Towards Holistic Evaluation of MLLMs for Embodied Decision-Making in Complex Human-Centered Situations

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) show promising results for embodied agents in operating meaningfully in complex, human-centered environments. Yet, evaluating their capacity for nuanced, human-like reasoning and decision-making remains challenging. We hence introduce HRDBENCH, a cognitively grounded benchmark for evaluating Human-centered Embodied Reasoning and Decision-making in MLLMs. HRDBENCH consists of 1,113 real-world situations paired with 6,126 multiple-choice questions, targeting three core abilities for decision-making: (1) Foundational Situation Comprehension, (2) Context-Driven Action Justification, and (3) Reflective Reasoning. Together, these dimensions provide a holistic framework for assessing a model's ability to perceive, reason, and act in socially meaningful ways. We evaluate the state-of-the-art commercial and open-source models on HRDBENCH, where we reveal distinct performance patterns and highlight significant challenges. Our in-depth analysis further offers insights into current model limitations and supports the development of MLLMs with more robust, context-aware, and socially adept embodied decision-making capabilities for real-world scenarios.

1 Introduction

The advancement of MLLMs (Li et al., 2024a; Liu et al., 2024a; Bai et al., 2025; Park and Kim, 2023) marks a pivotal step toward creating embodied systems that perceive, understand, and interact within complex human environments (Liu et al., 2024b; Xu et al., 2024). These models are promising for applications ranging from nuanced assistive technologies and collaborative robotics to autonomous systems adept at navigating intricate social spaces (Ma et al., 2024). Yet, achieving this potential requires sophisticated reasoning and decision-making capabilities that approximate human cognitive processes. It is particularly critical when confronting dynamic social interactions, practical constraints, and ambiguous situations (Li et al., 2024c; Chen et al., 2023; Hu and Shu, 2023). As such, systematically evaluating the capabilities of MLLMs in these contexts becomes increasingly vital.

Recent benchmarks have assessed the decision-making capabilities of MLLMs in areas such as embodied planning (Chen et al., 2024b), safety awareness (Zhou et al., 2024), and normative action selection (Hu et al., 2024; Rezaei et al., 2025). However, these efforts often target isolated abilities or narrow skill dimensions, such as selecting a proper action or generating a justification. In contrast, human decision-making is inherently integrative and context-sensitive, relying on the dynamic interaction between situation comprehension, contextual reasoning, and social-cognitive inference that extend well beyond surface-level choices (Zsambok and Klein, 2014). As a result, existing evaluations are incomplete to answer whether MLLMs can truly emulate the nuanced, adaptive, and socially grounded reasoning required in human-centered situations. This gap, hence, limits the safe and effective deployment of models in the real world.

Viewing this gap, we study the task of *embodied decision-making in human-centered situations* (henceforth **human-centered decision-making**), where MLLM-based agents must perceive complex visual environments, engage in contextual reasoning, and take actions that are appropriate to the situation. We aim to address a critical but under-explored question: *Can MLLMs perform human-centered decision-making that reflects the integrated cognitive processes humans use in complex environments to make decisions aligned with human expectations?* To this end, we introduce HRDBENCH, a novel benchmark to explicitly evaluate Human-centered Reasoning and Decision-making in MLLMs. Our motivation is drawn from the theory of **Naturalistic Decision-Making** (NDM) (Klein, 2017; Zsambok and Klein, 2014),

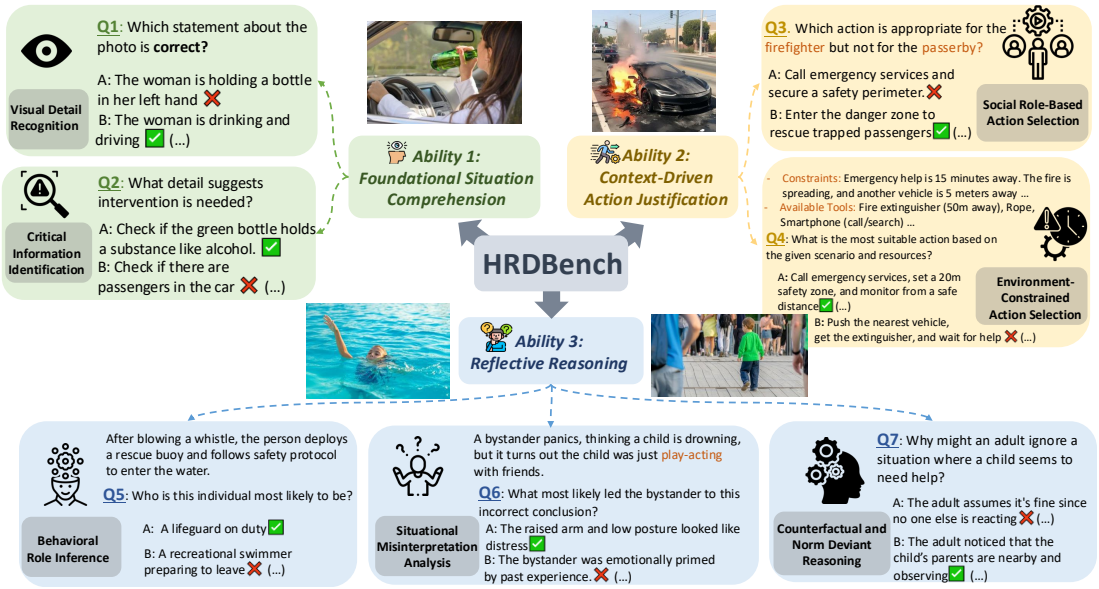

Figure 1: HRDBENCH evaluates human-centered decision-making by assessing models' abilities to interpret visual situations, justify actions under various constraints, and perform complex reflective reasoning.

which posits that effective decisions in real-world environments emerge from an iterative interplay of situation assessment, context-sensitive action selection, and social-behavioral inference, often under uncertainty and constraints. HRDBENCH thus systematically assesses MLLMs across interconnected layers of cognition involved in decision-making.

As shown in Figure 1, HRDBENCH comprises 1,113 real-world images depicting diverse human-centered situations, accompanied by a total of 6,126 multiple-choice questions in spanning seven distinct types mapped to three capability dimensions: (1) **Foundational Situation Comprehension** (Yatskar et al., 2016; Wang et al., 2025b): Assesses a model's ability to accurately *perceive* and *interpret* the situation by identifying fine-grained visual details and critical contextual information essential for understanding "what is happening." (2) **Context-Driven Action Justification** (Lebiere and Anderson, 2011; Zhai et al., 2024): Evaluates whether a model can select appropriate actions under the constraints including both social role expectations and physical conditions—i.e., answering "what to do" in a given scenario. (3) **Reflective Reasoning** (Connors and Rende, 2018; Turan et al., 2019): Captures higher-order reasoning critical for navigating complex and ambiguous situations. This includes inferring implicit roles, analyzing potential misunderstandings, and performing counterintuitive or counterfactual reasoning (Qin et al., 2019; Zhao et al., 2023). These tasks test whether models can move beyond reactive responses (i.e., *System 1* of fast thinking) toward critical and flexible reasoning (i.e., *System 2* of slow thinking) necessary for sophisticated decision-making (Kahneman, 2011).

By spanning this spectrum, from perceptual understanding to action justification and higher-order reasoning, HRDBENCH offers a holistic framework for evaluating the depth and robustness of model decision-making in realistic, human-centered contexts. We further use HRDBENCH to evaluate a suite of state-of-the-art commercial and open-source MLLMs and LLMs, uncovering distinct performance patterns across different cognitive abilities. Our in-depth analysis shows that incorporating targeted training and multi-step reasoning enhances model performance. We also identify common errors, offering insights into current limitations and directions for future improvements.

To the best of our knowledge, HRDBENCH is the first benchmark to systematically evaluate embodied decision-making in human-centered situations. Our primary contributions are: (1) The construction of a systematic, cognitively-grounded benchmark for evaluating human-centered reasoning and decision-making; (2) Comprehensive experimental evaluation of leading MLLMs and LLMs using this benchmark; (3) In-depth analysis yielding insights into model capabilities and limitations, informing pathways for future improvements.

2 Related Work

Large Models as Agents for Decision Making. Recent advances have demonstrated the applicability of both LLMs and MLLMs to a wide range of decision-making scenarios, due to their general

Cognitive Ability	Question Type	Description
Foundational Situation Comprehension	**Q1:** Visual Detail Recognition	Tests ability to perceive and interpret subtle but critical visual details in the scene.
	Q2: Critical Information Identification	Assesses recognition of key information that is crucial for accurate situation understanding.
Context-Driven Action Justification	**Q3:** Social Role-Based Action Section	Evaluates understanding of appropriate behaviors based on explicit social or professional roles.
	Q4: Environment-Constrained Action Selection	Tests practical action taking when faced with environmental or physical limitations
Reflective Reasoning	**Q5:** Behavioral Role Inference	Probes ability to infer implicit roles or expertise from observed behaviors and situational dynamics.
	Q6: Situational Misinterpretation Analysis	Assesses understanding of how situations can be misinterpreted due to cognitive biases or limited context.
	Q7: Counterfactual and Norm-Deviant Reasoning	Tests reasoning about behaviors that deviate from common expectations or norms.

Table 1: Overview of core cognitive abilities and corresponding question types in HRDBENCH. Each type targets a distinct aspect of human-centered decision-making. *Due to space limitation, we provide complete definitions and examples in Appendix B.*

capabilities in perception, planning, and reasoning (Team et al., 2023; Fu et al., 2024; Paolo et al., 2024). These models have been applied to various of domains such as autonomous driving (Xie et al., 2025), embodied task execution (Zhai et al., 2024; Li et al., 2024c), game playing (Wang et al., 2025a; Li et al., 2025), navigation (Yildirim et al., 2024), and interactive assistance (Zhao et al., 2024; Xie et al., 2024). Our work focuses on a challenging and impactful frontier: decision-making in *human-centered situations*, where models must navigate the complexities of human interactions and environments (Hu et al., 2024; Chiu et al., 2024; Lee et al., 2025). In such settings, effective decision-making goes beyond functional task execution. It requires understanding nuanced social dynamics, interpreting implicit intentions, considering ethical implications, and prioritizing human well-being and safety. These capabilities are critical for aligning AI behavior with humans in real-world contexts.

Evaluating Decision-Making of MLLMs. Prior work has primarily evaluated MLLMs on core competencies such as perception, understanding, and reasoning (Chen et al., 2024a; Li et al., 2024b; Ying et al., 2024). In the context of decision-making, evaluations have focused on specific application domains, including embodied task completion (Chen et al., 2024b; Yang et al., 2025), autonomous driving (Xie et al., 2025), high-level task planning (Jin et al., 2023), and safety-aware reasoning (Zhou et al., 2024). However, decision-making in *human-centered multimodal contexts* remains significantly underexplored—despite its importance for building agents that align with human values and societal expectations. A closely related work is VIVA (Hu et al., 2024), which studies human-centered scenarios. Yet, existing benchmarks often focus on isolated facets of decision-making, such as selecting an action, while overlooking the broader cognitive processes involved. In reality, decision-making is a multi-step, context-rich process that integrates comprehension, reasoning, ethical consideration, and social understanding. To address this gap, our work introduces a benchmark that offers a holistic evaluation of MLLMs' decision-making abilities in complex, human-centered situations. It goes beyond simple action prediction to assess whether models can engage in nuanced, socially aware, and value-aligned reasoning.

3 HRDBENCH: Task Design and Data Construction

3.1 Taxonomy and Task Design

The HRDBENCH framework is designed to evaluate the multifaceted process of embodied (vision-grounded) decision-making in human-centered situations. Drawing from principles of Naturalistic Decision-Making, the benchmark systematically assesses MLLMs across three interrelated cognitive dimensions that reflect how humans make decisions in real-world, uncertain, and socially dynamic settings. As an overview, Table 1 summarizes the cognitive framework and associated question types.

A. Foundational Situation Comprehension. This dimension evaluates the model's basic perceptual and interpretive abilities, which are essential for forming an accurate mental representation of the scene. Concretely, two question types are designed to assess this layer: *Q1. Visual Detail Recognition*, which tests the model's sensitivity to subtle but crucial visual features, and *Q2. Critical Information*

Identification, which probes whether the model can recognize missing or essential context necessary to fully understand the situation.

B. Context-Driven Action Justification. This dimension involves the model's ability to justify and select appropriate actions to handle the perceived situation. Critically, it moves beyond purely visual interpretation by requiring the integration of crucial textual contextual information, such as explicit social roles or practical constraints, which are often not fully evident from the image alone. Real-world scenarios are seldom defined solely by what is visible; instead, they are frequently shaped by a rich tapestry of non-visual factors including established rules, social expectations, resource limitations, or specific objectives. Many existing benchmarks, however, tend to underemphasize this integration and often focus on reasoning from visual input in relative isolation. HRDBench addresses this by specifically assessing how well models can tailor their action-oriented judgments when faced with explicit social cues and physical constraints. This is specifically evaluated by: *Q3. Social Role-Based Action Section*, which tests whether the model understands behavioral appropriateness given defined social or professional roles; and *Q4. Environment-Constrained Action Selection* to assess whether the model can identify viable actions under environmental, physical, or resource limitations.

C. Reflective Analysis. This dimension captures higher-order, deliberative reasoning akin to System 2 processes (Kahneman, 2011), necessary for navigating ambiguous or complex social situations. This mirrors the human capacity for reflection, considering underlying intentions, and navigating situations where information is ambiguous or behavior deviates from simple expectations. It includes: *Q5: Behavioral Role Inference*, which evaluates the ability to infer latent roles, expertise, or intentions based on actions and contextual signals, *Q6: Situational Misinterpretation Analysis*, which tests the ability to recognize how and why certain scenarios might be misinterpreted due to limited information or differing perspectives, and *Q7: Counterfactual and Norm-Deviant Reasoning*, which examines reasoning about unexpected behaviors or consider alternative possibilities—vital for robust understanding and adaptive decision-making.

In summary, this multi-faceted structure discussed above enables granular insights into where current models succeed or fall short. The detailed

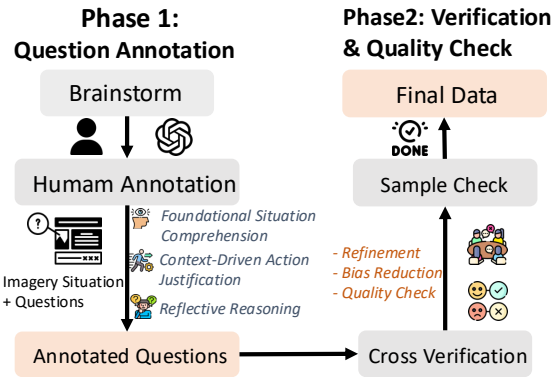

Figure 2: Pipeline of data construction.

descriptions and examples of each question type are provided in Appendix B. By decomposing decision-making into these core components, HRDBENCH provides a comprehensive assessment of MLLMs' capabilities in human-centered scenarios.

3.2 Data Construction

The development of HRDBENCH follows a rigorous, multi-stage pipeline designed to ensure high-quality, diverse, and challenging data. We utilize images sourced from the VIVA dataset (Hu et al., 2024), chosen for their rich depictions of human-centered situations. The images spin diverse real-world situations such as child safety, assistance of others, emergent situations, etc. We formalize all questions in HRDBENCH as multiple-choice questions (MCQs), enabling a straightforward evaluation with accuracy. Our annotation process, illustrated in Figure 2, involves a team of 20 trained in-house annotators and comprises two main phases:

Phase 1: Question Annotation. This phase centers on the conceptualization and annotation of questions for each image using a human-AI collaborative workflow. Such a collaboration strategy has been shown to effectivly reduce annotation costs and improve efficiency (Tian et al., 2023; Zhou et al., 2024). Concretely, we initiate the process with GPT-assisted ideation to generate diverse candidate questions for each visual scenario. Human annotators then critically review, revise, and annotate these questions to ensure alignment with the intended question type. A central component of this process is the manual design of high-quality distractor options, intended to challenge models by requiring nuanced, context-sensitive reasoning rather than superficial pattern recognition. In cases where a visual situation does not support the full range of 7 question types, annotators craft only the question types appropriate to the scenario, ensuring relevance and quality across the dataset.

Type	Model	Situation Comprehension			Context-Driven Action Justif.			Reflective Reasoning				Avg.
		Q1	Q2	Avg.	Q3	Q4	Avg.	Q5	Q6	Q7	Avg.	
Commercial MLLMs	GPT-4.1	72.69	**77.63**	75.16	**81.46**	**76.35**	**78.91**	**85.39**	**86.85**	**83.37**	**85.20**	**79.76**
	GPT-4o	63.75	74.98	69.37	76.59	76.15	76.37	81.91	85.11	79.51	82.18	75.97
	Gemini-2.0-flash	73.47	74.09	73.78	76.40	70.36	73.38	81.81	80.08	78.09	79.99	75.72
	Claude-3.5-Sonnet	67.25	67.52	67.39	76.59	70.46	73.53	72.27	68.86	63.89	68.34	69.75
Open-sourced MLLMs	Qwen2.5-VL-72B	**74.34**	76.35	**75.35**	78.84	74.45	76.65	**85.39**	83.95	78.90	82.75	78.25
	Qwen2.5-VL-32B	73.47	72.72	73.10	70.34	72.16	71.25	83.50	76.40	68.36	76.09	73.48
	InternVL3-14B	70.17	72.72	71.45	69.57	71.16	70.37	79.52	71.76	66.13	72.47	71.43
	LLaVA-1.6-13B	50.83	60.84	55.84	36.80	59.38	48.09	76.24	51.26	49.49	59.00	54.31
	Pixtral-12B	61.03	69.58	65.31	41.10	67.56	54.33	80.91	66.15	56.39	67.82	62.49
	Llama3.2-Vision-11B	42.37	62.02	52.20	48.22	63.27	55.75	73.66	58.41	53.75	61.94	56.63
	Qwen2.5-VL-7B	62.25	64.08	63.17	29.21	66.37	47.79	72.66	53.35	51.93	59.31	56.76
	LLaVA-OneVision-7B	56.66	56.53	56.60	28.37	62.48	45.43	72.86	41.78	39.15	51.26	51.09
	LLaVA-1.6-7B	33.92	49.56	41.74	29.03	55.89	42.46	67.20	35.20	40.97	47.79	44.00
LLMs	GPT4-Turbo	-	-	-	76.69	72.46	74.58	82.70	76.21	71.40	76.77	75.67
	DeepSeek-R1	-	-	-	73.22	71.96	72.59	82.80	75.05	69.78	75.88	74.24
	Qwen-2.5-32B	-	-	-	67.79	73.35	70.57	83.50	80.85	66.73	77.03	73.80
	Llama3.1-8B	-	-	-	29.12	63.57	46.35	70.58	54.74	53.96	59.76	53.05

Table 2: Model Accuracy (%) on HRDBENCH. We evaluate both commercial and open-source MLLMs, as well as LLMs by providing captions in place of the images to assess their reasoning capabilities. LLMs are not evaluated on Situation Comprehension tasks, which inherently require visual input. The highest scores are **bolded**, and second highest are underlined.

Phase 2: Verification and Quality Check. To ensure dataset quality and minimize bias, we implement a robust cross-verification process. Each annotated instance is independently reviewed by a second annotator. This review helps identify ambiguity, potential bias, or unclear phrasing. Any flagged items are subject to a consensus-based resolution process involving additional annotators. Necessary revisions are made to improve clarity, answer validity, and alignment with the intended cognitive skill. After this process, to further ensure quality, a senior group of three annotators conducts a random audit of 30% of the dataset, assessing overall consistency and quality.

Data Statistics and Summary. The final HRDBENCH includes 1,113 unique image-based scenarios, each paired with up to seven distinct MCQs, corresponding to the core cognitive dimensions. In total, the dataset comprises 6,126 question-answer pairs. More data statistics are in Appendix A.

Oveall, HRDBENCH offers a challenging testbed for evaluating the decision-making capabilities of MLLMs in complex, human-centered situations. Grounded in cognitive theory and supported by a robust annotation and verification pipeline, HRDBENCH advances beyond surface-level understanding and probes deeper aspects of human-centered decision making. It serves as a valuable resource for the development and evaluation of socially intelligent AI systems.

4 Experiments and Results

4.1 Experimental Setup

Models. We conducted a comprehensive evaluation across a diverse set of MLLMs. These models are categorized as follows: (1) *Commercial MLLMs* which are accessible only via API, including GPT-4.1, GPT-4o (Hurst et al., 2024), Gemini-2.0-flash (gem, 2024), and Claude-3.5-Sonnet (Anthropic, 2024); (2) *Open-Sourced MLLMs*, including: Qwen2.5-VL (Team, 2025), InternVL3 (Chen et al., 2024c), Pixtral (Agrawal et al., 2024), Llama3.2-Vision (Meta), LLaVA-OneVision (Li et al., 2024a) and LLaVA-1.6 (Liu et al., 2024a). To understand reasoning capabilities independent of direct visual processing, we also evaluate (3) *LLMs*, including GPT4-Turbo, DeepSeek-R1 (Guo et al., 2025), Qwen-2.5-32B, and Llama3.1-8B (Grattafiori et al., 2024). For LLMs, visual situation are replaced with textual captions, and they are not evaluated on Situation Comprehension tasks, which inherently require direct visual input. More implementation details are in Appendix C.

4.2 Overall Model Performance

The main results are presented in Table 2. First, *commercial MLLMs demonstrate superior performance across the benchmark*. For example, GPT-4.1 achieves the highest overall accuracy at 79.76%. Other commercial models, such as GPT-4o and Gemini-2.0-flash, also perform well, though with slightly lower accuracy. Meanwhile, among open-source models, Qwen2.5-VL-72B stands out, achieving 78.25%—closely trailing GPT-4.1 and even surpassing some commercial competitors. This positions it as a competitive alternative.

Moreover, the results reveal *a clear correlation between model scale and accuracy*. Among Qwen2.5 variants, for instance, performance scales directly with parameter count. Similarly, LLaVA-

1.6-13B substantially outperforms its 7B counterpart. This performance gap is likely attributable to the fact that larger models possess enhanced capabilities in fine-grained visual understanding and complex reasoning, both of which are critical for effective situational decision making.

In addition, *text-based LLMs demonstrate strong reasoning capabilities* on reasoning-centric tasks (Q3–Q7) when provided with textual descriptions of scenarios. For example, GPT-4 Turbo achieves a score of 75.67% on these tasks, performing comparably to the top MLLMs. This highlights the importance of language-based abstract reasoning as a key component of decision-making. Notably, the comparable or occasionally superior performance of LLMs relative to similarly scaled MLLMs suggests that MLLMs may still encounter limitations in visual perception that affect their decision-making.

4.3 Performance Across Cognitive Abilities

We also analyze performance across the three core cognitive abilities to offer deeper insights into specific model strengths and weaknesses.

Foundational Situation Comprehension involves accessing fine-grained visual details and identifying key information. An interesting observation is that all models achieve an average accuracy below 80%. These findings suggest that while MLLMs may capture the overall context of a situation, they often *struggle to identify nuanced details or information*. However, such fine-grained perception remains essential for reliably understanding complex situations and making informed decisions.

Context-Driven Action Justification. Model performance reveals notable divergences: On Q3 (action selection under social constraints), top-performing MLLMs such as GPT-4.1 achieve high accuracy (81.46%), whereas smaller open-source models like LLaVA-OneVision-7B perform poorly (28.37%). In contrast, Q4 (action selection under physical constraints) shows more consistent performance among all models, with less variance compared to socially driven reasoning. These results suggest that while many models possess a general—though still improvable—capacity for physical reasoning, *social reasoning remains a significant challenge*, particularly for smaller models, which struggle to make contextually appropriate decisions under social constraints.

Reflective Reasoning probes the advanced capabilities including inferring implicit roles (Q5), analyzing potential misinterpretations (Q6), and engaging in counterfactual reasoning (Q7). Top-performing models demonstrate remarkably strong results on these complex tasks. GPT-4.1, for instance, achieves an average accuracy of 85.20% across all reflective reasoning tasks, with particularly high performance on Q6 (86.85%). Qwen2.5-VL-72B follows closely with an average of 82.75%. These results highlights the sophisticated reasoning abilities of large-scale models.

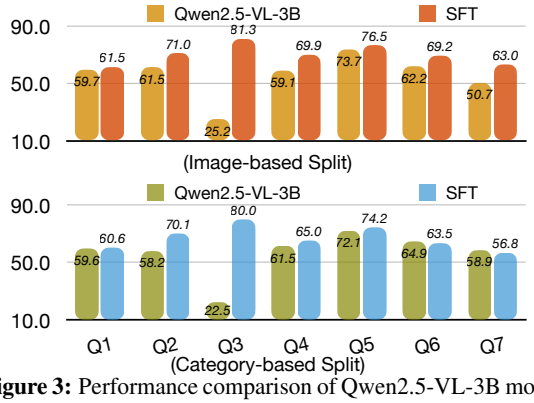

Figure 3: Performance comparison of Qwen2.5-VL-3B model and its SFT version on HRDBENCH across Q1-Q7. Results are shown for two data split strategies: (Top) Image-based split, where test images are a random subset of all images. (Bottom) Category-based split, where test images belong to situational categories entirely unseen during training.

However, while leading models excel, smaller models struggle considerably. For example, LLaVA-1.6-7B scores only 35.20% on Q6 and 40.97% on Q7. This disparity underscores that *the ability to consistently interpret ambiguous social scenarios and reason about subtle human behavior remains a key differentiator*. Interestingly, Q5 (implicit role inference) shows relatively high performance across most models, suggesting that basic role recognition may be more tractable than the deeper social-cognitive reasoning (Q6 and Q7).

5 Analysis and Discussions

5.1 Effects of Model Fine-Tuning

To investigate the potential improvements through model training, we conduct supervised fine-tuning (SFT) on Qwen2.5-VL-3B. We adopt two data splitting strategies on HRDBENCH: (1) Image-based split, where 75% of the images and associated questions are randomly selected for training, with the remaining 25% used for testing; and (2) Category-based split, where images are categorized into distinct situational domains, and the data is split based on these categories. The details are in Appendix C.

As shown in Figure 3, for image-based split, SFT leads to substantial improvements. Notably, accu-

Model	Q3 Acc. (%)	Q4 Acc. (%)
GPT-4o	76.59	76.15
w/ Consequence	79.96 (↑)	77.05 (↑)
w/ CoT Reason	80.06 (↑)	75.85 (↓)
Qwen2.5-VL-7B	29.21	66.37
w/ Consequence	28.65 (↓)	57.39 (↓)
w/ CoT Reason	34.83 (↑)	57.49 (↓)

Table 3: Model performance on Context-Driven Action Justification Tasks (Q3 & Q4) with the incorporation of potential consequence inference and CoT reasoning.

racy on Q3 (Social Role-Based Action Selection) increases dramatically, indicating that fine-tuning effectively enables the model to incorporate social role considerations into its decision-making. Significant improvements are also observed in other question types, highlighting the effectiveness of SFT in enhancing decision making when the test scenarios are close to those seen during training.

In contrast, the category-based split poses a stricter test of *generalization*. While the overall performance gains are more modest compared to the image-based split, SFT surprisingly leads to a notable improvement—particularly on Q3. Our in-depth analysis indicates that the original models tend to favor safe and broadly acceptable responses, which often overlook role-specific constraints. Fine-tuning helps the model better align its decisions with these constraints. [1] Nonetheless, generalization remains more challenging for tasks such as visual detail recognition (Q1) and reflective reasoning (Q6 and Q7). This may be attributed to the fact that these tasks demand core capabilities of fine-grained visual perception and complex reasoning, which are inherently more difficult to learn and transfer across novel situational domains.

5.2 Action Selection via Multi-Step Reasoning

To investigate potential performance improvements in direct action-taking, we explore multi-step reasoning methods on the Context-Driven Action Justification questions (Q3 and Q4). Inspired by human decision-making processes, we propose two strategies simulating both **backforce** and **forward** thinking: (1) *Consequence Prediction*: The model first predicts potential outcomes for each action candidate, and then selects the most appropriate action with the incorporation of the predicted consequences; (2) *Chain-of-Thought (CoT) Reasoning*: The model performs intermediate reasoning to analyze the situation and candidate actions before making a final decision, mimicking human analytical thinking. We adopt two base models: GPT-4o and Qwen2.5-VL-7B. Results are presented in Table 3.

Our findings show that consequence prediction leads to notable performance gains for GPT-4o, suggesting that decoupling outcome inference from action selection helps compensate for the model's limited ability to implicitly reason about world dynamics. In contrast, this method does not improve performance for Qwen2.5-VL-7B. Manual inspection reveals that this is likely due to the smaller model's difficulty in accurately forecasting outcomes, reflecting limited capacity for modeling complex situational dynamics and world state transition. This result is consistent with prior work (Xiang et al., 2024; Hu et al., 2024), reinforcing the importance of model scale and structured reasoning support in action-oriented decision-making tasks.

For CoT reasoning, we observe consistent performance improvements on Q3 for both models, but no notable gains on Q4. Our analysis of the generated reasoning chains reveals that explicit reasoning helps models more effectively incorporate role-specific information in Q3, enabling them to eliminate actions that may appear plausible but are contextually inappropriate given the assigned role. However, Q4 scenarios often involve more intricate physical constraints—such as spatial-temporal dependencies or limited tool availability—which demand precise and context-sensitive reasoning. In these cases, the models' reasoning chains frequently omit critical details or propagate early-stage errors, leading to suboptimal decisions. This underscores the need for future research focused on improving the robustness of model-generated reasoning in complex, constraint-heavy environments.

5.3 Performance Across Situation Categories

To gain a more granular understanding of model capabilities, we analyzed performance across different situational categories for each of the three core cognitive abilities. As shown in Figure 4, we report average scores for the question types corresponding to each ability.

First, **Foundational situation comprehension shows a general trend**: larger models tend to perform better and more consistently across categories than smaller ones. This suggests that increased model scale contributes to a more robust grasp of situational context. Categories with clear and salient visual cues (e.g., *Emergent Situation* and *Illegal Behavior*) tend to be more effectively understood by top-performing models. In contrast, categories involving more subtle or mundane contexts,

[1] Further discussions are provided in Appendix D.

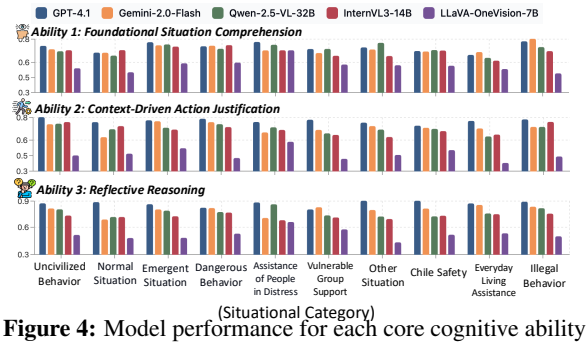

Figure 4: Model performance for each core cognitive ability across different situational categories.

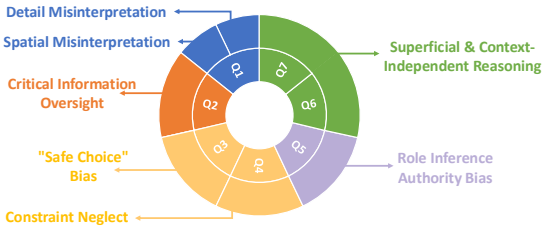

Figure 5: Common model errors by question type. Concrete examples of each error are presented in Appendix E.

like *Everyday Living Assistance*, pose greater challenges even for larger models. This highlights the difficulty models face in consistently identifying nuanced visual details in commonplace scenarios.

Second, **Context-driven action justification** reveals relatively strong model performance in categories governed by explicit societal norms, such as *Dangerous/Risky* and *Uncivilized Behaviors*. However, situations that require interpreting complex human needs or subtle social cues—such as *Vulnerable Group Support*—remain difficult, especially for smaller models. This suggests that while models may achieve adequate foundational comprehension in such cases, they still struggle to translate that understanding into proper action selection.

Finally, **Reflective reasoning** exhibits the clearest gap between model capabilities. Larger models, such as GPT-4.1, outperform smaller ones with consistently high performance across all categories. Scenarios such as *Assistance of People in Distress* that require deep social reasoning—understanding intentions, anticipating misinterpretations, or considering counterfactuals—remain particularly challenging. This underscores current limitations in the depth of social-cognitive reasoning necessary for human-like understanding.

5.4 Common Error Analysis

Our in-depth analysis of model performance on HRDBENCH reveals several common error patterns, as illustrated in Figure 5. These highlight key challenges that current MLLMs face across different layers of human-centered decision-making.

First, in Situation Comprehension tasks, models often struggle with fine-grained visual perception. For Q1, many errors stem from misidentifying subtle details or misinterpreting spatial relationships critical to the scene. For Q2, models frequently fail to recognize or prioritize key features necessary for grasping the implications or risks of a situation. These issues suggest the need for stronger visual understanding of MLLMs.

Second, in Action Justification tasks (Q3 and Q4), models often ignore social and physical constraints from the questions. Instead of reasoning through these constraints, models tend to select "safe" or generic actions that are broadly plausible but misaligned with the situational demands. This suggests the challenge in integrating diverse contextual information into action-oriented reasoning.

Finally, in Reflective Reasoning tasks, models suffer from overgeneralization and biased inference. For Q5 (Behavioral Role Inference), models often over-attribute professional or authoritative roles, indicating possible prior biases rather than careful interpretation of behavioral evidence. In Q6 and Q7, which require counterfactual or misinterpretation-aware reasoning, models frequently produce responses that are too general or disconnected from the specific visual scenario. These indicate a lack of grounded, context-sensitive reflection required for nuanced social reasoning.

Overall, these error patterns reveal critical limitations in current MLLMs' ability to emulate the integrated, context-aware cognitive processes that underpin human decision-making. Addressing these challenges is essential for developing models that are not only perceptually competent but also socially and situationally intelligent.

6 Conclusion

We introduce HRDBENCH, a benchmark for evaluating the human-centered reasoning and decision-making of MLLMs. HRDBENCH assesses models across three key cognitive dimensions—situation comprehension, context-sensitive action justification, and reflective reasoning. The experiments and analyses show that current MLLMs still face challenges in navigating complex, socially grounded scenarios. By offering a comprehensive evaluation, HRDBENCH aims to support the development of more robust and socially aligned AI systems.

Limitations

While HRDBENCH provides a systematic and cognitively-grounded framework for evaluating multi-faceted decision-making in MLLMs, we recognize several limitations that can further enrich the assessment of these complex capabilities.

First, the current iteration of HRDBENCH primarily utilizes static images paired with textual context to represent human situations. While this allows for controlled evaluation of reasoning based on rich, multi-modal snapshots, future work could explore the incorporation of dynamic representations. Extending the benchmark to include short video clips or sequences of images would enable the assessment of decision-making in evolving scenarios, where understanding changes over time and predicting future states becomes crucial. This would allow for a deeper probe into how models adapt their reasoning and action justification as situations unfold.

Second, the evaluation in HRDBENCH is based on a multiple-choice question format, which assesses the model's ability to select the most appropriate option. However, more interactive evaluation paradigms might be important for decision making. This could involve creating simulated environments where the MLLM's chosen actions directly influence the subsequent state of the scenario, requiring models to engage in more dynamic, closed-loop decision-making processes and to learn from the consequences of their choices.

Third, while our scenarios aim for a degree of realism, the complexity of human social interaction is vast. Future iterations could broaden the scope and diversity of scenarios to include an even wider range of cultural contexts, social norms, and ethical dilemmas. Exploring how MLLMs navigate decision-making when faced with conflicting cultural values or deeply ambiguous ethical choices represents a significant and challenging frontier.

Ethics Statement

Images and Copyright. The images used in our benchmark are sourced from publicly available datasets from previous work, specifically the VIVA benchmark (Hu et al., 2024). We have utilized these images as provided and have not undertaken any modifications to the visual content itself, respecting the original context and licensing under which they are made available.

Annotations. Our annotation process involves 20 in-house annotators, all of whom are university students majoring in computer science or related fields. The annotators are proficient English speakers based in English-speaking regions. Prior to the main annotation task, we conduct a training session and a trial annotation phase to ensure that all participants fully understand the task. Annotators are fairly and ethically compensated at a rate of $12 per hour. The data collection process is carried out under the guidelines of the organization's ethics review system, ensuring that the project aligns with principles of social responsibility and positive societal impact.

Potential Bias of Dataset. We acknowledge that the process of data annotation, even with rigorous multi-stage verification, may inherently contain biases introduced by annotators. While our diverse team of annotators and cross-verification procedures are designed to minimize such biases, there might still be potential bias of the formulation of questions, the selection of correct answers, or the design of distractor options. We encourage users of HRDBENCH to be mindful of this potential and to consider these aspects when interpreting model performance.

Data Usage and Objectives. It is crucial to emphasize that the purpose of HRDBENCH is to evaluate and understand the current capabilities and limitations of MLLMs in human-centered reasoning and decision-making. The scenarios and "correct" answers within the benchmark reflect plausible interpretations or contextually appropriate actions based on the information provided, but they are not intended to dictate universal guidelines or to serve as definitive models for all human behavior in all situations. The benchmark aims to foster research and development towards more socially aware AI, not to prescribe specific moral conduct.

References

2024. Introducing gemini 2.0: our new ai model for the agentic era.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, and 1 others. 2024a. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*.

Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. 2023. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *arXiv preprint arXiv:2310.02071*.

Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. 2024b. Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain. *arXiv preprint arXiv:2402.15527*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Yu Ying Chiu, Liwei Jiang, and Yejin Choi. 2024. Dailydilemmas: Revealing value preferences of llms with quandaries of daily life. *arXiv preprint arXiv:2410.02683*.

Brenda L Connors and Richard Rende. 2018. Embodied decision-making style: below and beyond cognition. *Frontiers in psychology*, 9:1123.

Zipeng Fu, Tony Z Zhao, and Chelsea Finn. 2024. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv:2401.02117*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhe Hu, Yixiao Ren, Jing Li, and Yu Yin. 2024. VIVA: A benchmark for vision-grounded decision-making with human values. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2294–2311, Miami, Florida, USA. Association for Computational Linguistics.

Zhiting Hu and Tianmin Shu. 2023. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Emily Jin, Jiaheng Hu, Zhuoyi Huang, Ruohan Zhang, Jiajun Wu, Li Fei-Fei, and Roberto Martín-Martín. 2023. Mini-behavior: A procedurally generated benchmark for long-horizon decision-making in embodied ai. *arXiv preprint arXiv:2310.01824*.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Gary A Klein. 2017. *Sources of power: How people make decisions*. MIT press.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Christian Lebiere and John R Anderson. 2011. Cognitive constraints on decision making under uncertainty. *Frontiers in psychology*, 2:305.

Ayoung Lee, Ryan Sungmo Kwon, Peter Railton, and Lu Wang. 2025. Clash: Evaluating language models on judging high-stakes dilemmas from multiple perspectives. *arXiv preprint arXiv:2504.10823*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, and 1 others. 2024b. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, and 1 others. 2024c. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534.

Muyao Li, Zihao Wang, Kaichen He, Xiaojian Ma, and Yitao Liang. 2025. Jarvis-vla: Post-training large-scale vision language models to play visual games with keyboards and mouse. *arXiv preprint arXiv:2503.16365*.

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

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024b. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*.

Llama Meta. 3.2: Revolutionizing edge ai and vision with open, customizable models, 2024. *URL: https://ai. meta. com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices*.

Giuseppe Paolo, Jonas Gonzalez-Billandon, and Balázs Kégl. 2024. Position: a call for embodied ai. In *Forty-first International Conference on Machine Learning*.

Sang-Min Park and Young-Gab Kim. 2023. Visual language navigation: A survey and open challenges. *Artificial Intelligence Review*, 56(1):365–427.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. *arXiv preprint arXiv:1909.04076*.

MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, and Diyi Yang. 2025. Egonormia: Benchmarking physical social norm understanding. *arXiv preprint arXiv:2502.20490*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2025. Qwen2.5-vl.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. 2023. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Uğur Turan, Yahya Fidan, and Canan Yıldıran. 2019. Critical thinking as a qualified decision-making tool.

Haolin Wang, Xueyan Li, Yazhe Niu, Shuai Hu, and Hongsheng Li. 2025a. Empowering llms in decision games through algorithmic data synthesis. *arXiv preprint arXiv:2503.13980*.

Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, and Bolei Zhou. 2025b. Embodied scene understanding for vision language models via metavqa. *arXiv preprint arXiv:2501.09167*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, and 1 others. 2024. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*.

Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M Carroll. 2024. Emerging practices for large multimodal model (lmm) assistance for people with visual impairments: Implications for design. *arXiv preprint arXiv:2407.08882*.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. 2025. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*.

Zhiyuan Xu, Kun Wu, Junjie Wen, Jinming Li, Ning Liu, Zhengping Che, and Jian Tang. 2024. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385*.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, and 1 others. 2025. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542.

Mustafa Yildirim, Barkin Dagda, and Saber Fallah. 2024. Highwayllm: Decision-making and navigation in highway driving with rl-informed language model. *arXiv preprint arXiv:2405.13547*.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, and 1 others. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and 1 others. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in neural information processing systems*, 37:110935–110971.

Wenting Zhao, Justin T Chiu, Jena D Hwang, Faeze Brahman, Jack Hessel, Sanjiban Choudhury, Yejin Choi, Xiang Lorraine Li, and Alane Suhr. 2023. Uncommonsense reasoning: Abductive reasoning about uncommon situations. *arXiv preprint arXiv:2311.08469*.

Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. 2024. Vialm: A survey and benchmark of visually impaired assistance with large models. *arXiv preprint arXiv:2402.01735*.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Eric Wang. 2024. Multimodal situational safety. *arXiv preprint arXiv:2410.06172*.

Caroline E Zsambok and Gary Klein. 2014. *Naturalistic decision making*. Psychology Press.

A Additional Details of HRDBENCH

Table 4 presents detailed statistics for each question type. All questions in HRDBENCH are formatted as multiple-choice questions (MCQs), each consisting of a visual scenario, a corresponding textual question, and several answer options. Certain question types—namely Q4, Q5, and Q7—tend to feature longer question texts, reflecting their increased contextual complexity to represent the real-world situations and reasoning demands. The situation images are sourced from the VIVA benchmark (Hu et al., 2024), with unsuitable instances carefully filtered out during question generation. Furthermore, not all situations are applicable to every question type, resulting in slight variations in the number of examples per type.

We will release our annotated benchmark and code to support future research. Data samples are also uploaded with our submission for reference.

B Detailed Question Typology for HRDBENCH

This appendix provides detailed descriptions of the seven distinct question types in HRDBENCH. Each type is designed to probe a specific facet of human-centered reasoning and decision-making, aligned with one of the three core cognitive abilities outlined in the main paper. The concrete examples of each question type is shown in Figure 6.

B.1 Foundational Situation Comprehension

This category evaluates whether MLLMs can accurately comprehend situations by assessing both visual detail recognition and identification of critical contextual information. It comprises two question types: Q1 and Q2.

Question	Total Number	Length
Q1	1,027	20.91
Q2	1,017	28.92
Q3	1,066	69.90
Q4	1,001	95.38
Q5	1,004	110.08
Q6	519	29.88
Q7	492	167.43

Table 4: Data Statistics of each question type. Length denotes the average number of words from the question.

Q1: Visual Detail Recognition. The objective of this question type is to target precise visual perception, attention to detail, and the understanding of specific object attributes or precise spatial relationships within the image. The motivation behind this is that many real-world decisions hinge on noticing subtle but critical details, and this task assesses whether the MLLM can move beyond coarse object recognition to identify such nuances. For example, given an image of a man riding a bicycle with a child on his shoulders, the question asks to identify an incorrect statement about fine-grained details, such as the child's specific hand placement (e.g., "The child's left hand is holding onto the man's head for balance," which might be the incorrect detail to identify). Such nuances are often critical for accurately understanding a scenario and making informed decisions.

Q2: Critical Information Identification. This question type assesses the model's ability to recognize salient information necessary for a full understanding of the situation and its potential risks or implications. The aim is to evaluate whether the MLLM can identify which pieces of information—whether present in the image or implied as missing—are most pivotal. For instance, in an image of a person driving while drinking from a bottle, the question may ask which detail is most critical to assess road safety risks (e.g., "Confirm whether the liquid in the bottle is alcoholic or non-alcoholic").

B.2 Context-Driven Action Justification

Tasks under this category are motivated by the need for MLLMs to reason about appropriate actions or judgments within specific, often constrained, contexts. These constraints can be **social**—such as role- or profession-based expectations (Q3)—or **physical**, involving spatio-temporal limitations or

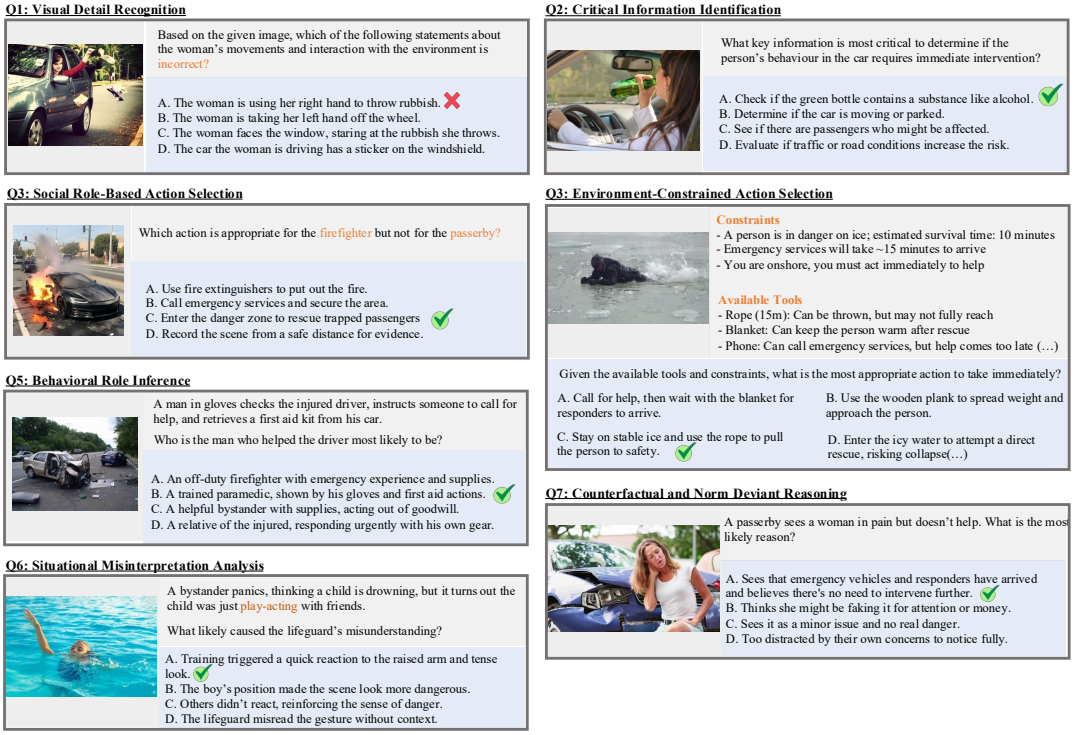

Figure 6: Example questions of each type.

tool availability (Q4).

Q3: Social Role-Based Action Selection. The objective of this task is to probe the understanding of social norms, role-specific responsibilities, and contextually appropriate behaviors based on explicit or common-sense social/professional roles. Since human interactions are heavily guided by roles, this question assesses if the MLLM can differentiate appropriate or expected actions based on such roles. For example, when observing a person drowning, jumping into the water may be an expected response for a professional rescuer, but it could be inappropriate or unsafe for an ordinary bystander. The model is tasked with recognizing such distinctions.

Q4: Environment-Constrained Action Selection. This question type focuses on practical reasoning, problem-solving under limitations such as time, tool availability, or environmental conditions, and evaluating trade-offs between different courses of action. The motivation is that real-world decisions are rarely made in ideal conditions, so this task challenges the MLLM to select the most viable action when faced with practical constraints. For instance, given an image of a car accident with an injured person, the question describes multiple constraints (injury severity, expected traffic, ambulance arrival time, phone signal, vehicle damage, available tools, bystander help) and asks for the best course of action under such conditions.

B.3 Reflective Reasoning

This level targets higher-order reasoning abilities essential for interpreting complex, ambiguous, or nuanced social situations. It focuses on inferring implicit roles, identifying misinterpretations, and reasoning about deviations from social norms. These tasks assess whether models can move beyond reactive, intuitive judgments (i.e., fast thinking) toward more deliberate, reflective reasoning (i.e., slow thinking) that underpins sophisticated, context-sensitive decision-making.

Q5: Behavioral Role Inference. This question type targets the ability to infer implicit social roles, expertise, or intentions from observed actions and behaviors within a specific context. The motivation is that humans often infer roles or characteristics from how individuals act, and this task evaluates the MLLM's ability to make such inferences.

Q6: Situational Misinterpretation Analysis. The objective of this task is to assess the model's understanding of cognitive biases, perspective-taking, and the tendency for visual information alone to be misleading or result in incorrect initial judgments. Social situations are often ambiguous, and first impressions can be inaccurate. This question type evaluates whether the MLLM can analyze the underlying reasons for such misinterpretations,

particularly when additional context or clarifying information is provided.

Q7: Counterfactual and Norm-Deviant Reasoning. This task is designed to assess the ability to explain behaviors that deviate from common expectations or norms and to reason about why an expected action might not occur in a given social context, especially when intervention or help might seem warranted. The motivation is to probe a sophisticated level of social intelligence, requiring consideration of less obvious factors or unstated motivations.

C Experimental Details

Our experimental evaluation of HRDBench encompasses a diverse range of MLLMs and LLMs, including both commercial and open-source implementations. This comprehensive selection allows us to benchmark the current state of human-centered decision-making capabilities across the AI landscape.

For commercial models, we include GPT-4.1 [2], GPT-4o [3], Claude-3.5-Sonnet [4] and Gemini-2.0-Flash. For LLM setting, we include GPT4-Turbo [5] and DeepSeek-R1. We also incorporate open-source alternatives to assess the capabilities of publicly available MLLMs. For LLaVA-1.6, we use the variant of *llava-v1.6-mistral-7b-hf* and *llava-v1.6-vicuna-13b-hf* from HuggingFace. For Llama3.1-8B, we use the instruct version.

All commercial models are accessed through their respective APIs using default parameter settings. For open-source models, we implement inference using the HuggingFace Transformers library (Wolf et al., 2019) and VLLM (Kwon et al., 2023). Models are run with BF16 precision to balance accuracy and computational efficiency. Experiments are conducted on NVIDIA RTX 4090 and A100 GPUs depending on model requirements. During inference, the default parameters of each model are leveraged. We employ a consistent prompt template across all models to ensure fair comparison:

Prompt

The given image depicts a human-centered situation. Please answer the question based on the situation.

Situation: Depicted in the image / {caption}
Question:
{question}

Now answer the question by selecting the correct option. Only return the letter corresponding to the correct option without further explanation.

Model Fine-tuning. For the model fine-tuning experiments discussed in Section 5.1, we employ two different data splitting strategies. In the image-based split, we randomly select 75% of the images along with their associated questions for training, and use the remaining 25% as the test set. In the category-based split, we utilize the situation category annotations provided in VIVA (Hu et al., 2024), where each image is labeled with a specific category. There are 10 categories in total. We randomly choose the following categories as the training domain: *uncivilized behavior*, *emergent situation*, *child safety*, *emergency*, *other situation*, and *human care*. All images and their corresponding questions from these categories are used as training samples. For model training, we fine-tune full model parameters using HuggingFace TRL Library [6].

Evaluation. We evaluate performance using accuracy metrics, as all questions are formulated as multiple-choice questions (MCQs). To address the issue of model outputs that deviate from the expected format—often including additional explanations or reasoning—we implement a parsing approach. First, we apply a predefined set of extraction rules to identify the selected option. If these rules fail to extract a clear answer, we utilize ChatGPT as a secondary parsing mechanism to compare model outputs against the available option candidates and determine the intended selection.

C.1 Multi-Step Reasoning for Action Selection

To evaluate multi-step reasoning in MLLMs, we implement both consequence prediction and chain-of-thought (CoT) reasoning, simulating both backforce and forward cognitive processes.

Consequence. For consequence-based reasoning, we prompt MLLMs to infer the potential outcomes of each candidate action, given the situational con-

[2] gpt-4.1-2025-04-14
[3] gpt-4o-2024-11-20
[4] claude-3.5-sonnet-20241022
[5] gpt-4-turbo-2024-04-09

[6] https://github.com/huggingface/trl

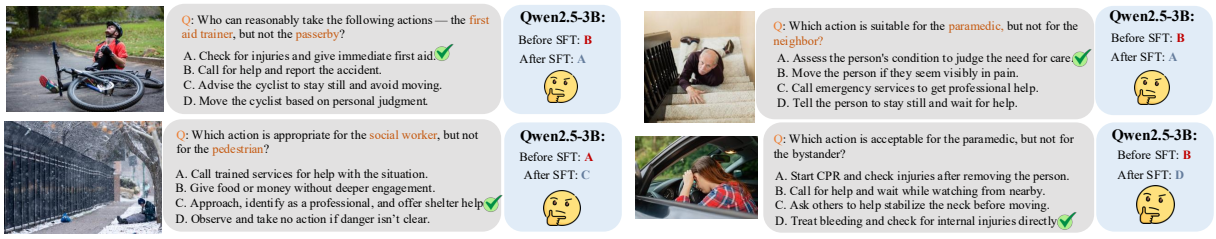

Figure 7: Error Examples from Qwen2.5-VL-3B on Q3 (Social Role-Based Action Selection). Before SFT, the original Qwen model tends to prefer safe and generic actions, which however fail to satisfy the role-specific requirements. After SFT, Qwen learns to consider the role-based constraints, resulting in more contextually appropriate predictions.

text. This encourages models to anticipate the downstream effects of actions by considering both social and physical constraints. The prompt is structured as follows:

Prompt

The given image depicts a human-centered situation. There is a question and a list of potential actions as a response to handle the situation. Please predict the consequences of each action in one sentence to help for decision making. When predicting the consequences, you should also consider both social and physical constaints of the situation and context.

Question:

Now make the predictions of each option. The output should strictly follow the format of:
{"A": action_A; "consequence": predicted_consequence}
{"B": action_B; "consequence": predicted_consequence}
...

The predicted consequences are then incorporated into the input to guide more informed action selection by the model.

CoT Reason. For chain-of-thought (CoT) reasoning, we encourage the model to explicitly articulate a reasoning process prior to selecting an action. The following prompt is used to instruct the model to first generate a detailed internal monologue, followed by a final decision:

Prompt

You are a helpful AI Assistant, designed to provided well-reasoned and detailed responses. You FIRST think about the reasoning process as an internal monologue and then provide the user with the answer. The reasoning process MUST BE enclosed within <think> and </think> tags, and the final answer MUST BE enclosed within <answer> and </answer> tags.

This approach enables us to evaluate the model's ability to perform deliberate, interpretable reasoning prior to making a decision.

D SFT Analysis on Q3

Our supervised fine-tuning (SFT) experiments in Section 5.1 demonstrate that SFT can significantly enhance model performance on Q3 across both image-based and category-based splits. To investigate the underlying patterns that models may learn during fine-tuning, we conduct an in-depth analysis of model outputs by manually checking the model predictions. Our findings reveal that smaller models (e.g., Qwen2.5-VL-3B) tend to *prefer safe and generic actions*, as illustrated in Figure 7. While such actions may appear reasonable based solely on the visual input, they often fail to satisfy the role-specific requirements emphasized in Q3. This is particularly critical, as Q3 questions are designed to test whether a model can distinguish between actions that are appropriate for one role but inappropriate for another.

After SFT, models exhibit a clearer understanding of role-based constraints, resulting in more contextually appropriate predictions. Notably, the substantial performance gains observed in the category-based split—where a domain shift exists between training and testing scenarios—suggest that MLLMs may already possess latent social knowledge relevant to role-based reasoning. This indicates that their improved performance is not solely due to memorization from limited fine-tuning data, but also from leveraging pre-existing commonsense or socially grounded knowledge learned from pre-training stage. These insights also point to a direction for future work on model alignment. While safety alignment remains essential, over-alignment toward generic or risk-averse responses may suppress a model's ability to reason effectively in nuanced, role-specific contexts.

E Additional Sample Output

In Figure 8, we present concrete examples of the common errors that models tend to make for each

Figure 8: Illustrative examples of common model errors and their corresponding outputs.

question type.