EmModBench: Benchmarking Cross-Modal Decision Consistency for Embodied Agents

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) are increasingly used as the cognitive core of embodied agents. In real deployments, agents often receive fragmented observations: the same decision-critical cue may appear as a raw visual signal in one episode, but as a structured textual message from upstream modules (maps or planners) in another. When each modality alone is sufficient, a reliable agent should make modality-invariant decisions, producing the same correct action across modalities. Existing benchmarks mainly test static knowledge retrieval or unimodal perception, leaving cross-modal decision consistency for embodied reasoning underexplored. To bridge this gap, we introduce **EmModBench**, a diagnostic benchmark for decision consistency under embodied input fragmentation. EmModBench evaluates TEXT-ONLY, IMAGE-ONLY, and IMAGE+TEXT settings, built via a rigorous pipeline with human auditing and strengthened by bidirectional verification. It contains 4,406 multiple-choice questions spanning five embodied competency domains, and our evaluation runs approximately 280,000 inference calls across 15 representative VLMs. Results show multimodal fusion is not reliably beneficial and can even hurt due to modality interference; models may also be consistently wrong across modalities, indicating brittle cross-modal grounding even with high unimodal accuracy.

1 Introduction

Vision-language models (VLMs) are increasingly serving as the cognitive core of embodied agents, interpreting multimodal observations to generate actions and plans (Ahn et al., 2022; Driess et al., 2023; Huang et al., 2022). A practical yet under-evaluated challenge arises at deployment: embodied systems rarely receive a single, fixed-format input. In real-world modular architectures, the agent's input is rarely confined to a single sensory stream. Instead, it must process heterogeneous representations passed from upstream subsystems,such as topological graphs from a SLAM system or structured sub-goals from a high-level planner. Consequently, the agent must be capable of reasoning over both raw visual observations and the structured symbolic messages that represent them. While these inputs differ in modality, they can preserve the same decision-critical cues needed to determine the next action (a visual Stop sign vs. a textual command "Halt"). In such settings, **cross-modal decision consistency** becomes a safety-critical requirement: a reliable agent should yield the same correct action regardless of input format, as long as each modality alone provides sufficient decision-critical evidence to answer the decision point. This requirement is especially important under the prevailing paradigm

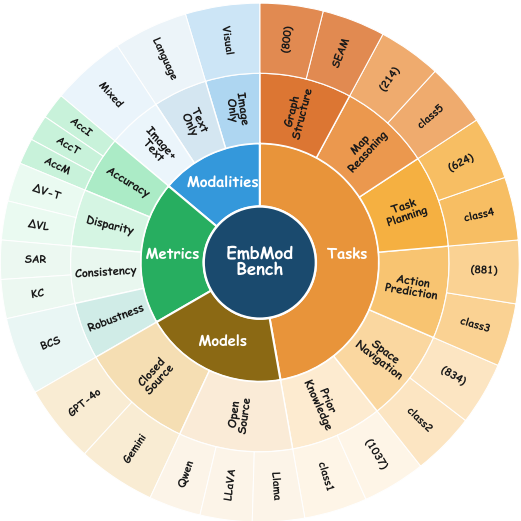

Figure 1: **Composition of EmModBench.**A multi-dimensional view of the benchmark's 4,406 samples across varying Tasks and Modalities, alongside the evaluated Models and the multi-perspective Metrics used to diagnose embodied decision consistency.

of embodied navigation and planning. Many modern frameworks operationalize long-horizon missions as sequences of atomic decision steps supported by external state (e.g., memory, maps, or structured representations) (Rana et al., 2023; Yin et al., 2024a, 2025). In systems such as SG-Nav and UniGoal, the agent observes the environment, queries or updates a stateful representation, and performs a discrete reasoning step to select the next action. Consequently, long-horizon reliability hinges on the robustness of these single-step decisions: even a modality-induced decision flip can cascade into inconsistent subgoals, corrupted state, and unsafe behavior. However, despite this modular structure, current evaluations often treat modality as part of a fixed task definition rather than a variable that probes the stability of the reasoning core under real deployment conditions.

While the community has begun to investigate cross-modal inconsistency, existing efforts leave a significant gap for embodied decision-making. Benchmarks for foundation models (e.g., PRISM, SEAM) predominantly focus on static knowledge and factoid QA, and do not reflect the goal-conditioned nature of agentic decisions (Wei et al., 2025; Tang et al., 2025). Conversely, end-to-end embodied benchmarks (Majumdar et al., 2024; Yang et al., 2025) evaluate full-stack performance (e.g., success rate), which is valuable but often conflates perceptual noise, control dynamics, and reasoning errors. As a result, it remains difficult to pinpoint whether a failure stems from partial observations or from modality-sensitive decision making. This motivates controlled diagnostic evaluations that isolate the agent's decision module and directly test whether it remains stable under input fragmentation.

To bridge this gap, we introduce **EmModBench**, a benchmark designed to diagnose cross-modal decision inconsistency in embodied scenarios. Unlike generic VQA, EmModBench frames each instance as a goal-conditioned decision point, spanning five competency domains: Prior Knowledge, Spatial Understanding, Next-Action Planning, Long-Horizon Mission Prioritization, and Structured Path Reasoning. Constructed via a rigorous pipeline with human auditing, each item is instantiated under three observation configurations (IMAGE-ONLY, TEXT-ONLY, and IMAGE+TEXT) to emulate fragmented sensing while preserving decision-critical content. **EmModBench** does not measure full closed-loop interaction (e.g., exploration, control, or compounding perception errors), but instead isolates the reasoning stability of the decision module.

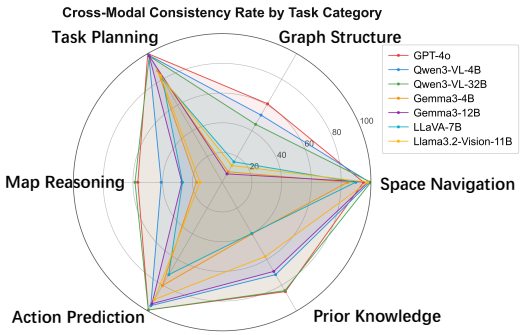

Figure 2: **Cross-modal consistency by task category.** Radar plot of cross-modal consistency rates (SAR) for different models across several EmModBench task families, measuring how often a model outputs the same option under different modal inputs. Note that high consistency does not imply correctness; models may exhibit "Consistent Failure" (high SAR but low Accuracy), reinforcing the need to evaluate consistency jointly with correctness metrics like KC (Knowledge Consistency).

EmModBench contains **4,406** unique multiple-choice questions (unique embodied decision points). Under our full evaluation protocol , each model is evaluated on **18,680** instances. Across **15** models, this amounts to approximately **280,000** inference calls.

In summary, our contributions are:

- **Task: embodied cross-modal decision consistency.** We formalize cross-modal decision consistency under embodied input fragmentation, highlighting its importance for long-horizon coherence and safety.
- **Benchmark: EmModBench.** We construct a benchmark of 4406 unique embodied decision points aligned across modalities and instantiated under three observation settings, spanning five embodied competency domains.
- **Evaluation and findings.** We provide a comprehensive evaluation spanning open-weight and proprietary SOTA models, revealing that modality fusion is non-monotonic and often degrades performance in smaller models. We also conduct an initial closed-loop pilot suggesting that offline consistency metrics may be indicative of online navigation stability.

2 Related Work

Embodied Navigation and Benchmarks. Embodied navigation relies on photorealistic simula-

tors (Matterport3D, Habitat, AI2-THOR; Chang et al., 2017; Savva et al., 2019; Kolve et al., 2022; Deitke et al., 2020) and standard evaluation suites. While classic benchmarks focus on instruction following (R2R, RxR, CVDN, REVERIE; Anderson et al., 2018; Ku et al., 2020; Thomason et al., 2019; Qi et al., 2020) or grounded execution (ALFRED, OpenEQA; Shridhar et al., 2020; Das et al., 2017; Majumdar et al., 2024), recent efforts like EmbodiedBench (Yang et al., 2025) evaluate foundation agents at scale. However, these end-to-end evaluations often conflate perceptual noise with reasoning failures, lacking fine-grained diagnosis of why a decision flips under different modalities.

Foundation Models for Embodied Agents. LLMs and VLMs increasingly serve as the cognitive core for agents, utilizing zero-shot planning (Huang et al., 2022), affordance alignment (SayCan, PaLM-E; Ahn et al., 2022; Driess et al., 2023), and structured scene representations (SayPlan, SG-Nav; Rana et al., 2023; Yin et al., 2024a). Other works explore improved reasoning–acting loops and uncertainty awareness (Song et al., 2023; Ren et al., 2023; Yao et al., 2023; Zawalski et al., 2025; Xiang et al., 2023). Despite these advances, agents remain brittle under conflicting observations, where linguistic priors often override visual grounding, leading to hallucinations and accumulated errors (Chakraborty et al., 2025; Guan et al., 2024; Yin et al., 2024b; Ji et al., 2023; Rawte et al., 2023a,b; Tonmoy et al., 2024; Rahmanzadehgervi et al., 2025).

Cross-Modal Consistency. Recent benchmarks explicitly probe consistency in foundation models. PRISM (Wei et al., 2025) and SEAM (Tang et al., 2025) reveal cross-modal knowledge inconsistencies, while XModBench (Wang et al., 2025) and others (Zhang et al., 2024; Zhu et al., 2024; Selch et al., 2025; Yan et al., 2025; Fu et al., 2025; Liu et al., 2024) systematically study conflict reasoning. However, these works predominantly focus on static knowledge QA. **EmModBench differs** by shifting the focus to embodied decision-making. Unlike prior work, we model input fragmentation (Text/Image/Fused) via a rigorous bidirectional verification protocol, targeting the specific reliability requirements of long-horizon navigation agents rather than just factual correctness.

3 Benchmark

In this section, we introduce **EmModBench** and describe how the benchmark is designed and constructed.Rather than evaluating end-to-end execution where control noise confounds reasoning errors, we aim to isolate and evaluate the **cross-modal decision reasoning** capability of multimodal models. EmModBench is a comprehensive multiple-choice question answering benchmark for embodied scenarios, aiming to evaluate multimodal models in terms of cross-modal decision-making and, crucially, cross-modal decision consistency. In real embodied systems, inputs at the same decision point are often fragmented: an agent may receive only RGB egocentric observations, only structured textual inputs (e.g., user instructions, semantic cues, object relations/scene graphs, or topological/room maps), or their combination. To reflect this reality under a unified evaluation protocol, each EmModBench instance is evaluated under three settings: **Text-only**, **Image-only**, and **Image+Text**.

EmModBench spans five embodied competency categories: prior knowledge and commonsense, spatial understanding and local navigation, task decomposition and next-action planning, long-horizon planning and mission prioritization, and structured path reasoning. The benchmark contains 4406 carefully curated QA instances. Each instance is evaluated under the three modality configurations, resulting in 18,680 total inference cases.

3.1 Design Principles

The key challenge in building EmModBench is to construct a cross-modal parallel evaluation set where each modality is individually sufficient to answer the decision point (i.e., preserves the decision-critical cues needed for the correct action). This ensures that any performance gap across modalities more likely reflects differences in a model's representation and reasoning, rather than missing essential evidence in the input.

We follow four principles:

1. **Decision-sufficient alignment.** Each question is answerable from either the image-only or text-only input; we validate this with near-ceiling human accuracy (Table 9).

2. **Decision-critical cue preservation.** Since cross-modal conversion inevitably introduces noise, we prioritize preserving decision-relevant entities, relations, and scene cues (rather than full information parity) via structured extraction,

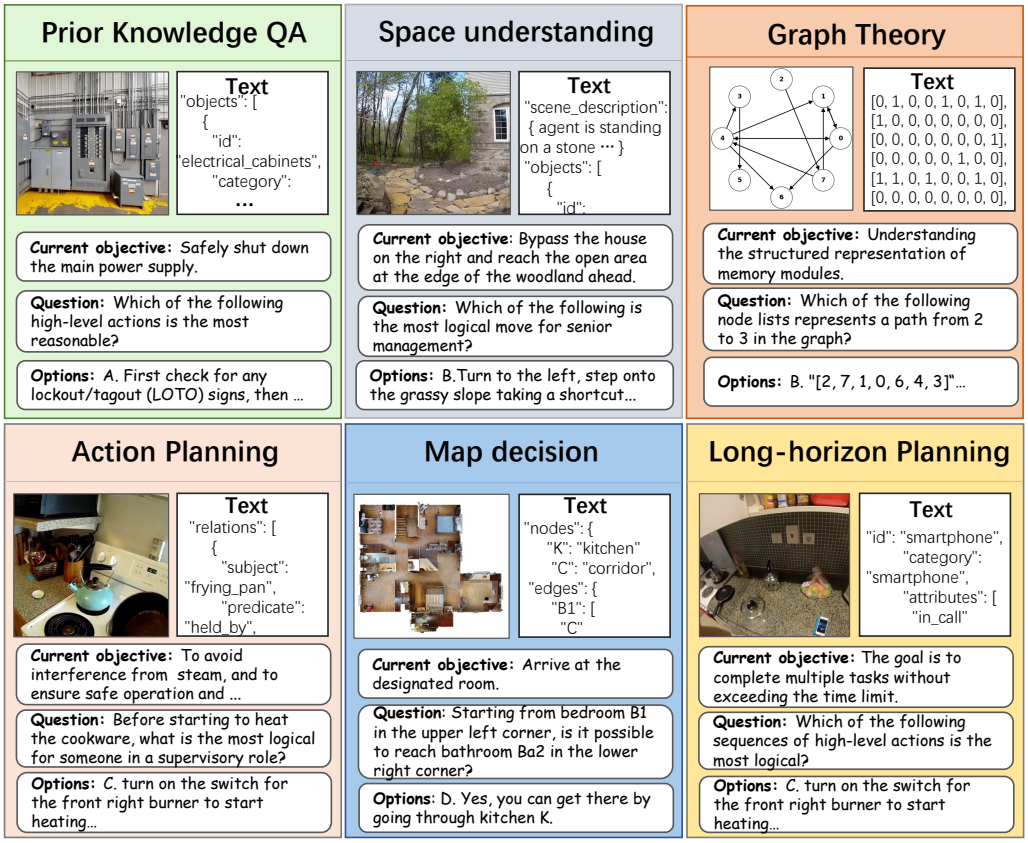

Figure 3: **Sample instances demonstrating competency domains of EmModBench.** Each block displays a unique embodied scenario with paired modalities: a Visual Input and a corresponding Structured Textual Proxy (e.g., scene object lists, topological adjacency). Crucially, the Questions and Options remain identical across modalities, strictly isolating the agent's ability to ground diverse inputs into the same goal-oriented decision.

cross-model verification, and human auditing. Our pipeline also includes human auditing on stratified subsets to reduce hallucination artifacts common in automated captioning.

3. **Diagnosability.** We adopt a unified multiple-choice format to decouple decision reasoning from action execution mechanics,and keep the question and options identical across modality settings, enabling systematic diagnosis of modality gaps and multimodal synergy/interference.

4. **Quality control.** We combine automated filtering with expert verification to remove ambiguous, under-specified, or cross-modally misaligned items.

To realize cross-modal alignment in practice, we employ three conversion routes: (1) simulator-backed structured extraction, (2) human-verified scene synthesis, and (3) structured data structure conversion. Simulator backends provide traceable object inventories and relations; caption synthesis generates global scene descriptions; and graph-based tasks naturally support bidirectional mappings between visual graphs and textual adjacency encodings, enabling more controlled, information-preserving mappings across modalities.

3.2 Dataset Construction

Instance format. Each EmModBench instance consists of: (i) an **image modality** observation (egocentric frame), (ii) a **text modality** context (a global caption plus a list of salient objects and their relations), (iii) a **goal/task context**, and (iv) a **multiple-choice QA pair** (question stem and four candidate actions/decisions). Since embodied agents act toward explicit objectives, we explicitly include the current goal to ensure that decisions are evaluated in a goal-conditioned manner.

Data sources. **EmModBench** is built from three major sources: **Matterport3D** simulator environments, **Ego4D** (an open-source egocentric vision dataset), and canonical map/graph topology resources. For indoor embodied scenes, we use Mat-

Table 1: **Overall results on EmModBench.** We report $Acc_{\mathcal{I}}$, $Acc_{\mathcal{T}}$, $Acc_{\mathcal{M}}$, Δ, SAR/KC, C_3/C_0, and Acc_{Rev}/BCS. Bold: best; ↑ higher is better, ↓ lower is better. Gray rows group model types.See Sec. 4.4 for metric definitions.

Model	**Forward Performance ↑**			**Disparity**		**Consistency ↑**		**Robustness**		**Understanding ↑**	
	$Acc_{\mathcal{I}}$	$Acc_{\mathcal{T}}$	$Acc_{\mathcal{M}}$	$\Delta_{V\text{-}T}$	Δ_{VL} ↑	SAR	KC	C_3 ↑	C_0 ↓	Acc_{Rev}	BCS
Open Weights Models											
Gemma 3 (4B)	65.9	69.9	69.3	-4.0	-0.5	67.0	70.4	56	11	56.7	52.1
Gemma 3 (12B)	77.2	78.4	79.2	-1.3	+0.7	77.7	81.0	70	7	76.3	74.7
Llama 3.2-Vision (11B)	76.1	76.5	72.5	-0.5	-4.0	74.4	77.3	65	9	55.9	54.1
LLaVA v1.5 (7B)	67.2	63.4	67.1	+3.8	-0.1	74.6	75.3	57	18	45.3	39.5
LLaVA v1.5 (13B)	70.0	61.9	67.3	+8.1	-2.7	64.3	67.0	53	11	52.4	46.1
Qwen2.5-VL (3B)	74.7	75.8	74.8	-1.1	-0.9	80.6	84.2	70	11	55.4	53.3
Qwen2.5-VL (7B)	79.4	79.2	81.0	+0.2	+1.7	84.7	87.4	75	10	79.6	78.7
Qwen3-VL (2B)†	39.1	18.1	40.7	+21.0	+1.6	25.5	37.1	12	13	22.6	20.2
Qwen3-VL (4B)	77.4	86.4	85.2	-9.0	-1.3	87.2	87.3	85	**2**	63.2	61.5
Qwen3-VL (8B)	77.0	85.0	87.0	-8.0	+2.0	84.7	84.6	82	**2**	68.4	66.8
Qwen3-VL (30B)	84.8	76.8	92.6	+8.0	**+7.8**	86.6	86.8	83	4	72.2	68.8
Qwen3-VL (32B)	88.3	85.7	95.4	+2.7	+7.1	**89.9**	**89.9**	**87**	3	74.5	68.1
Proprietary / Closed Models											
Gemini-3	**89.0**	87.2	95.0	+1.9	+6.0	88.6	88.9	**87**	**2**	**89.2**	**84.3**
GPT-4o	**89.0**	**88.2**	**95.6**	+0.8	+6.5	88.9	89.1	**87**	**2**	86.7	82.5
Qwen3-VL-Plus	84.9	81.8	91.5	+3.2	+6.5	88.6	89.5	84	5	86.1	79.8
Reference Baselines											
Human Baseline	98.9	98.6	99.8	+0.3	+0.9	99.2	99.8	98.9	0.02	97.8	95.6
Constant-A (Degenerate)	24.7	24.7	24.7	0.0	0.0	100	100	24.7	75.3	25.1	6.2

terport3D as the 3D environment source and render egocentric observations through the Habitat simulator. To obtain an aligned text modality for each view, we derive a structured representation consisting of an object list and coarse relations like room containment and basic spatial relations.

Ego4D frames do not naturally come with paired structured text. We therefore construct the textual modality via a human-validated synthesis pipeline. For the Ego4D subset (**1,505** instances; **34.15%** of **EmModBench**), we adopt a cross-verification procedure with two high-performing models of different architectures (**Claude Opus 4.5** as generator and **Grok-4.1** as verifier), and complement it with strict human auditing on a random **20%** sample to mitigate hallucinations and subtle misalignment. Further details of the pipeline, audit protocol, and robustness analyses are provided in Appendix D.

For structured planning tasks, we construct strictly aligned cross-modal representations: the *visual* input is a node–edge diagram (e.g., floorplan topology or abstract graphs), and the *text* input is an adjacency list or adjacency matrix encoding. This bidirectional mapping preserves information across modalities and enables controlled evaluation of cross-modal reasoning and consistency. See the Appendix D for more details on data preparation.

Question design and quality assurance. EmModBench is designed to probe *decision coherence* in complex embodied settings rather than mere object or scene recognition. Each question is constrained to be fully solvable under either single modality. Distractor options are designed to be challenging yet unambiguous. We adopt a two-stage QA pipeline: (1) automated filtering (rules and base-model screening) to remove under-specified, low-quality, or multi-answer items; and (2) expert verification with internal pilot testing to iteratively resolve ambiguities and cross-modal misalignment until the final item set meets our solvability and consistency criteria.Details in Table 9 and Appendix D.Human performance is near-ceiling across modalities (Table 9), suggesting items are decision-sufficient and largely unambiguous; ambiguous or multi-answer cases are removed during expert verification.

3.3 Task Taxonomy

EmModBench covers five key embodied competencies:

- **Prior Knowledge & Commonsense.** Tests scene-function understanding, object–place associations, feasibility judgments, and commonsense reasoning for goal-conditioned decisions.
- **Spatial Understanding & Local Navigation.** Tests spatial relations, safe navigation choices, path feasibility under constraints, and efficiency–safety trade-offs.
- **Task Decomposition & Next-action Planning.** Tests logical decomposition of multi-step tasks and the selection of locally optimal, constraint-aware next actions.
- **Long-horizon Planning & Prioritization.** Tests multi-goal coordination, prioritization under time/resource constraints, and long-horizon action sequencing.
- **Map/Graph Planning.** Tests reachability, path reasoning, shortest-path inference, and structured planning in strictly aligned cross-modal map/-graph representations.

4 Experiments

4.1 Research Questions

EmModBench is designed to diagnose the cognitive reliability of agents facing embodied input fragmentation. While real-world deployment requires handling incomplete or redundant inputs, the core decision module must remain coherent and safety-aware regardless of the input modality. We therefore structure our experiments around the following research questions.

RQ1: Single-modality robustness under sensor fragmentation. When an embodied agent receives only one modality at a decision point (e.g., camera occlusion, communication loss, only a symbolic/semantic map available), how does performance degrade under **image-only** ($\mathcal{I}$) and **text-only** ($\mathcal{T}$) settings?

RQ2: Fusion gain vs. modality interference. Does providing both modalities (**image+text**, $\mathcal{M}$) reliably improve decision accuracy, or can fusion introduce interference and reduce performance compared to the stronger unimodal input?

RQ3: Cross-modal decision consistency under decision-sufficient alignment. When each modality individually provides sufficient decision-critical evidence to determine the correct action, does the model output the same action across modalities? Inconsistency here directly translates to unpredictable embodied behavior.

RQ4: Safety-relevant failure profiles (collective hallucination vs. disagreement). When models fail, are they (i) collectively wrong (high-confidence, same wrong action across modalities), or (ii) inconsistently wrong (different wrong actions depending on modality)? These correspond to very different risks in embodied systems.

4.2 Models

Our evaluation strategy covers two distinct categories of models to provide a comprehensive analysis of the current landscape.

1. Efficient Open-Weight Models (Edge-Compatible). We target embodied decision modules intended for edge devices (e.g., mobile robots, on-device compute), where memory and latency budgets are strict. Accordingly, we evaluate a wide range of lightweight to mid-sized multimodal models: **Gemma 3** (4B/12B), **Llama 3.2-Vision** (11B), **LLaVA** (7B/13B), **Qwen2.5-VL** (3B/7B), and **Qwen3-VL** (2B/4B/8B/30B/32B).

2. Proprietary Frontier Models (Capability Ceiling). To determine whether cross-modal inconsistency is merely a symptom of limited model capacity or a fundamental issue in VLM alignment, we also evaluate state-of-the-art proprietary models: **GPT-4o**, **Gemini-3**, and **Qwen3-VL-Plus**. Evaluating these "ceiling" models allows us to diagnose whether scaling laws automatically resolve embodied decision inconsistency.

All evaluated models and their overall results are summarized in Table 1.

4.3 Evaluation Settings and Protocol

Three forward modality settings. We evaluate each instance under three settings: (1) **Image-only** ($\mathcal{I}$), where the agent receives the visual representation (egocentric view or diagram) and the question; (2) **Text-only** ($\mathcal{T}$), providing the structured textual representation (caption, object relations, or graph text) and the question; and (3) **Multimodal** ($\mathcal{M}$), which provides both image and text inputs. All instances are multiple-choice with four options

$\{A, B, C, D\}$, requiring the model to output exactly one option.We randomize option order per instance and verify that the gold-answer letter distribution is near-uniform (25% each); we also include a Constant-A degenerate baseline in Table 1 to expose option-prior exploitation.

Answer extraction for direct-answer models. we parse the final answer by strict regex matching of the last valid option letter in $\{A, B, C, D\}$. If parsing fails (e.g., the model outputs no valid option), the prediction is counted as incorrect.

4.4 Metrics

We employ a comprehensive set of metrics to evaluate correctness, stability, and safety.

Forward Performance & Disparity. For each setting $m \in \{\mathcal{I}, \mathcal{T}, \mathcal{M}\}$, we report standard accuracy Acc_m. We quantify modality dominance via $\Delta_{V\text{-}T} = Acc_{\mathcal{I}} - Acc_{\mathcal{T}}$ and fusion effectiveness via $\Delta_{VL} = Acc_{\mathcal{M}} - \max(Acc_{\mathcal{I}}, Acc_{\mathcal{T}})$, where $\Delta_{VL} < 0$ highlights **modality interference**.

Decision Consistency. We measure stability using **Simple Agreement Rate (SAR)**, defined as the proportion of instances where $\hat{y}^{(\mathcal{I})} = \hat{y}^{(\mathcal{T})} = \hat{y}^{(\mathcal{M})}$. However, SAR can be inflated by degenerate strategies (e.g., consistently outputting "A"). Following PRISM (Wei et al., 2025), we also report **Knowledge Consistency (KC)**: $KC = P(\text{all-correct} \mid \text{not-all-wrong})$, capturing reliable correctness among instances that are not uniformly failed across modalities.

Robustness & Safety. We categorize failure modes into **Correctness Consistency** ($C_3 = N_{\text{all-correct}}/N$) and **Collective Hallucination** (C_0), which counts instances where all modalities predict the same wrong option. A high C_0 indicates a risky lack of cross-modal error correction. For additional robustness evaluations, including our modality-conflict stress tests, see Appendix C.

Bidirectional Understanding. To verify grounding, we report **Reverse Accuracy** (Acc_{Rev}) via a context-retrieval task: given the question and its gold answer, the model selects the matching scene from four candidates (1 true + 3 distractors sampled from the same task family). We query each candidate independently with a 1–10 match score and take the argmax as prediction; candidate order is randomized. We combine this with forward performance into the **Bidirectional Consistency Score (BCS)**,defined as $BCS = P(\text{Fwd} \cap \text{Rev})$. Details of the reverse protocol are provided in Appendix B.

5 Main Results

Table 1 summarizes overall accuracy under three realistic input settings (*image-only* $\mathcal{I}$, *text-only* $\mathcal{T}$, and *image+text* $\mathcal{M}$), together with fusion gain (Δ_{VL}) and cross-modal agreement (SAR/KC). To understand where inconsistency comes from, we further break results down by task family (Table 2) and by pairwise disagreement rates (Table 7). Across all models, a consistent picture emerges: current LVLMs can be strong in local embodied perception, yet remain brittle when the system forces modality substitution, naive fusion, or self-verification.

RQ1: Single-modality robustness under sensor fragmentation. When an embodied agent receives only one channel at a decision step (camera degraded, or only a symbolic map is available), model behavior is far from uniform. We observe a clear task-dependent modality preference rather than a universal "vision-first" or "text-first" rule: vision tends to be more reliable for local spatial judgments and immediate action choices, while structured text is often more dependable for global/topological planning such as map/graph reasoning (Table 2). This implies that in real deployment, "fallback to any available modality" is not enough: agents likely need skill-aware routing; for example, use vision to stay safe locally, but switch to structured text when the problem is inherently topological or symbolic. Further task-level preferences are detailed in Appendix C.

RQ2: Fusion gain vs. modality interference. Contrary to the assumption that fusion monotonically improves accuracy, our results reveal a distinct divide: Frontier proprietary models achieve genuine synergy, showing consistent positive fusion gains ($\Delta_{VL} > 0$). However, many lightweight open-weight models pay a noticeable "fusion tax," where adding the second modality hurts performance (negative Δ_{VL}). Qualitatively, this suggests that while advanced models can effectively arbitrate conflicting cues, smaller models often struggle with decision boundaries when channels are concatenated. For on-device embodied systems, the takeaway is practical: always-on fusion is not guaranteed to be safer; conditional fusion (gating by confidence) is a more robust default. We further

Table 2: Comprehensive Macro Accuracy Heatmap. Colors indicate performance levels: Green (>80%), Yellow (60-80%), Orange (40-60%), and Red (<40%). SEAM is categorized under Task 5.

Model	Overall Avg			Task 1 (Perc)			Task 2 (Space)			Task 3 (Next)			Task 4 (Mission)			Task 5 (Map & Graph)					
				Class 1			Class 2			Class 3			Class 4			*Sub 1: Map*			*Sub 2: Graph*		
	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T
Gemma3-12B	67.2	71.8	71.7	78.6	77.2	79.8	83.9	84.2	84.0	85.7	86.8	86.6	90.9	91.0	91.0	34.1	64.5	58.9	29.8	27.1	30.1
Gemma3-4B	55.8	58.3	58.5	51.4	57.6	56.3	68.1	73.3	72.9	72.3	75.5	74.7	76.8	75.8	75.8	36.9	42.5	45.3	29.2	25.2	25.8
Llama3.2-11B	62.1	61.2	57.9	74.5	74.4	65.5	77.7	77.6	77.6	74.5	74.6	73.8	76.4	77.6	70.4	42.1	32.2	24.8	27.5	31.0	35.4
LLaVA-13B	57.7	51.3	54.5	63.4	49.3	55.6	72.9	69.1	71.9	69.0	67.7	69.9	73.3	65.6	71.4	42.5	31.3	31.8	25.0	24.6	26.1
LLaVA-7B	54.5	51.9	55.3	59.0	51.0	59.4	70.7	69.4	70.1	61.6	60.3	61.1	69.5	65.4	68.6	45.3	42.1	50.0	21.1	23.4	22.4
Qwen2.5-VL-3B	61.4	63.3	62.5	72.3	73.1	71.1	70.5	71.8	70.4	77.5	77.5	76.9	77.4	76.7	76.8	43.0	51.9	49.1	28.0	28.7	30.9
Qwen2.5-VL-7B	69.7	70.7	72.7	82.7	81.4	84.7	84.2	83.9	83.9	90.7	90.2	90.7	93.4	93.2	93.3	37.4	44.9	49.5	29.9	30.6	33.8
Qwen3-VL-2B	39.1	18.1	40.7	51.7	12.5	46.7	50.3	20.4	52.1	50.3	21.4	47.7	54.4	27.1	45.4	11.8	10.1	11.5	4.1	17.4	16.1
Qwen3-VL-4B	64.3	78.3	77.0	82.1	80.4	81.0	80.7	79.1	79.3	83.9	82.6	83.1	80.2	80.1	80.3	36.0	77.6	79.0	22.6	69.9	59.6
Qwen3-VL-8B	65.4	79.9	81.2	80.7	76.4	78.7	82.3	80.5	81.3	91.2	89.8	90.4	80.5	80.5	79.9	35.0	87.4	86.0	22.6	64.9	71.1
Qwen3-VL-30B	85.7	78.6	92.7	83.5	81.9	83.5	85.5	85.7	85.1	94.8	93.4	92.8	91.3	91.0	91.3	47.8	83.4	82.8	19.2	40.4	65.5
Qwen3-VL-32B	89.1	86.8	95.5	89.8	89.0	91.3	87.1	88.5	88.3	93.1	94.4	94.1	90.7	87.4	89.7	59.9	91.7	90.4	28.1	44.2	64.8
Qwen3-VL-Plus	85.8	83.1	91.6	83.5	85.0	83.5	95.4	93.2	95.6	96.4	95.7	96.4	91.2	89.9	91.0	49.0	78.3	79.0	35.7	39.3	71.1
Gemini3	89.9	88.1	95.0	92.9	89.8	87.4	93.7	92.1	94.0	97.1	96.4	97.1	92.4	91.1	92.2	59.2	88.5	90.4	41.3	54.0	75.5
GPT-4o	89.9	89.1	95.5	92.9	89.8	88.2	93.8	93.9	93.5	97.1	97.1	97.1	92.1	89.1	92.2	59.2	91.1	92.4	34.6	56.4	73.9

diagnose the mechanism of this **attention dispersion**, via qualitative case studies (Figure 4) and attention map visualizations in Appendix A.

RQ3: Cross-modal decision consistency under decision-sufficient alignment. Even when each modality is constructed to be individually sufficient (i.e., contains the cues needed to determine the correct action), models can still behave inconsistently, and the consistency signal itself can be misleading; we observe two qualitatively different regimes. In the desired regime, models are both accurate and stable across modality substitutions (low disagreement rates in Table 7, with strong forward accuracy in Table 1), suggesting the agent is using modality-invariant abstractions. In the problematic regime, a model may look "stable" because it repeatedly outputs the same option across all inputs, yet that stability is not grounded in understanding, a degenerate strategy that produces consistent but wrong actions. Hence, consistency must be interpreted jointly with correctness, not as a standalone robustness guarantee.

Although some models exhibit high consistency (SAR), this visual stability can be misleading. As shown in Table 1, high SAR often masks "Collective Hallucination" (C_0)where models consistently predict the same wrong action. This gap between SAR and Knowledge Consistency (KC) indicates that models often rely on stubborn parametric priors rather than robust cross-modal grounding.

RQ4: Safety-relevant failure profiles in embodied decision making. Not all errors are equally dangerous for embodied agents. A particularly risky pattern is when all three inputs (image-only, text-only, and fused) lead to the same wrong action, because multimodal redundancy can no longer serve as an internal alarm. In contrast, when a model disagrees across modalities, the system can at least detect uncertainty and trigger safeguards (ask for clarification, slow down, or replan). Empirically, models vary widely in how often they exhibit "same-wrong" failures versus disagreement-driven failures (see the robustness columns in Table 1). This highlights an embodied-specific evaluation point: beyond average accuracy, we should prefer models whose failures remain detectable through cross-modal disagreement checks, since undetectable consistent mistakes are harder to mitigate without external supervision. A small-scale closed-loop pilot (Appendix E) provides preliminary evidence that these offline consistency metrics are associated with online navigation stability.

6 Conclusion

We introduce **EmModBench** to diagnose a critical yet overlooked requirement for embodied agents: cross-modal decision consistency at decision points. Our comprehensive evaluation reveals that strong unimodal performance does not guarantee reliable grounding—multimodal fusion can be non-monotonic, sometimes resulting in unexpected performance degradation rather than the anticipated gain. We hope this benchmark serves as a cognitive stress test, guiding the development of future agents that are not only accurate but also robustly aligned across sensory modalities.

7 Limitations

EmModBench is designed as a controlled diagnostic for cross-modal decision consistency. While it provides precise signals on cognitive reliability, we outline three boundary conditions of our current scope.

Diagnostic Isolation vs. Online Dynamics. We deliberately adopt an offline, discrete decision protocol to isolate cognitive reasoning errors from the noisy control dynamics of continuous simulation. While this design enables precise "unit testing" of the decision module, it does not capture the temporal error accumulation characteristic of closed-loop execution (e.g., how a small inconsistency drifts into a navigation failure over time). Mainstream embodied navigation evaluations often operate on image streams and interactive rollouts (e.g., R2R/VLN-style tasks, ObjectNav, CVDN). We position EmModBench as a complementary diagnostic to end-to-end benchmarks (like R2R or ObjectNav), focusing on the cause (inconsistent reasoning) rather than the symptom (trajectory failure).

Information Density Asymmetry. Although we enforce **decision-critical alignment** via a rigorous human-in-the-loop pipeline, we do not claim strict information parity between an image and a text description. Visual inputs contain dense, unstructured signals (e.g., lighting, texture) that structured text—even when expert-verified—must inherently compress. Therefore, our consistency metric measures stability under decision-sufficient alignment (answer-critical cue preservation), rather than pixel-level information preservation.

Sensitivity to Textual Phrasing. Our text modality relies on high-quality descriptions with human auditing. However, current VLMs can be sensitive to subtle stylistic variations like list formatting or synonym choices. While our focus is on semantic grounding, we acknowledge that inconsistent behavior could partially stem from prompt fragility. Future work could incorporate adversarial text perturbation to further disentangle grounding failures from prompt overfitting. We view these limitations as opportunities for future extensions, including tighter structured-text alignment, stream-based evaluation, and multi-source text generation to stress-test robustness.

8 Ethical considerations

We strictly adhered to ethical guidelines for research involving human subjects.

- **Voluntary Participation:** All annotators were recruited voluntarily. No academic credits, grades, or professional evaluations were tied to their participation.
- **Fair Compensation & Workload:** The study was conducted with a manageable workload distributed over multiple sessions to prevent fatigue, ensuring it did not interfere with the students' primary research duties.
- **Content Safety:** The dataset consists of simulated household environments (e.g., Matterport3D, Ego4D) and contains no personally identifiable information (PII) or offensive content.

Dataset Usage and Privacy. Our benchmark incorporates visual data from the Ego4D dataset. We strictly adhere to the dataset's licensing agreements. While these egocentric videos depict human hands performing manipulation tasks,we take care to avoid personally identifiable information (PII) in the selected frames. The presence of human hands serves as a standard proxy for egocentric agent manipulation in current embodied AI research, given the scarcity of large-scale, diverse robotic interaction datasets.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, and 26 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *Preprint*, arXiv:2204.01691.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *Preprint*, arXiv:1711.07280.

Trishna Chakraborty, Udita Ghosh, Xiaopan Zhang, Fahim Faisal Niloy, Yue Dong, Jiachen Li, Amit K. Roy-Chowdhury, and Chengyu Song. 2025. Heal: An empirical study on hallucinations in embodied agents driven by large language models. *Preprint*, arXiv:2506.15065.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *Preprint*, arXiv:1709.06158.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2017. Embodied question answering. *Preprint*, arXiv:1711.11543.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. 2020. Robothor: An open simulation-to-real embodied ai platform. *Preprint*, arXiv:2004.06799.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: An embodied multimodal language model. *Preprint*, arXiv:2303.03378.

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

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *Preprint*, arXiv:2201.07207.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. 2022. Ai2-thor: An interactive 3d environment for visual ai. *Preprint*, arXiv:1712.05474.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methodsin Natural Language Processing (EMNLP)*, pages 4392–4412, Online. Association for Computational Linguistics.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, and 5 others. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. *Preprint*, arXiv:1904.10151.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2025. Vision language models are blind: Failing to translate detailed visual features into words. *Preprint*, arXiv:2407.06581.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. 2023. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. *Preprint*, arXiv:2307.06135.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M Towhidul Islam Tonmoy, Aman Chadha, Amit P. Sheth, and Amitava Das. 2023a. The troubling emergence of hallucination in large language models – an extensive definition, quantification, and prescriptive remediations. *Preprint*, arXiv:2310.04988.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023b. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. *Preprint*, arXiv:2307.01928.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A platform for embodied ai research. *Preprint*, arXiv:1904.01201.

Lukas Selch, Yufang Hou, M. Jehanzeb Mirza, Sivan Doveh, James Glass, Rogerio Feris, and Wei Lin. 2025. Prismm-bench: A benchmark of peer-review grounded multimodal inconsistencies. *Preprint*, arXiv:2510.16505.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. *Preprint*, arXiv:1912.01734.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *Preprint*, arXiv:2212.04088.

Zhenwei Tang, Difan Jiao, Blair Yang, and Ashton Anderson. 2025. Seam: Semantically equivalent across modalities benchmark for vision-language models. *Preprint*, arXiv:2508.18179.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. *Preprint*, arXiv:1907.04957.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Xingrui Wang, Jiang Liu, Chao Huang, Xiaodong Yu, Ze Wang, Ximeng Sun, Jialian Wu, Alan Yuille, Emad Barsoum, and Zicheng Liu. 2025. Xmodbench: Benchmarking cross-modal capabilities and consistency in omni-language models. *Preprint*, arXiv:2510.15148.

Mingjie Wei, Wei-Nan Zhang, Chen Zhang, Yifeng Ding, Donglin Di, Lei Ren, Wei Chen, and Ting Liu. 2025. Prism: A benchmark for unveiling cross-modal knowledge inconsistency in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11121–11129, New York, NY, USA. Association for Computing Machinery.

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. 2023. Language models meet world models: Embodied experiences enhance language models. *Preprint*, arXiv:2305.10626.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. 2025. Multimodal inconsistency reasoning (MMIR): A new benchmark for multimodal reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18829–18845, Vienna, Austria. Association for Computational Linguistics.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. 2024a. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Preprint*, arXiv:2410.08189.

Hang Yin, Xiuwei Xu, Lingqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. 2025. Unigoal: Towards universal zero-shot goal-oriented navigation. *Preprint*, arXiv:2503.10630.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024b. Woodpecker: hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12).

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. 2025. Robotic control via embodied chain-of-thought reasoning. *Preprint*, arXiv:2407.08693.

Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2024. Cross-modal consistency in multimodal large language models. *Preprint*, arXiv:2411.09273.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *Preprint*, arXiv:2410.03659.

A Qualitative Analysis and Case Studies

In this section, we analyze specific failure modes to understand why multimodal fusion does not always yield performance gains. A prevailing assumption in embodied AI is that integrating multiple sensory inputs (e.g., Vision + Text) monotonically reduces uncertainty. However, our experimental results, particularly with smaller-scale models (e.g., Qwen3-VL-8B), reveal a counter-intuitive phenomenon: **Modality Interference**.We select this example because it exhibits a typical *fusion tax* pattern ($\Delta_{VL} < 0$) and illustrates how high agreement can still coincide with incorrect decisions (Collective Hallucination, C_0).

We posit that when a model's cross-modal alignment mechanism is insufficient, the introduction of a second modality can shift the cognitive load from "information integration" to "conflict resolution," potentially distracting the model from clear unimodal signals.

A.1 Case Study: The Burden of Fusion

We illustrate this phenomenon using a representative sample from the *Long-Horizon Planning* category (Item ID: `003`).

Analysis of Prediction Error. As illustrated in Figure 4, the task imposes a strict goal: prioritize the immediate physical action (finishing the wash) under a time constraint. Crucially, Option 2 ("Check phone") is designed as a **Hallucination Probe**—a distractor involving an object that is factually absent from both the visual feed and the textual object list.

- **Unimodal Discriminative Robustness:** In unimodal settings, Qwen3-VL-8B demonstrates strong grounding capabilities that allow it to correctly reject this distractor.
  - In the $\mathcal{I}$**-only** setting, the high-frequency visual signals of "running water" and "active hand movements" serve as a **strong physical anchor**. The model perception is dominated by the ongoing interaction, leaving no latent space for hallucinating a non-existent phone.
  - Similarly, in the $\mathcal{T}$**-only** setting, the model relies on the strict boundaries of the structured object list. Since "phone" is explicitly missing from the provided inventory (which lists only sink, pan, and glove), the model utilizes **negative reasoning** to logically rule out Option 2 as an impossible action.
- **Multimodal Grounding Dilution:** However, in the $\mathcal{M}$ setting, we observe a paradoxical collapse in this discriminative reasoning. The introduction of cross-modal inputs does not reinforce the signal; instead, it introduces **alignment noise**. The cognitive effort required to map the dense textual description (containing background objects) to the visual scene appears to **dilute** the model's focus on the primary actor (the hands).

  In this state of heightened uncertainty, the model appears to rely less on the immediate evidence (the washing action) and more on a **parametric prior**—a generic, high-probability "displacement behavior" (checking a phone) often seen in pre-training data. Thus, multimodal fusion may introduce alignment noise that increases susceptibility to hallucinated affordances, coinciding with failures that do not occur in unimodal settings.

Mechanism of Interference: Attention Dispersion. Why does adding correct text lead to hallucination? We answer this by visualizing the model's attention in Figure 5.

- **Focused vs. Dispersed Attention:** In the $\mathcal{I}$-only setting (Fig. 5b), the model's attention is tightly locked on the hands and the pan—the center of the action. However, when the textual description lists peripheral objects (e.g., "cleaning supplies," "dish rack"), the multimodal attention (Fig. 5c) scatters towards these background elements.
- **Loss of Grounding:** The Attention Difference map (Fig. 5d) explicitly shows this trade-off: the **Red regions** indicate a significant loss of attention on the hands. By attending to the background clutter mentioned in the text, the model effectively "loses sight" of its own active hands. In this state of lowered visual confidence, the model fails to anchor its decision in the current physical reality, causing it to hallucinate an unrelated activity (checking the phone).

Remark on Visual Domain Gap (Human vs. Robot). Readers may notice that the visual input depicts human hands (sourced from Ego4D)

Case Study: Kitchen Cleanup Sequence (Item 003)

Task Context: You are washing a frying pan at the sink with the water running. **Goal:** You are expecting a guest in 20 minutes and need the kitchen to look tidy without delay.

Modality Inputs:

- **Visual ($\mathcal{I}$):** Egocentric view showing agent actively washing a black pan under a running faucet. [Signal: Action in progress].
- **Textual ($\mathcal{T}$):** "Scene Description: You are standing in front of the kitchen sink washing a black pan... Objects: [sink, frying_pan (being_washed), water_faucet (running), rubber_glove (unused), drying_rack, cleaning_supplies]... Goal: Tidy up for a guest arriving in 20 mins." [Signal: Detailed State + Time Constraint].

Model Predictions (Qwen3-VL-8B):

- **Image-Only ($\mathcal{I}$): Option A (Correct)** - "Quickly finish rinsing the pan and rack it, turn off the tap; then wipe the sink edge..."
- **Text-Only ($\mathcal{T}$): Option A (Correct)** - "Quickly finish rinsing the pan and rack it... turn off the tap..."
- **Multimodal ($\mathcal{M}$): Option C (Incorrect)** - "Immediately stop washing, pick up the phone nearby to check for unread messages... if important, postpone dishwashing." *(Note: No phone is present in the image or the structured object inventory (verified during QA); this is a hallucinated distractor.)*

Figure 4: An example of **Modality Interference**. The "Phone" option is a distractor not present in either modality. While the model correctly ignores it in unimodal settings, the fusion of modalities triggers a hallucination, leading to a reasoning failure.

rather than robotic end-effectors. We explicitly retain this human-centric observation for two reasons: (1) **Scale and Diversity:** Ego4D provides a diversity of real-world interaction scenarios currently unmatched by robotic datasets. (2) **Cognitive Isomorphism:** Our benchmark targets the cognitive planning layer (e.g., "what to do next") rather than the low-level motor control layer (e.g., "how to move joints"). We posit that the high-level decision logic—such as recognizing that "a soapy pan needs rinsing"—is largely independent of the manipulator morphology at the planning level. Thus, human demonstrations serve as a rigorous proxy for evaluating the "brain" of an embodied agent, independent of its physical body.

Conclusion. This case demonstrates that multimodal consistency is not merely about feature aggregation. When fusion strategies are brittle, the introduction of a second modality can paradoxically dilute the model's attention to the primary task. This dispersion weakens the grounding necessary to reject distractors, leading to "consistent failure" or hallucinations that would not occur in unimodal settings.

B Bidirectional Verification and Grounding Analysis

While the forward evaluation assesses an agent's ability to select actions, it provides limited insight into whether the decision is derived from the specific observation. To probe grounding versus reliance on parametric priors, we employ a **Reverse Verification Protocol** as a controlled scene-retrieval task.

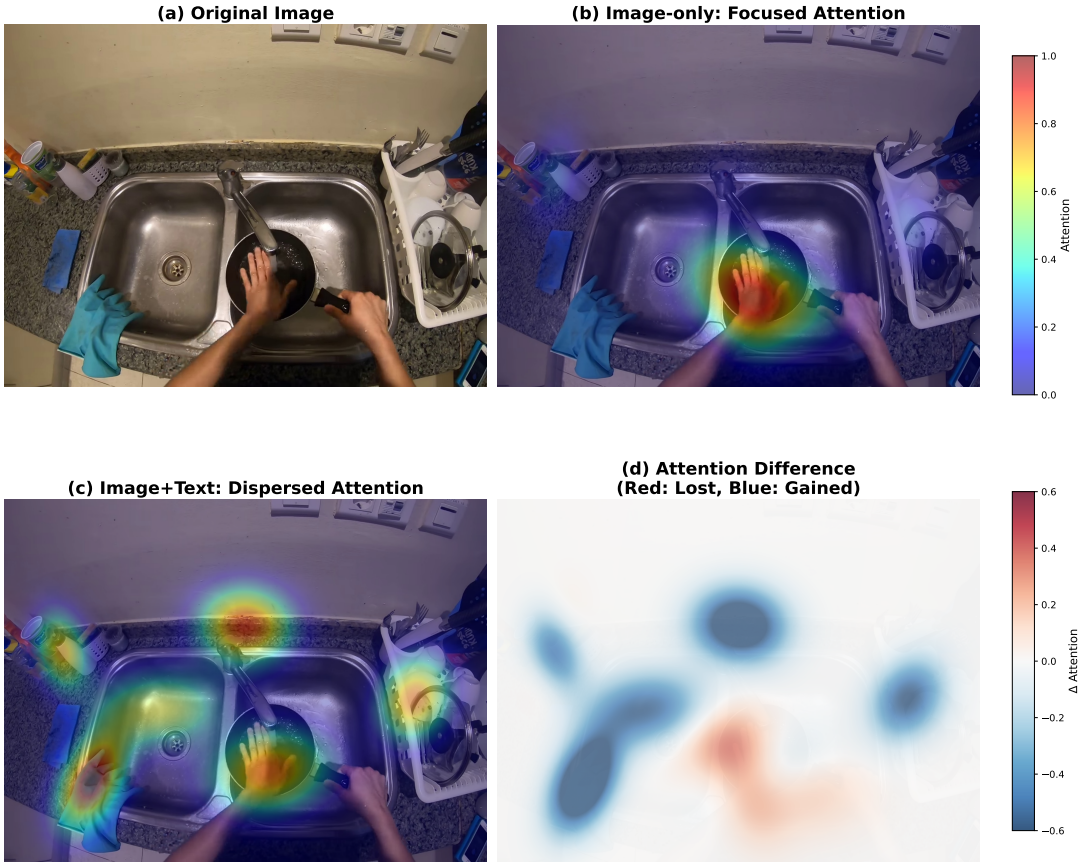

Figure 5: **Visualization of Modality Interference (Attention Dispersion)**.We compare gradient-based visual relevance maps (Grad-CAM-style) for the Image-Only (b) and Image+Text (c) settings using Qwen3-VL-8B. While the unimodal model focuses on the decision-critical action (washing the pan), the introduction of text causes attention drift towards irrelevant background objects mentioned in the description (cleaning supplies, dish rack). (d) The difference map highlights the suppression of task-relevant features (Red) and the amplification of background clutter (Blue), consistent with degraded visual grounding under naive fusion.Relevance is computed w.r.t. the predicted option logit on the vision backbone.

B.1 Experimental Methodology

For each instance that the model answers correctly in the forward setting (we use the multimodal $\mathcal{M}$ setting unless otherwise noted), we construct a reverse query by providing the and its gold answer and asking the model to retrieve the matching scene from **four candidates** (1 true scene + 3 distractors). Distractors are sampled from the same task family (and the same data source split when applicable) to avoid superficial cues. We query each candidate independently using a fixed template and obtain a 1–10 match score; We adopt this independent scoring protocol for practical and fairness reasons: most evaluated VLM APIs only support a single image per call, and forcing a four-way forced-choice by concatenating candidates into one image would compress each candidate and consume the limited visual-token budget, disproportionately hurting smaller models. Independent scoring keeps the input interface consistent across model sizes while avoiding layout/resolution artifacts introduced by concatenation. the candidate with the highest score is selected as the prediction (ties are broken by the earliest candidate index). Candidate order is randomized.

We report **Reverse Accuracy** (Acc_{Rev}) as the scene-retrieval accuracy. We further report the **Bidirectional Consistency Score (BCS)**, defined as the probability that the model is correct in both forward decision-making and reverse retrieval: $BCS = P(\text{Fwd} \cap \text{Rev})$. Finally, we analyze the **Consistency Gap** ($\Delta_C = Acc_{Fwd} - \text{BCS}$), where a larger gap indicates a divergence between decision correctness and grounded context identification.

Anti-shortcut controls. (i) We use the same prompt template across all candidates; (ii) candidate order is randomized; (iii) distractors are sam-

pled from the same task family to reduce trivial mismatches.

B.2 Quantitative Analysis

Table 3 presents results on the forward-correct subset used for reverse verification (i.e., instances where the model answers correctly in the forward $\mathcal{M}$ setting). While most models achieve high Forward Accuracy on this subset by construction, BCS provides a more nuanced view of whether correct decisions are grounded in the corresponding scene.

Observations on Cross-Modal Alignment. We observe varying degrees of consistency across architectures. Mid-sized open-weight models, such as **Llama-3.2-11B** and **Qwen2.5-VL-3B**, show strong Forward Accuracy, comparable to larger proprietary models. However, their BCS scores are lower (ranging between 53-54%), resulting in a Consistency Gap of approximately 41.8%.

- **Interpretation:** This performance delta suggests that for a portion of instances, these models may arrive at the correct answer through language priors or option selection strategies, rather than fully resolving the visual-textual correspondence. This highlights a potential area for improvement in smaller architectures: ensuring that decision-making is robustly grounded in the input observation.

Scaling and Grounding. A comparison within the Qwen2.5-VL family illustrates the impact of model scale. The 7B model shows a notably smaller gap (20.2%) compared to the 3B model (41.8%), with BCS improving from 53.3% to 78.7%. This trend implies that increasing model capacity may contribute to more coherent bidirectional mappings between visual states and semantic actions.

C Robustness Under Modality Conflict

In embodied deployment, agents may encounter discrepancies between sensor inputs (e.g., a visual feed vs. a textual map update). To understand how models weigh conflicting information, we analyze their behavior on a subset of instances where visual ($\mathcal{I}$) and textual ($\mathcal{T}$) inputs describe semantically distinct scenes.

We construct this modality-conflict subset by pairing an image from one instance with a text description from a different instance (same task family), such that the two modalities correspond to different underlying scenes. We define "ground-truth aligns with Image" if the correct option is determined by the image-source instance, and "aligns with Text" if it is determined by the text-source instance. Models are evaluated in the multimodal $\mathcal{M}$ setting on these conflicted pairs, and P_{Img}/P_{Txt} report accuracy on the two labeled subsets, respectively.

C.1 Model-Level Preference

Table 4 displays the preference tendencies across different models. Across the models shown in Table 4, we observe relatively balanced arbitration between modalities.

The results indicate that preferences are generally model-specific but moderate. For example, **Llama-3.2** shows a slight visual tendency (+2.0%), while **Qwen** and **Gemma** models show a minor textual tendency ($\sim -1.0\%$). These small margins suggest that models generally attempt to process both signals rather than defaulting entirely to one modality.

C.2 Modality Preference in Map-Based Tasks

A more distinct pattern appears when analyzing preferences by **Embodied Task Category**. As shown in Table 5, the type of task appears to influence which modality the model relies on.

Visual Preference in Egocentric Tasks. For tasks involving immediate spatial awareness—such as Spatial Understanding and Mission Prioritization,models consistently show a slight visual preference. This aligns with the expectation that navigation tasks primarily rely on visual observation.

Textual Shift in Map Decisions. In the **Map & Graph Decision** category, we observe a noticeable shift: models exhibit a higher accuracy when following the textual description ($\Delta_{Pref} = -17.8\%$).

- **Analysis:** Visual inputs in this category often consist of node-link diagrams or floorplans. The preference for text suggests that current architectures may find structured textual representations (e.g., adjacency lists) easier to interpret for topological reasoning compared to extracting connectivity from visual diagrams.

- **Implication:** This observation supports the multi-modal design of **EmModBench**. It suggests that for abstract planning tasks, text

Table 3: **Comprehensive Bidirectional Consistency Analysis.** Models are ranked by their Bidirectional Consistency Score (BCS). The **Consistency Gap** (Δ_C) reveals the divergence between decision-making capability and context understanding. Note the dramatic gap in mid-sized models compared to the robustness of larger architectures.

Model	Forward Acc	Reverse Acc	BCS	Consistency Gap (Δ_C)
Tier 1: Robustly Grounded Agents				
Qwen2.5-VL-7B	98.9%	79.6%	**78.7%**	**20.2%**
Gemini-3-12B	97.9%	76.3%	**74.7%**	**23.2%**
Tier 2: High Performance, Low Grounding				
Qwen3-VL-8B	97.1%	68.4%	**66.8%**	30.3%
Qwen3-VL-4B	95.9%	63.2%	**61.5%**	34.4%
Llama-3.2-11B	**95.9%**	55.9%	**54.1%**	**41.8%**
Qwen2.5-VL-3B	95.1%	55.4%	**53.3%**	**41.8%**
Gemma-3-4B	91.4%	56.7%	**52.1%**	39.3%
LLaVA-v1.5-13B	88.4%	52.4%	**46.1%**	42.3%

Table 4: **Model-Level Modality Preference.** Accuracy breakdown when ground-truth aligns with Image (P_{Img}) vs. Text (P_{Txt}) under conflicting inputs. $\Delta_{Pref} > 0$ indicates a Visual tendency; $\Delta_{Pref} < 0$ indicates a Textual tendency. Most models show a balanced preference with small gaps.

Model	Image Pref. (P_{Img})	Text Pref. (P_{Txt})	Gap (Δ_{Pref})	Tendency
Llama-3.2-11B	82.1%	80.1%	+2.0%	*Image*
LLaVA-v1.5-13B	78.2%	77.2%	+0.9%	*Image*
Qwen2.5-VL-7B	90.1%	90.9%	-0.7%	*Text*
Gemma-3-12B	89.7%	90.7%	-1.0%	*Text*
Qwen3-VL-8B	85.9%	88.4%	-2.6%	*Text*

is a critical carrier of information, and robust agents may need to dynamically weight modalities based on the task type.

D Data Generation and Rigorous Quality Assurance

Connection to the main paper. In the main paper, we briefly describe the three data sources of **EmModBench**—Matterport3D simulator environments, Ego4D egocentric data, and canonical graph/topology resources—and refer readers here for the full data generation and quality assurance details. This appendix focuses on the components that are most important for reproducibility and trustworthiness, especially the construction and verification of the textual modality for Ego4D, where paired structured text is not natively available.

Data sources (recap). EmModBench is built from three major sources: Matterport3D simulator environments, Ego4D egocentric data, and canonical graph/topology resources.

- **Matterport3D subset.** For indoor embodied scenes, we use the Matterport3D dataset as the 3D environment source and render egocentric observations through the Habitat simulator. Note that the standard Matterport3D simulator interface mainly provides visual observations (RGB/depth) and navigation states, but does not directly expose object lists or object-relation graphs. To obtain a structured text modality aligned with each view, we leverage **Matterport3D object instance annotations** and **room/category metadata**, and perform backend processing to map annotations into the agent's current viewpoint. Specifically, given the camera pose and field-of-view, we identify the decision-relevant object instances (e.g., visible instances and/or instances within a distance threshold), assign each instance its category and room membership, and derive coarse spatial/functional relations needed by our questions. The resulting text modality consists of an object list and coarse relations (e.g., room containment and basic spatial relations).

- **Ego4D subset.** Ego4D frames typically lack paired structured text. The Ego4D subset contains **1,505** instances (**34.15%** of **EmModBench**). To construct the text modality while

Table 5: **Task-Level Modality Preference.** Standard navigation tasks (Classes 1–4) generally show a slight visual preference or balance. In contrast, the **Map & Graph Decision** category shows a marked performance advantage for the textual representation , suggesting that models may find structured textual descriptions more reliable for topological reasoning than visual diagrams.

Task Category	Image Pref.	Text Pref.	Gap (Δ_{Pref})
Prior Knowledge (Class 1)	68.6%	67.2%	+1.4%
Spatial Understanding (Class 2)	94.8%	94.0%	+0.8%
Mission Prioritization (Class 4)	88.1%	87.3%	+0.8%
Next-Action Planning (Class 3)	86.6%	86.3%	+0.3%
Map & Graph Decision (Class 5)	35.4%	53.2%	**-17.8%**

minimizing hallucinations, we adopt a **human-verified generation pipeline**. Specifically, we first use a high-performing vision–language model to synthesize (i) salient object lists and relations and (ii) global scene descriptions. To reduce model-specific biases and detect potential hallucinations, we then validate the generated texts with a different-architecture language model, and additionally perform human audits on a randomly sampled **20%** of the instances (**301** samples). This multi-stage verification is designed to ensure that the textual descriptions preserve decision-relevant cues with high fidelity. We detail the complete protocol, audit criteria, and observed error types in the remainder of this appendix.

- **Structured path reasoning (Map/Graph) subset.** For structured planning tasks, we construct strictly aligned cross-modal representations: the *visual* input is a node–edge diagram (e.g., floorplan topology or abstract graphs), and the *text* input is an adjacency list or adjacency matrix encoding. This bidirectional mapping preserves information across modalities and enables rigorous evaluation of cross-modal reasoning and consistency.

Scope of this appendix. The remainder of this appendix describes our data generation pipeline and quality assurance procedures, with an emphasis on the Ego4D textual modality. These safeguards are designed to keep the generated text both visually grounded and decision-relevant, while remaining scalable for benchmark construction.

Overview. To ensure the scalability of **EmMod-Bench** without compromising semantic fidelity, we developed a data generation pipeline that produces textual modalities (including structured object-centric descriptions and global scene summaries) and supports the construction of challenging QA pairs. Recognizing the risk of hallucinations in Large Language Models (LLMs), we go beyond one-shot generation by combining automated cross-verification with human auditing. Concretely, we implement a **Dual-Architecture Cross-Verification** protocol followed by a strict **Human-in-the-Loop (HITL)** audit. This multi-stage procedure is intended to ensure that the textual modality is well-formed, internally consistent, and aligned with decision-critical visual cues.

D.1 Protocol 1: Dual-Architecture Cross-Verification

We hypothesize that models from distinct architectures and training corpora are less likely to produce the same unsupported details. Based on this intuition, we adopt a Generator–Verifier workflow:

1. The Generator (Claude Opus 4.5). We utilize **Claude Opus 4.5** as the primary generator for constructing structured object-centric descriptions and macro (global) scene summaries. The generator is prompted to follow a fixed schema, which supports automatic parsing and downstream evaluation.

2. The Verifier (Grok 4.1). To audit the generated content, we employ **Grok 4.1** as an independent verifier. Grok 4.1 is assigned a critic role and checks the output from three angles:

- **Syntax Check:** Verifying strict adherence to the JSON schema (e.g., ensuring no trailing commas and correct nesting).
- **Logic Check:** Identifying semantic contradictions (e.g., an object cannot be both "inside" and "under" the same container simultaneously).
- **Consistency Check:** Ensuring the macro description aligns with the generated object list and

relations.

Regeneration rule. Any instance flagged by the Verifier is automatically discarded and regenerated. This adversarial setup filters the majority of structural and logical errors before human inspection.We allow up to a small fixed number of retries regeneration attempts per instance (we use $K = 3$ in our implementation). An instance is flagged if it violates the schema or contains logical/consistency errors under the verifier rubric.

D.2 Protocol 2: Human-in-the-Loop (HITL) Audit

Following automated cross-verification, we conduct a human audit to estimate residual issues and validate "gold-standard" quality for a representative subset.

Sampling Strategy. We randomly sample **20%** of the generated Ego4D subset for manual review (**301** out of **1,505** instances). This sample is used to estimate residual issues and to identify recurring failure patterns that can be addressed by prompt/schema refinement.

Audit Results. Expert annotators evaluate the sampled instances against three criteria: **Accuracy (Grounding)**, **Format**, and **Relevance**. Overall, the pass rate is **98%** on the audited sample. Each sampled instance is reviewed by two annotators with access to the original Ego4D frame and the generated text. Disagreements are resolved by a third senior annotator.

The rejected cases primarily involve:

1. *Redundancy:* repetitive phrasing in the macro scene description;

2. *Granularity Issues:* descriptions that are too vague for the decision context.

Handling of failed cases. When an instance fails the audit, we either regenerate it under stricter constraints or exclude it from the final benchmark, depending on the error type. Since the dominant issues are stylistic, we expect limited impact on the benchmark's measured decision consistency; nevertheless, we include this audit to provide transparency and to guide ongoing refinement of the pipeline.

D.3 Robustness to Residual Noise in Synthesized Text

To verify that our main conclusions are not driven by residual noise in the synthesized Ego4D text modality, we re-compute key metrics on the human-audited pass-only subset. Across models, the qualitative trends (e.g., modality interference patterns and the prevalence of consistent-but-wrong cases) remain stable. We report the full results in Table 6.

D.4 Prompt Templates

To facilitate reproducibility, we provide representative prompt templates used in our pipeline below. The first template illustrates the strict instructions given to the Generator model for converting visual inputs into structured textual modality (Scene Graphs + Descriptions), primarily used in *Task Planning* scenarios. Additionally, the second template demonstrates the protocol for synthesizing spatial reasoning questions from topological graph data, specifically designed for the *Map & Graph Reasoning* category.

This structured prompt enforces a separation between the unstructured "narrative" (for implicit reasoning) and the structured "graph" (for symbolic reasoning), ensuring that the Textual Modality in *EmModBench* is both descriptive and computable.

E Online Predictive Validity (Pilot Bridge Study)

To examine whether the cognitive inconsistencies diagnosed by **EmModBench** manifest as behavior-level instability in embodied deployment, we conduct a small-scale *closed-loop bridge study* as a sanity check. Specifically, we test whether offline consistency/grounding diagnostics (SAR, BCS) exhibit a consistent trend with rollout stability in a dynamic navigation task. We emphasize that this pilot is intended to provide corroborative evidence rather than establish causality.

E.1 Experimental Setup

Testbed and Architecture. We use **UniGoal** (Yin et al., 2025) as the execution framework. UniGoal maintains an online structured scene graph during navigation, matching the modular *visual perception vs. structured context* setting targeted by EmModBench. We keep UniGoal's overall pipeline fixed and only swap its underlying LLM/VLM decision modules with two

Table 6: **Robustness check on synthesized Ego4D text.** We compare performance on the full Ego4D subset produced by our pipeline vs. a human-audited subset. The small gaps (<1%) suggest that residual synthesis noise has limited impact on aggregate metrics; see Appendix D.2 for audit statistics.

Eval Set	Source	Size	$Acc_{\mathcal{T}}$ (Avg)	Consistency (SAR)
Ego4D (Full Pipeline)	All Generated	1,505	74.8%	78.5%
Audited Sample (Raw)	Random 20%	301	75.0%	78.8%
Audited (Pass-only)	**Human Verified**	**295**	**75.1%**	**79.0%**

Table 7: **Pairwise Inconsistency Rates.** Lower is better. Tri-modal inconsistency corresponds to $100 - \text{SAR}$. Note clearly that Qwen3-VL (2B)† shows low inconsistency due to degenerate failure (repeating one option).

Model	Vis-Text	Vis-Mix	Text-Mix	Tri-modal
	$\text{Inc}_{\mathcal{I},\mathcal{T}}$	$\text{Inc}_{\mathcal{I},\mathcal{M}}$	$\text{Inc}_{\mathcal{T},\mathcal{M}}$	$\text{Inc}_{\mathcal{I},\mathcal{T},\mathcal{M}}$
Gemma 3 (4B)	28.9	26.7	13.4	33.0
Gemma 3 (12B)	19.8	15.0	13.0	22.3
Llama 3.2-Vis (11B)	18.7	18.1	17.8	25.6
LLaVA v1.5 (7B)	20.6	13.6	18.2	25.4
LLaVA v1.5 (13B)	30.2	19.9	27.5	35.7
Qwen2.5-VL (3B)	15.9	13.0	11.7	19.4
Qwen2.5-VL (7B)	**12.9**	**10.4**	**8.4**	15.3
Qwen3-VL (2B)†	*2.1*	*2.0*	*2.4*	*2.3*
Qwen3-VL (4B)	12.1	10.7	4.5	12.8
Qwen3-VL (8B)	14.3	10.7	7.8	15.3

representative models that exhibit a clear gap in our main evaluation:

- **Qwen3-VL (8B)**: a *high-consistency* model (high SAR with positive fusion gain Δ_{VL}).
- **LLaVA v1.5 (7B)**: a *baseline/lower-consistency* model (lower overall accuracy and more susceptible to modality interference).

We keep perception, mapping, action space, and prompt templates identical across runs; only the underlying VLM used for decision selection is swapped.

Task and Dataset. We evaluate on HM3D under the standard **ObjectNav** setting. We randomly sample $N = 50$ episodes from the validation split. We report standard navigation metrics: Success Rate (SR) and Success weighted by Path Length (SPL).

Addressing Information Equivalence (T_{local} vs. T_{global}). A key challenge when bridging offline QA to online navigation is that structured context in modular agents can be memory-augmented (e.g., accumulated mapping), whereas visual observations are instantaneous. To better control for this confound, we distinguish two structured-text variants:

- T_{local} **(Local Graph):** an instantaneous structured proxy constructed by filtering the scene graph to the agent's current view/nearby region at the current step. This is designed to be information-matched to the current image observation (I) as much as possible. We explicitly exclude privileged signals from T_{local} (e.g., global maps, shortest-path hints, goal coordinates, or future observations).
- T_{global} **(Memory-Augmented):** the full accumulated scene graph used by UniGoal for mapping. We use this only in the memory ablation setting.

We report episode-averaged metrics with 95% bootstrap confidence intervals over episodes.

E.2 Protocol: Shadow Mode & Random Fragmentation

To simulate the input fragmentation scenarios in EmModBench within a continuous rollout, we adopt a **Shadow Mode** protocol:

1. **Execution Policy:** At each step, the agent follows a Random Fragmentation policy where the input format is sampled from $\{I\text{-only}, T_{local}\text{-only}, M(I + T_{local})\}$ with probabilities $\{0.4, 0.4, 0.2\}$. This stresses robustness under input-format switching.
2. **Decision-Point Probing:** At decision points (steps with > 1 navigable waypoints), we additionally query all three inputs (I, T_{local}, M) to obtain their *counterfactual* intended actions, while executing only the action selected by the sampled input format.

We measure behavioral stability using two intuitive metrics:

- **Action Match (Cross-Modal Agreement):** the proportion of decision points where the proposed actions are identical across all three modalities ($Action_I = Action_{T_{local}} = Action_M$).

Table 8: **Bridge Study Results.** Comparing offline diagnostics with closed-loop navigation behavior. **Action Match**: rate of identical decisions at critical junctions; **Jitter**: frequency of vacillating moves. Note that **Success Rate** is lower than clean-input benchmarks due to fragmented inputs.

Model	Offline Diagnostics		Online Metrics (HM3D Pilot)			
	Consistency (SAR, % ↑)	Grounding (BCS, % ↑)	**Action Match** (↑)	**Success Rate** (% ↑)	**Eff.** (SPL, % ↑)	**Behav. Jitter** (↓)
Qwen3-VL (8B) *(High Consistency)*	**84.6**	**66.8**	**0.58**	**48.0**	**20.5**	**0.14**
LLaVA v1.5 (7B) *(Lower Consistency)*	74.6	39.5	0.36	34.0	14.2	0.24

- **Jitter (Oscillation Rate):** the frequency of vacillating behaviors (e.g., Left–Right–Left sequences), normalized by trajectory length.

E.3 Results and Trend Analysis

Table 8 summarizes the pilot results. As a reference anchor, the UniGoal paper reports an SR of 54.5% on HM3D under standard (non-fragmented) inputs.

Observations. We observe a qualitative trend consistent with the offline diagnostics:

1. **Higher offline grounding aligns with more stable decisions under input-format switching.** Qwen3-VL achieves substantially higher offline grounding (BCS: 66.8% vs. 39.5%) and correspondingly higher **Action Match** (0.58 vs. 0.36), suggesting it is less sensitive to modality switching at critical junctions.

2. **Lower consistency associates with increased behavioral jitter.** LLaVA exhibits higher **Jitter** (0.24 vs. 0.14), indicating more frequent vacillation under fragmentation, which is accompanied by lower SR/SPL in this pilot.

Conclusion. Overall, this pilot bridge study provides corroborative evidence that **EmModBench**'s offline consistency/grounding scores provide preliminary evidence that offline consistency/grounding scores are associated with closed-loop stability in modular embodied systems, particularly under fragmented observations. A larger-scale study with more models and episodes is left for future work.

F Human Baseline and Ethical Considerations

To validate the premise of **EmModBench**—that the inconsistency observed in VLMs stems from model limitations rather than dataset ambiguity—we conducted a human performance study. This serves to establish a "Human Upper Bound" and verify that decision-critical information is preserved across modalities.

F.1 Human Baseline Results

We recruited three graduate students with expertise in embodied AI to evaluate a stratified sample (10% of instances). As shown in Table 9, human performance is nearly saturated, contrasting sharply with model averages.

Protocol. We sample 10% of items with stratification across the five task categories. Each annotator answers each item under $\mathcal{I}$, $\mathcal{T}$, and $\mathcal{M}$ separately with randomized order and no feedback. Human SAR is computed as the fraction of items where an annotator's three answers agree ($\hat{y}^{(\mathcal{I})} = \hat{y}^{(\mathcal{T})} = \hat{y}^{(\mathcal{M})}$), averaged over annotators.

Analysis of Human Data.

- **Information Sufficiency:** Humans achieved $> 98.6\%$ accuracy across all three modality settings. The slight variations (e.g., 99.8% in Image vs. 98.6% in Multimodal) are within the margin of minor human error, confirming that each modality independently contains sufficient information to solve the task.

- **No Modality Interference:** Unlike models, which often suffer from "Modality Interference" (where fusion drops performance), humans maintained consistent correctness. The high SAR (99.2%) indicates that humans almost never flip their decision based on input format, validating the semantic alignment of our dataset.

Table 9: **Human vs. Model Performance.** Human annotators achieve near-perfect consistency (99.2%), confirming that items are unambiguous and each modality is decision-sufficient for answering correctly. In contrast, current models exhibit a significant performance and consistency gap.

Metric	Human Baseline	Avg. Model	Performance Gap
Image-Only Accuracy ($\mathcal{I}$)	98.9%	75.3%	+24.5%
Text-Only Accuracy ($\mathcal{T}$)	98.6%	74.8%	+24.1%
Multimodal Accuracy ($\mathcal{M}$)	99.8%	74.1%	+24.5%
Consistency (SAR)	**99.2%**	**78.5%**	**+20.7%**

Prompt Box : Prompt for Scene Graph and Description Generation

[Task] Input is a first-person or top-down view image from an embodied agent. The scene usually depicts an intermediate state of an environment (e.g., having breakfast, organizing a room, tidying a desk).
Your goal is to convert visual information into a JSON format without losing information, avoiding words like "image," "camera," or "photo." The output must include: 1. 2–3 sentences of macro scene description (first-person "You"). 2. Task-relevant scene graph: Scene Info + Key Objects + Key Relations.
[Output Format] You must output ONLY a valid JSON with the following structure:

```
{
  "notation_type": "scene_graph",
  "notation": {
    "scene_description": "< 2-3 sentences, 1st person macro description>",
    "scene": {
      "location": "<English scene type, e.g., kitchen>",
      "viewpoint": "<e.g., first_person>",
      "ego_agent": "implicit",
      "task_context": "<English short description of current sub-task>"
    },
    "objects": [
      { "id": "counter", "category": "kitchen_counter" }
    ],
    "relations": [
      { "subject": "plate", "predicate": "on_top_of", "object": "counter" }
    ]
  }
}
```

[Specific Requirements] 1. scene_description: - first-person perspective. - Describe as if the agent is narrating its own environment (e.g., "You are...", "In front of you..."). - Do NOT use "image/camera/screen". - Mention: Room/Area, Current Sub-task Status, Key items for long-horizon planning (dirty dishes, trash, items to be folded).
2. Focus on Planning-Relevant Relations: - Status: "being eaten", "finished", "untouched". - Attributes: "dirty", "needs disposal", "needs refrigeration". - Spatial: Important relations between zones (Table vs. Sink vs. Bin).
3. JSON Rules: - No comments // or ellipses ... - Must be valid JSON (double quotes, no trailing commas).
[Example Output (Style Reference)]

```
{
  "notation_type": "scene_graph",
  "notation": {
   "scene_description": "You are sitting at the kitchen island, with a plate of fried eggs in front..."
    "scene": {
      "location": "kitchen",
      "viewpoint": "first_person_top_down",
      "ego_agent": "implicit",
      "task_context": "ego agent is having breakfast..."
    },
    "objects": [
      { "id": "counter", "category": "kitchen_counter" },
      { "id": "fried_egg", "category": "cooked_egg", "attr": ["ready_to_eat"] },
      { "id": "dirty_dishes", "category": "dishes" }
    ],
    "relations": [
      { "subject": "fried_egg", "predicate": "on_top_of", "object": "plate" },
      { "subject": "dirty_dishes", "predicate": "inside", "object": "sink" }
    ]
  }
}
```

Prompt Box : Prompt for Floorplan Adjacency Parsing and QA Generation

[Role] You are an annotator building a benchmark dataset for multimodal models. You are provided with a top-down view of an indoor room (e.g., floorplan or 3D top-down render).
[Task] 1. Abstract a labeled undirected graph from the image to represent the connectivity between rooms (i.e., which rooms are directly connected by a door or opening). 2. Based on this graph structure, design a single-choice question to test spatial/path reasoning capabilities. 3. Output a strictly valid JSON object without any additional text.
[Requirements]
1. Graph Notation (notation) - **notation_type:** Fixed as "adjacency_list". - **notation.nodes:** A dictionary where keys are short room IDs (e.g., "L", "K", "B1") and values are English room types (e.g., "living_room", "kitchen", "bedroom", "bathroom", "corridor", "balcony"). - **notation.edges:** An adjacency list representing the undirected graph. Keys are room IDs; values are lists of adjacent room IDs. Connect two rooms ONLY if there is a visible door or opening. - **Scope:** Include only clearly visible and relevant rooms. Do not hallucinate rooms.
2. Question Design (question & options) - **question_type:** Fixed as "multiple_choice". - **question:** An English question focusing on spatial reasoning based on the graph structure. Examples: - "What is the shortest path from B1 to Ba2?" - "Is it possible to reach the kitchen from the bedroom without passing through the living room?" - "Which room requires the fewest steps to reach the balcony?" - **Constraints:** There must be exactly one correct answer. The other three options must be plausible distractors but strictly incorrect. - **options:** An array of 4 option strings (indices 0-3). Do not include prefixes like "0." or "A.". - **correct_answer:** The index string of the correct option ("0", "1", "2", or "3"). - **answer_idx:** The integer index (0-3).
3. Metadata Fields - **id:** Format "room_path_XXXX" (use "room_path_0001" as a placeholder). - **task:** Fixed as "path_reasoning". - **domain:** Fixed as "indoor_floorplan". - **split:** Fixed as "train". - **image_path:** Relative path, e.g., "floorplan/images/room_path_0001.png".
[Output Format] - Output ONLY a valid JSON object. - Use double quotes. No comments (//) or trailing commas.
[Example Output]

```
{
  "id": "room_path_0001",
  "task": "path_reasoning",
  "domain": "indoor_floorplan",
  "split": "train",
  "question_type": "multiple_choice",
 "question": "Starting from the Living Room (L), what is the shortest path to the Twin Bedroom (B1)?",
  "notation_type": "adjacency_list",
  "notation": {
    "nodes": {
      "L": "living_room",
      "K": "kitchen_dining",
      "B1": "twin_bedroom",
      "B2": "double_bedroom",
      "Ba1": "bathroom_1",
      "C": "corridor"
    },
    "edges": {
      "L": ["K", "C"],
      "K": ["L", "B1", "B2"],
      "B1": ["K"],
      "B2": ["K", "Ba1"],
      "Ba1": ["B2"],
      "C": ["L"]
    }
  },
  "options": [
    "L -> K -> B1",
    "L -> C -> K -> B1",
    "L -> K -> B2 -> Ba1 -> B1",
    "L -> C -> B1"
  ],
  "correct_answer": "L -> K -> B1",
  "answer_idx": 0,
  "image_path": "floorplan/images/room_path_0001.png"
}
```