-task clips. Utilizing the dynamic Observation Window (Sec. 4.2), we further derive tens of thousands of truncated slices across varying temporal phases, each automatically annotated with logically progressive CoT ground truths.

While existing benchmarks advance perception capabilities, they mostly overlook real-world robot observation constraints according to Tab. 1. As illustrated in Fig. 1A, datasets like Ego-Exo4D use high, stable human-worn cameras, leaving targets clearly visible. Conversely, CoT-HRC natively simulates a 0.4m low-angle quadruped view, inducing severe dynamic torso occlusions and FoV truncations. Furthermore, unlike HARPER and PARTNR, CoT-HRC uniquely equips heavily truncated visual inputs with multi-step logical attributions (Fig. 1C). This fills a critical gap in diagnosing VLM reasoning under partial observability.

5 Evaluation Methodology

To objectively evaluate the realistic reasoning capability of VLMs under incomplete spatiotemporal information and partial observability, this benchmark devises a step-by-step evaluation methodology and proposes a set of quantitative evaluation metrics based on structured outputs.

5.1 Step-by-Step QA Protocol

Traditional end-to-end evaluations obscure whether VLM failures stem from visual perception loss or deficient physical commonsense. To address this black-box issue, we introduce a Step-by-Step Reasoning QA protocol. Aligned with the annotations in Sec. 4.3, this method decouples intent inference into a multi-step chain by sequentially prompting the model with progressive questions: **1) Target Zone Anchoring:** What is the final target zone that the Humanoid in the video is approaching?; **2) Held Object Identification:** If an object is being held, can the category of the object be determined?; **3) Target Furniture Identification:** Which piece of furniture within the target zone is the Humanoid most likely to interact with?; **4) Affordance Extraction:** What is the functional affordance of this piece of furniture?; and **5) Intent Inference:** Based on the analysis of the preceding steps, output the final Intent Triplet $\langle Action, Object, Target \rangle$.

For rigorous evaluation, responses are constrained to a restricted vocabulary, including an *unknown* option for severe occlusions. Recording predictions hierarchically provides a transparent basis for diagnosing cognitive breakpoints.

5.2 Evaluation Metrics

To quantitatively evaluate the structured JSON outputs from the step-by-step reasoning QA, we adopt a soft-matching mechanism combined with functional commonsense for basic elements and propose a set of commonsense-aware metrics.

Base Perception Metrics. We evaluate each reasoning node independently (Acc_{area}, Acc_{status}, Acc_{obj}, Acc_{aff}). For furniture (Acc_{furn_func}), targets are mapped into broad functional groups such as surface or storage to prevent penalizing functionally equivalent predictions. Finally, macro-intent (Acc_{macro_intent}) requires jointly predicting the target region and affordance, prioritizing the overarching goal over heavily occluded objects.

Step-wise Consistency Assessment. To ensure the intent prediction is grounded in a valid logical foundation, we introduce Reasoning Consistency SCA_R. It requires the model to correctly predict the action affordance supported by either the correct functional furniture or the correct spatial area:

$$SCA_R = Acc_{aff} \wedge (Acc_{furn\_func} \vee Acc_{area}) \tag{5}$$

We also report Full-chain Consistency SCA_F, a strict upper-bound metric requiring $SCA_R \wedge Acc_{obj}$.

Conditional Reasoning Accuracy. A core challenge in egocentric evaluation is distinguishing whether a failure stems from physical occlusion or deficient cognitive reasoning. To decouple perception from reasoning, we introduce CRA, defined as the probability of correctly inferring the macro-intent given successful perception of the prerequisite area and object:

$$CRA = P(Acc_{macro\_intent} = True \mid Acc_{area} = True \wedge Acc_{obj} = True) \tag{6}$$

This metric isolates and evaluates the model's pure abductive reasoning ability when visual prerequisites are met.

6 Experiments and Analysis

6.1 Experimental Setup

Evaluated Models. To establish comprehensive baselines across diverse architectures, we evaluate six VLMs: Qwen2-VL-7B-Instruct [35], InternVL2.5-8B [6], LLaVA-NeXT-Video-7B [38], LLaVA-OneVision (Qwen2-7B) [19], GPT-4o and Gemini-2.5-pro. All are tested strictly zero-shot to assess generalization against unseen egocentric occlusions without task-specific fine-tuning.

Implementation Details. We uniformly sample 8 frames per video slice to fit VLM context windows, setting the decoding temperature to 0 for reproducibility. Dual-view inputs are spatially downsampled to prevent out-of-memory errors while maintaining temporal alignment. All experiments run on NVIDIA GPUs.

6.2 Main Results: The Viewpoint Bottleneck

Table 2 presents the quantitative results across Egocentric (*Ego*), Third-person (*Third*), and Dual-View (*Dual*) which inputs the videos from the third-person view and top-down view.

Our quantitative evaluation grid reveals several critical insights regarding how current vision-language models navigate severe visual degradation under embodied view constraints.

The Impact of Physical Occlusion. Our results demonstrate a clear performance gap across different viewpoints. For highly capable models like InternVL and Qwen2-VL, we observe a consistent performance hierarchy. For instance, InternVL drops from 59.6% SCA_R under the dual-view Oracle setting to 44.8% under the egocentric view. This highlights the severe challenge posed by low-angle truncation and dynamic torso occlusion, supporting our primary hypothesis that physical occlusion severely disrupts the spatial reasoning chain.

Architecture Sensitivity. Interestingly, LLaVA-family models exhibit unexpected performance drops under the Oracle view compared to their egocentric performance. This likely stems from a severe spatial domain gap: the downsampled, overhead Oracle view diverges significantly from the human-centric, first-person datasets these models were heavily optimized on during visual-language pre-training. Thus, simply providing more multi-view visual information does not guarantee better intent reasoning if the underlying vision encoder lacks robust cross-view generalization.

Human Baseline and Proprietary Models. To establish the absolute performance upper bound for this task, a human baseline study was conducted to yield an average intent accuracy of 83.7%$\pm$1.7%. Furthermore, extended evaluations on advanced proprietary models including GPT-4o and Gemini-2.5-Pro

Table 2: Zero-shot intent anticipation performance of state-of-the-art VLMs on the CoT-HRC benchmark. We report results across three distinct viewpoints: Egocentric (*Ego*), Third-person (*Third*), and Dual-view Oracle (*Oracle*). Metrics include component accuracies, the macro-intent accuracy (*Intent*), the Conditional Reasoning Accuracy (*CRA*), and the Reasoning Consistency (SCA_R). All values are in percentages (%). Parentheses indicate *CRA* values calculated with a valid sample size of ≤ 5.

Viewpoint	Model	Area	Status	Obj	Furn	Aff	Intent	CRA	SCA_R
Ego	InternVL2.5-8B	51.6	37.7	3.1	41.5	52.9	31.6	66.7	**44.8**
	Qwen2-VL-7B-Instruct	54.9	49.5	1.4	20.2	50.4	27.1	(50.0)	34.1
	LLaVA-OneVision-7B	60.3	57.6	8.7	33.0	37.9	24.9	43.3	33.4
	LLaVA-NeXT-Video-7B	27.8	21.7	0.4	12.1	72.9	18.9	(0.0)	25.3
	GPT-4o	61.2	72.7	14.0	51.6	72.3	47.7	80.0	62.9
	Gemini-2.5-pro	59.4	73.4	19.5	49.3	61.4	42.2	73.1	54.3
Third	InternVL2.5-8B	61.0	64.8	10.5	43.1	56.5	38.8	88.6	**52.4**
	Qwen2-VL-7B-Instruct	63.5	69.3	7.6	27.6	51.3	33.0	63.0	41.5
	LLaVA-OneVision-7B	65.0	90.1	22.6	36.6	30.7	20.2	25.0	27.6
	LLaVA-NeXT-Video-7B	29.8	31.2	0.2	14.6	76.9	18.9	(0.0)	26.0
	GPT-4o	67.9	95.2	26.8	55.0	71.5	50.2	76.7	62.8
	Gemini-2.5-pro	66.8	96.8	34.9	56.4	71.3	47.5	74.6	63.6
Dual	InternVL2.5-8B	59.4	62.5	8.5	43.9	64.3	45.3	95.8	**59.6**
	Qwen2-VL-7B-Instruct	60.3	66.6	7.6	24.9	56.3	35.2	64.0	41.7
	LLaVA-OneVision-7B	60.3	92.4	18.4	31.1	30.0	17.9	31.8	23.6
	LLaVA-NeXT-Video-7B	22.2	15.0	0.0	11.9	71.8	14.4	(0.0)	21.7
	GPT-4o	70.2	95.2	27.3	52.5	72.0	52.3	73.6	64.2
	Gemini-2.5-pro	70.2	95.9	34.4	56.9	71.5	50.2	76.8	65.6

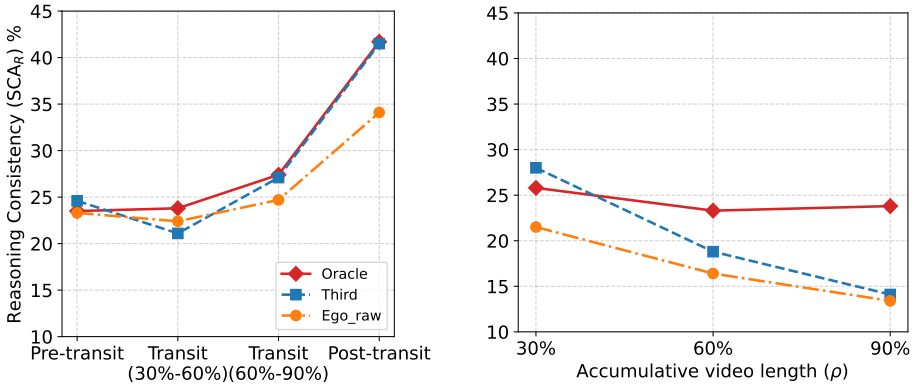
demonstrate that although their absolute reasoning accuracy improves, cross-view integration remains a persistent bottleneck, as detailed alongside complete datasets and comprehensive error analysis in Appendix F.

Sim-to-Real Fidelity. To ensure the view bottleneck observed in simulation authentically reflects real-world constraints, we conducted physical pilot studies using a Unitree quadruped robot. Evaluations across 60 real-world clips strictly corroborate our simulation findings, confirming that the severe egocentric performance degradation persists in reality. Furthermore, to validate the dataset's practical utility for model adaptation, fine-tuning Qwen2-VL-7B on 1,000 sampled instances yielded significant zero-shot capability gains, *e.g.*, SCA_R surged from 34.1% to 68.6%. Detailed real-world experimental setups and complete evaluation tables are provided in Appendix G and Appendix H.

6.3 Temporal Dynamics

To analyze VLM performance across different action stages and observation lengths, we conduct a temporal analysis using Qwen2-VL (Fig. 3).

Impact of Physical Phases. Fig. 3a illustrates SCA_R across four sequential action phases. Following initial state captures in the *pre-transit* phase, perfor-



(a) Performance across temporal phases (b) Impact of accumulative observation

Fig. 3: Temporal dynamics of intent anticipation. (a) Reasoning consistency (SCA_R) across four sequential action phases. (b) The impact of accumulative observation video length (ρ) on reasoning performance. Both evaluations are conducted using Qwen2-VL-7B-Instruct.

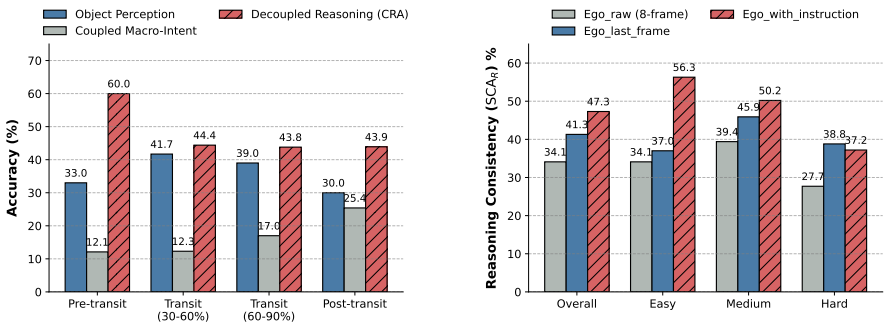
mance generally drops during the *transit (30%-60%)* phase. This decline occurs because active human movement introduces severe dynamic occlusions and motion blur, corrupting target visual features. Performance rebounds during the *post-transit* phase when the target furniture becomes visible, highlighting that intent anticipation is strictly constrained by instantaneous spatial relationships.

Attention Dilution in Long Observations. Fig. 3b evaluates performance as the observation window progressively accumulates ($\rho = 30\%$ to 90%). While longer videos intuitively provide more context, data reveals a monotonic SCA_R decay. For instance, Third-person SCA_R drops from 28.0% ($\rho = 30\%$) to 14.1% ($\rho = 90\%$). This indicates VLMs struggle to robustly aggregate extended temporal cues. Redundant frames and background shifts dilute attention, causing models to lose focus on critical fine-grained features necessary for logical reasoning.

6.4 Disentangling Perception and Reasoning

To investigate whether model failures stem from deficient physical commonsense or mere visual breakdown, we introduce an annotated egocentric setting (Ego_ann) where bounding boxes explicitly highlight active objects. Using Qwen2-VL, we compare Object Perception (Acc_{obj}), coupled Macro-Intent, and Conditional Reasoning Accuracy (CRA).

As shown in Fig. 4a, with explicit bounding boxes, low-level object perception (Acc_{obj}) reaches 30%-40%. However, the end-to-end Macro-Intent remains relatively low (12.1%-25.4%) because the target furniture remains severely occluded by the human torso, preventing complete logical deduction. Crucially, when isolating reasoning capability via CRA (intent accuracy given successful



(a) Disentangling Perception and Reasoning (b) Ablations on Semantic Priors & Difficulty

Fig. 4: Diagnostic ablation studies on the CoT-HRC benchmark using Qwen2-VL. **(a)** Disentangling visual perception (Acc_{obj}) and cognitive reasoning (CRA) under the *Ego_ann* setting across different temporal phases. **(b)** Ablation results evaluating the impacts of temporal inputs (1-frame vs. 8-frame), generalized semantic instructions, and spatial difficulty levels under the *Ego_raw* setting.

area and object perception), performance surges to 43.8%–60.0%. This gap provides a key diagnostic insight: VLMs inherently possess the required physical commonsense. Their poor HRC performance is predominantly a visual perception failure; low-angle occlusions sever prerequisite visual links, fundamentally preventing models from triggering latent reasoning chains.

6.5 Ablation Studies and Further Analysis

We further evaluate temporal and semantic ablations, visualized in Fig. 4b.

Temporal Aggregation vs. Single Frame. Comparing the 8-frame input (*Ego_raw*) to a single final-frame baseline (*Ego_raw_last_frame*) reveals an unexpected SCA_R drop (34.1% vs. 41.3%). This exposes inadequate temporal aggregation: rather than modeling causality, models are distracted by early redundant frames and motion blur, making the single final frame yield a higher signal-to-noise ratio.

Compensating Vision with Semantic Priors. Introducing a generalized language instruction (*e.g.*, "help me tidy up the room") significantly improves SCA_R to 47.3% (*Ego_raw_with_instruction*). This abstract instruction acts as a top-down constraint, effectively narrowing the affordance search space when bottom-up visual features are severely truncated.

Difficulty Levels & Priors. Under *Ego_raw*, *Medium* difficulty (39.4%) surprisingly outperforms *Easy* (34.1%), reflecting a perceptual trade-off: distant *Easy* frames avoid truncation but suffer severe resolution loss, while closer *Medium* frames offer clearer details despite minor occlusions. Introducing semantic priors resolves this resolution bottleneck via language reasoning, restor-

ing strict monotonic degradation (Easy > Medium > Hard) and confirming geometric constraints as the ultimate bottleneck.

7 Discussion

While CoT-HRC serves as a rigorous diagnostic framework, transitioning to open-world Human-Robot Collaboration (HRC) introduces limitations that chart future research directions.

Sim-to-Real Generalization. Although real-world pilots validate our simulated constraints, the large-scale benchmark relies on synthetic Habitat 3.0 environments. Scaling explicitly annotated real-world datasets remains crucial to fully close the sim-to-real gap and ensure robust physical deployment.

Task and Intent Diversity. While the Rearrange task acts as a rigorous atomic primitive, future intent taxonomies should expand beyond spatial object-transfers to encompass nuanced social gestures, articulated manipulations, and complex tool usage.

Algorithmic Innovations. Having exposed critical perception-reasoning bottlenecks under partial observability, developing novel models or learning methods to resolve cross-view integration and temporal aggregation remains an open frontier. Addressing these algorithmic limitations will catalyze the community's shift toward robust reasoning paradigms tailored for real-world physical constraints.

8 Conclusion

We introduced CoT-HRC, a benchmark evaluating human intent inference under constrained robot egocentric views. By coupling simulated spatiotemporal incompleteness with deterministic, hierarchical CoT annotations, our proposed diagnostic metrics (SCA_R and CRA) quantitatively disentangle visual perception failures from inherent cognitive deficiencies.

Extensive experiments reveal that while state-of-the-art VLMs possess robust physical commonsense, their performance is severely bottlenecked by dynamic occlusions and inadequate temporal aggregation. Ultimately, CoT-HRC establishes a definitive diagnosis of current cognitive boundaries under embodied visual constraints, providing a vital stepping stone for robust, interaction-aware embodied AI.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62376280 and T2521006).

References

1. Avogaro, A., Toaiari, A., Cunico, F., Xu, X., Dafas, H., Vinciarelli, A., Li, E., Cristani, M.: Exploring 3d human pose estimation and forecasting from the robot's perspective: The harper dataset. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2024). https://doi.org/10.1109/IROS58592.2024.10802238
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
3. Bian, T.: Robust understanding of human-robot social interactions through multimodal distillation. In: ACM MM (2025)
4. Caba Heilbron, F., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR (2015). https://doi.org/10.1109/CVPR.2015.7298698
5. Chang, M., Chhablani, G., Clegg, A., Cote, M.D., Desai, R., Hlavac, M., Karashchuk, V., Krantz, J., Mottaghi, R., Parashar, P., Patki, S., Prasad, I., Puig, X., Rai, A., Ramrakhya, R., Tran, D., Truong, J., Turner, J.M., Undersander, E., Yang, T.Y.: Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. In: ICLR (2025)
6. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821v2 (2024)
7. Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., Liu, Y.: Egothink: Evaluating first-person perspective thinking capability of vision-language models. In: CVPR (2024)
8. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. In: CVPR (2024)
9. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: An embodied multimodal language model. In: ICML (2023)
10. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: CVPR (2024). https://doi.org/10.1109/CVPR52733.2024.01356
11. Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., Kim, K., Wang, E., Lingelbach, M., Curtis, A., Feigelis, K., Bear, D.M., Gutfreund, D., Cox, D., Torralba, A., DiCarlo, J.J., Tenenbaum, J.B., McDermott, J.H., Yamins, D.L.: Threedworld: A platform for interactive multi-modal physical simulation. In: NeurIPS (2021)
12. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., González, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A.,

Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbeláez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives pp. 19383–19400 (2024). https://doi.org/10.1109/CVPR52733.2024.01834

13. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: CVPR (2024). https://doi.org/10.1109/CVPR52733.2024.01363
14. Guo, H., Agarwal, N., Lo, S.Y., Lee, K., Ji, Q.: Uncertainty-aware action decoupling transformer for action anticipation. In: CVPR (2024). https://doi.org/10.1109/CVPR52733.2024.01764
15. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: CVPR (2024). https://doi.org/10.1109/CVPR52733.2024.01274
16. Jia, B., Chen, Y., Huang, S., Zhu, Y., Zhu, S.C.: Lemma: A multi-view dataset for learning multi-agent multi-task activities. In: ECCV (2020)
17. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474 (2017)
18. Lei, Q., Wang, B., Tan, R.T.: Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection. In: NeurIPS (2024)
19. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
20. Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Lingelbach, M., Sun, J., Anvari, M., Hwang, M., Sharma, M., Aydin, A., Bansal, D., Hunter, S., Kim, K.Y., Lou, A., Matthews, C.R., Villa-Renteria, I., Tang, J.H., Tang, C., Xia, F., Savarese, S., Gweon, H., Liu, C.K., Wu, J., Fei-Fei, L.: Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In: Conference on Robot Learning (CoRL) (2023)
21. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Lou, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: CVPR (2024). https://doi.org/10.1109/CVPR52733.2024.02095
22. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2023)
23. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2024)
24. Liu, H., Shah, R., Liu, S., Pittenger, J., Seo, M., Cui, Y., Bisk, Y., Martín-Martín, R., Zhu, Y.: Casper: Inferring diverse intents for assistive teleoperation with vision language models. In: Conference on Robot Learning (CoRL) (2025)
25. Ma, Y., Song, Z., Zhuang, Y., Hao, J., King, I.: A survey on vision-language-action models for embodied ai. IEEE Transactions on Neural Networks and Learning Systems (2026). https://doi.org/10.1109/TNNLS.2025.3650584

26. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: ICCV (2019)
27. Mittal, H., Agarwal, N., Lo, S.Y., Lee, K.: Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In: CVPR (2024)
28. Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., Luo, P.: Embodiedgpt: Vision-language pre-training via embodied chain of thought. In: NeurIPS (2023)
29. Puig, X., Undersander, E., Szot, A., Cote, M.D., Yang, T.Y., Partsey, R., Desai, R., Clegg, A.W., Hlavac, M., Min, S.Y., Vondruš, V., Gervet, T., Berges, V.P., Turner, J.M., Maksymets, O., Kira, Z., Kalakrishnan, M., Malik, J., Chaplot, D.S., Jain, U., Batra, D., Rai, A., Mottaghi, R.: Habitat 3.0: A co-habitat for humans, avatars and robots. In: ICLR (2024)
30. Roy, D., Rajendiran, R., Fernando, B.: Interaction region visual transformer for egocentric action anticipation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024). https://doi.org/10.1109/WACV57701.2024.00660
31. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A platform for embodied ai research. In: ICCV (2019). https://doi.org/10.1109/ICCV.2019.00943
32. Shen, B., Xia, F., Li, C., Martin-Martin, R., Fan, L., Wang, G., Perez-D'arpino, C., Buch, S., Srivastava, S., Tchapmi, L., Tchapmi, M., Vainio, K., Wong, J., Fei-Fei, L., Savarese, S.: igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2021). https://doi.org/10.1109/IROS51168.2021.9636667
33. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016)
34. Wang, G., Guo, Y., Xu, Z., Kankanhalli, M.: Bilateral adaptation for human-object interaction detection with occlusion-robustness. In: CVPR (2024). https://doi.org/10.1109/CVPR52733.2024.02642
35. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191v2 (2024)
36. Zhang, X., Tian, S., Liang, X., Zheng, M., Behdad, S.: Early prediction of human intention for human–robot collaboration using transformer network. Journal of Computing and Information Science in Engineering **24**(5), 051004 (2024). https://doi.org/10.1115/1.4064258
37. Zhang, Y., Li, B., Liu, H., Lee, Y.J., Gui, L., Fu, D., et al.: Llava-next: A strong zero-shot video understanding model (April 2024), https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, accessed: 2026-06-29
38. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
39. Zhao, Q., Zhang, C., Wang, S., Fu, C., Agarwal, N., Lee, K., Sun, C.: Antgpt: Can large language models help long-term action anticipation from videos? In: ICLR (2024)