ENACT: Evaluating Embodied Cognition with World Modeling of Egocentric Interaction

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce ENACT, a scalable benchmark for studying *Embodied Cognition* via world modeling through egocentric interaction, probing how spatial perception, physical interaction, and language cohere in modern Vision–Language Models (VLMs). Grounded in a POMDP view of decision making, ENACT comprises two complementary permutation tasks: **forward world modeling** (reorder future observations to match a given action sequence) and **inverse world modeling** (reorder actions to explain a given observation sequence). Data are generated by replaying diverse household activities in the reproducible simulator (Behavior) that aligns symbolic scene graphs with egocentric RGB, yielding 8,972 QA items. Predictions are validated by an online verifier that accepts any sequence consistent with task constraints, and we report Task Accuracy (exact ordering) and Pairwise Accuracy (adjacent consistency). Across evaluated VLMs, performance degrades with longer interaction horizons, and inverse is consistently easier than forward. Targeted probes of GPT-5 mini and InternVL-3.5 show limited sensitivity to image realism and robot appearance, and GPT-5 mini exhibits marked sensitivity to camera-distribution shifts (elevated viewpoints, extreme apertures, fisheye). Both models display a handedness asymmetry with fewer right-hand errors. Overall, ENACT offers a scalable proxy for studying embodied cognition and a tool to inform models that better bind perception to action over long horizons.

1 Introduction

Intelligence is not simply computed, but enacted through interaction. The theory of *Embodied Cognition* holds that flexible thought grows from an agent's continuous sensorimotor dialogue with a physical, social, and linguistic world [1]. This perspective suggests that the mind is woven from the interplay between the body and its environment. Spatial perception shapes the boundaries of what is knowable [2], physical interaction uncovers what is possible, and language provides the tools to compress, name, and plan from this grounded experience. In this synthesis, space gives structure, contact gives evidence, and words bind these elements into shareable models. Following this thread, our work investigates the egocentric interaction experience to understand how perception and action co-author a world model, one step at a time.

The quest for intelligence has seen remarkable advancements with the scaling of large foundation models [3, 4]. Yet these models are fundamentally trained in a disembodied manner, learning from a volume of non-interactive data that vastly exceeds any single human's lived experience. This raises a natural and intellectually curious question: **Does embodied cognition emerge from such training?** Valuable prior work has begun to probe the parts: spatial perception in static scenes [2], physical interactions in contrived settings [5, 6] (e.g., a rolling ball hits another ball in a clean environment), and purely linguistic reasoning [7]. While these lines of work have been insightful, a gap still remains:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

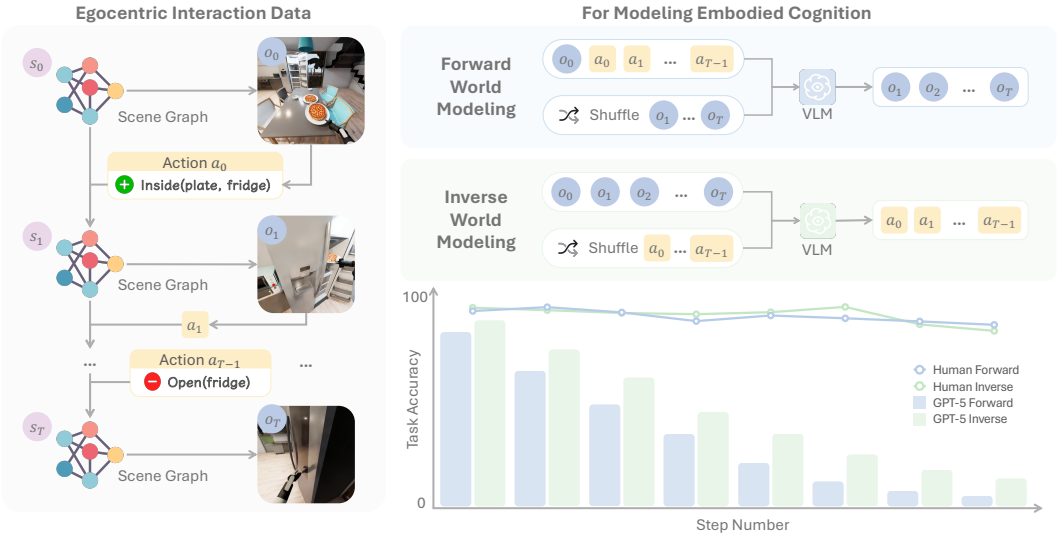

Figure 1: Grounded in a POMDP framework, **ENACT** probes embodied cognition in a **simple and scalable** way via world modeling through egocentric interaction (left). It poses two tasks (right top): **forward world modeling** (ordering observations given actions) and **inverse world modeling** (ordering actions given observations). Evaluation (right bottom) shows that GPT-5 performance drops as step length scales, solves better on inverse task, and lags behind humans.

the interplay among physical interaction, spatial perception, and linguistic understanding is seldom examined within a unified, egocentric experience where perception and action must cohere over time.

To bridge this gap, we introduce ENACT, a benchmark that studies the synergy of embodied capabilities through a simple, powerful objective: **world modeling through egocentric interaction**. Grounded in a Partially Observable Markov Decision Process [POMDP, 8], ENACT poses two complementary tasks (shown in Figure 1). In **forward world modeling**, the model receives an initial observation, a sequence of abstract actions, and a shuffled set of future observations, and must reorder the observations to match the actions. In **inverse world modeling**, it receives an ordered sequence of observations and must reorder a shuffled set of actions that explains the progression. Though conceptually simple, they demand a rich synthesis: fine-grained spatial understanding, long-horizon memory, and a tight binding between perception and interaction. The benchmark leverages a reproducible simulator to capture a robot's egocentric interactions from diverse household tasks, which also allows controlled experiments for probing existing models' data biases.

Our curation pipeline is **simple** and **scalable**. We replay long-horizon household tasks in a simulator that records aligned symbolic scene graphs (states) and egocentric RGB observations. We segment each replay at abstract state changes, prune near-duplicates via predicate-level changes, and assemble validated key-frame trajectories with visible transitions. From these, we use question templates to build QAs for forward and inverse world modeling. We report two metrics: Task Accuracy (exact ordering) and Pairwise Accuracy (adjacent consistency). Predictions are validated by an online verifier that accepts any sequence consistent with the constraints.

Our experiments reveal several key findings. Across all evaluated VLMs, two general trends emerge: performance significantly degrades as the interaction step length increases, and models consistently perform better on the inverse world modeling task than on the forward one.

Overall, our contributions are threefold: (1) We introduce a simple yet insightful objective for studying embodied cognition, scalable by design, that probes complex reasoning about world interaction through simple question-answering tasks. (2) We built a scalable data generation process on simulated interactions with autonomous annotations, from which we uniformly sampled 8,972 QAs to form the main body of ENACT. (3) We conduct experiments that reveal key limitations in current VLMs, offering insights for improving their long-horizon reasoning, grounding, and embodiment.

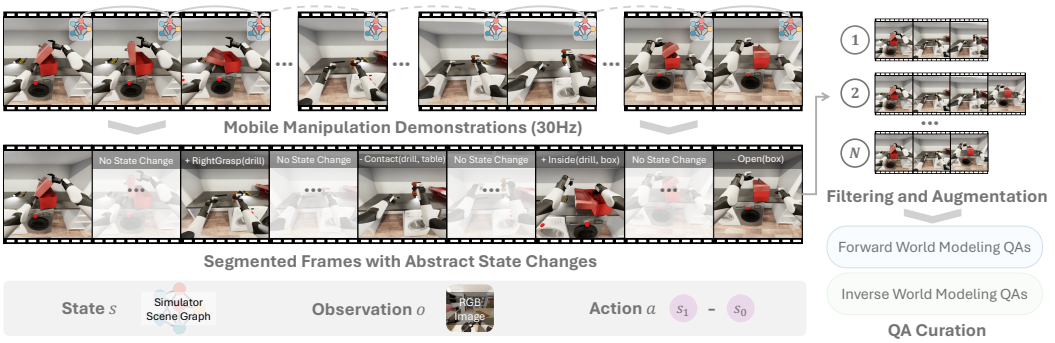

Figure 2: **Overview of ENACT data curation pipeline.** We first replay robot trajectories to obtain aligned scene graphs (states) and RGB observations. The raw trajectory is then segmented by identifying frames where an abstract state change occurs (i.e., the scene graph difference is non-empty). From this set of segmented frames, we sample multiple key-frame trajectories, which are finally used to construct the forward and inverse world modeling questions.

2 ENACT: Egocentric Interactive Embodied Cognition Test

2.1 Problem Formulation

We investigate the *Embodied Cognition* of VLMs by framing it as a world modeling problem, which we probe using egocentric, interactive reasoning tasks. We formulate our benchmark from robot raw dense trajectories, comprised of state-observation pairs $\{(s_t, o_t)\}$. The state s_t is a symbolic scene graph from the simulator state space $\mathcal{G}$, while the observation $o_t \in \mathbb{R}^{H\times W\times 3}$ is the corresponding egocentric RGB image. We view the underlying embodied task as a Partially Observable Markov Decision Process (POMDP, **(author?)** [8]). As shown in Figure 2, we first filter this raw data to identify all timestamps where a semantic change occurs (i.e., the scene-graph difference $\delta(s_t, s_{t-1}) \neq \varnothing$). This process yields a smaller, chronologically ordered set of segmented frames, which serve as the candidate pool for our benchmark.

From the pool of segmented frames, we sample R trajectories, each with a chronologically ordered tuple $\pi = (i_0, \cdots, i_{L-1})$ of L key frames. This initial abstraction into discrete decision epochs is similar to a semi-MDP [9]. However, we treat each of these final key-frame trajectories as a self-contained POMDP instance with scene graphs S_π and observations O_π. For $k = 0, \cdots, L-2$, the action connecting consecutive key frames is the visible scene-graph delta $a_k := \Delta_{\text{Vis}}(s_{i_{k+1}}, s_{i_k})$, where Δ_{Vis} returns the subset of differences in $\delta(s_{i_{k+1}}, s_{i_k})$ that are visible in both images. Together, these actions form a discrete symbolic action space $\mathcal{A}$. For notation simplicity, we relabel indices in π for each key-frame trajectory to $\pi = (0, \cdots, L-1)$ and $(s_k, o_k) := (s_{i_k}, o_{i_k})$.

Building on these trajectories, we formalize two tasks. For **forward world modeling**, given the current image o_0, the correct ordered action sequence $(a_0, \ldots, a_{L-2})$, and a *shuffled list* of next-state images $O' = (o'_1, \ldots, o'_{L-1})$, the model outputs a permutation $\sigma \in \text{Sym}([L-1])$ that orders the images to match the actions: $(o'_{\sigma(1)}, \ldots, o'_{\sigma(L-1)}) = (o_1, \ldots, o_{L-1})$. For **inverse world modeling**, given o_0, the correctly ordered state images $(o_1, \ldots, o_{L-1})$, and a *shuffled list* of actions $A' = (a'_0, \ldots, a'_{L-2})$, the model outputs a permutation $\tau \in \text{Sym}([L-1])$ that orders the actions to be consistent with the state progression: $(a'_{\tau(1)}, \ldots, a'_{\tau(L-1)}) = (a_0, \ldots, a_{L-2})$.

2.2 Key-Frame Trajectories Synthesis for Scalable Data Generation

Segmented Frames with Semantic Changes. Raw robot replays often contain long stretches with no meaningful semantic change (e.g., gripper motion when opening the toolbox in Figure 2) . We mark a timestamp t whenever the simulator state makes a minimal semantic edit (e.g., the robot is now right grasping the drill). The Behavior simulator [10] exposes boolean and relational predicates, where flipping one predicate or updating a relation is our atomic change. A time t enters the candidate pool if the scene-graph difference $\delta(s_t, s_{t-1})$ is nonempty. To avoid near-duplicate frames, we compare each new change with the last accepted segmented frame: we form a predicate-level *change signature*

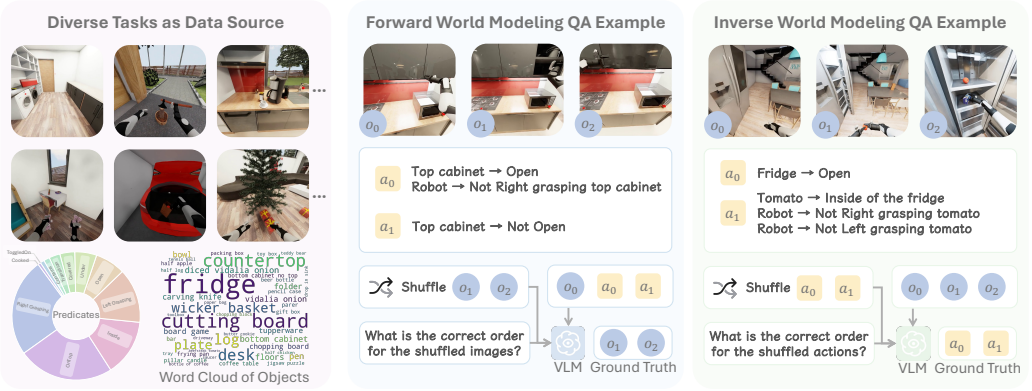

Figure 3: **Data sources and QA examples.** ENACT is built from diverse, long-horizon tasks performed by real robots (Left). We provide examples for (mid) forward world modeling and (right) inverse world modeling.

and keep t only if its cosine similarity with the previous signature is below a threshold. This yields a chronological set of segmented frames $\mathcal{K} = \{t_1 < \cdots < t_M\}$ with (s_{t_i}, o_{t_i}).

Key-Frame Trajectories Synthesis. From the segmented M frames, we sample length-L key-frame trajectories $\pi = (i_0, \ldots, i_{L-1})$ with $1 \leq i_0 < \cdots < i_{L-1} \leq M$, so indices do not need to be adjacent. Each candidate is strictly validated: for every k, the visible state change $\Delta_{\text{Vis}}(s_{i_{k+1}}, s_{i_k})$ is nonempty, and the edited objects are visible in both images, except for object transitioning events (e.g., pineapple being diced), where transient occlusion is permitted. We then treat each valid key-frame trajectory as an individual POMDP instance, with S_π and A_π as defined in the problem formulation. To make data generation *scalable*, we exploit that typically $L < M$ (in practice $L \leq 10$ while $M \gtrsim 30$), and we use skipping to convert trajectory construction into a "seat selection" combinatorics problem, choosing L seats out of M, which yields at most $\binom{M}{L}$ distinct candidates from a single replay. These trajectories are later converted into the forward and inverse world-modeling tasks by shuffling future states or actions, as specified in the problem formulation.

2.3 Dataset Overview and Evaluation Design

Dataset overview. We construct the benchmark from the Behavior simulator and challenge [10]. Behavior Challenge provides 50 long-horizon tasks with up to 200 trajectories per task. We use 29 tasks and replay one trajectory per task to recover aligned pairs $\{(s_t, o_t)\}$. Each replay is segmented into *segmented frames* $\mathcal{K}$, then converted into key-frame trajectories and finally into two QA types: forward world modeling and inverse world modeling (examples in Figure 3). Across step lengths $L \in \{3, \ldots, 10\}$ we sample about 560 items per L for each QA type, yielding 8,972 total questions. The data uses 11 predicate classes (e.g., Inside, Open, Cooked, Grasping) over 149 object categories, and distributions are shown in Figure 3.

Evaluation design. Multiple valid answers can exist for a given question. We therefore use an *online verifier* that accepts any predicted permutation, σ or τ, that is consistent with the corresponding input description constraints. Furthermore, we report two complementary metrics: *Task accuracy* captures exact ordering, while *Pairwise accuracy* grants partial credit for near-correct sequences. Specifically, (1) *Task accuracy* measures exact success at the question level. A question receives score 1 if the verifier accepts the full prediction and 0 otherwise. The dataset score is the average over questions, $\text{TA} = (1/|\mathcal{D}|) \sum_{x \in \mathcal{D}} \mathbf{1}\{\text{accepted}(x)\}$. (2) *Pairwise accuracy* measures stepwise consistency. For a question with length L, we count how many adjacent pairs pass the verifier's local check (state–action for forward; action–state for inverse) and divide by L. We report the micro-average across the split, $\text{PA} = \big(\sum_x \#\text{correct pairs in } x\big) / \big(\sum_x L_x\big)$, which is equivalent to averaging per-item pairwise scores when L is fixed.

Model	Forward World Modeling								Inverse World Modeling							
	3	4	5	6	7	8	9	10	3	4	5	6	7	8	9	10
Proprietary Models																
GPT-5	84.62	75.26	69.96	64.18	57.48	52.16	49.45	46.93	86.28	80.37	76.09	68.78	65.71	62.13	57.12	55.33
GPT-5 mini	87.50	76.25	70.65	63.41	58.14	52.38	46.65	44.11	85.05	76.77	75.43	67.67	63.79	57.04	55.04	50.02
GPT-5 nano	67.83	50.29	38.61	30.35	25.97	21.90	17.59	16.84	72.81	53.95	42.48	36.45	31.68	28.20	24.11	20.33
Gemini 2.5 Pro	86.10	76.42	69.83	60.80	53.26	48.12	40.12	36.98	87.94	81.18	75.39	70.03	66.03	62.91	57.78	56.62
Gemini 2.5 Flash	81.64	67.94	54.17	43.38	37.43	32.73	29.88	28.07	82.78	72.18	60.83	58.19	53.14	51.78	47.99	44.98
Gemini 2.5 Flash-Lite	64.34	49.07	38.70	33.87	27.81	25.44	23.31	20.31	69.58	57.55	46.04	39.09	34.06	30.18	27.51	23.16
Claude Sonnet 4	65.65	45.82	36.65	30.52	26.61	22.78	21.49	20.16	73.25	56.85	48.87	43.07	37.00	32.71	30.50	28.49
Open-Weight Models																
GLM-4.5V	74.30	59.99	47.65	38.78	30.83	25.69	21.60	19.67	80.59	69.28	57.04	51.53	46.95	41.68	37.36	37.93
Llama-4-Mav-17B-128E-Ins	72.47	52.09	43.87	35.30	29.90	25.89	22.79	20.49	72.55	62.60	50.52	43.10	35.17	31.68	28.10	25.80
InternVL3.5-241B-A28B	75.79	62.25	50.83	45.85	37.84	32.88	27.85	25.24	82.26	70.09	60.61	53.38	45.90	39.35	34.12	30.56
Gemma-3-27b-it	63.29	44.66	32.04	25.82	22.11	19.50	16.74	16.29	64.95	48.37	40.04	33.87	28.53	23.63	21.74	19.36
QVQ-72B-Preview	69.14	52.96	40.83	36.27	33.16	30.63	26.30	24.76	71.33	58.77	48.43	44.36	40.26	39.30	36.66	36.58
Qwen2.5-VL-72B-Ins	78.15	60.05	49.87	41.92	36.77	31.73	28.03	25.07	77.80	65.85	53.30	48.19	44.07	37.57	33.76	36.27
Qwen2.5-VL-32B-Ins	67.83	55.46	44.35	35.75	27.52	26.42	22.01	18.07	63.55	59.70	54.57	51.01	49.36	47.17	41.47	40.16
Ovis2.5-9B	58.39	42.51	34.96	31.08	24.61	20.78	18.11	16.96	64.86	51.74	41.65	35.47	30.95	26.64	23.70	23.25
MiniCPM-V-4.5	60.75	38.73	33.65	25.47	24.81	21.40	21.56	18.33	69.23	53.08	47.35	39.55	34.87	30.63	27.05	25.71
Idefics3-8B-Llama3	60.23	36.99	31.83	24.25	21.29	20.80	20.46	17.71	47.38	33.86	27.26	23.48	19.87	18.50	17.04	15.16
Cosmos-Reason1	56.28	41.86	34.75	28.40	26.46	26.49	25.41	24.88	58.30	45.93	44.25	38.50	35.72	34.56	31.50	28.64
Human Performance	**93.62**	**95.30**	**95.04**	**93.87**	**95.43**	**95.41**	**94.75**	**95.13**	**92.15**	**93.85**	**94.77**	**94.58**	**96.23**	**97.74**	**95.21**	**95.46**

Table 1: **Evaluation on ENACT (Pairwise Accuracy).** Dark gray indicates the best result within each category (Proprietary or Open-Weight Models), and Light gray denotes the second-best result within the category.

3 Experiments and Analysis

3.1 World Modeling as a Proxy for Evaluating Embodied Cognition

Experimental Setup. (1) *VLM evaluation setup.* We evaluate 7 proprietary VLMs from 3 families [3, 4, 11] and 22 open-weight models from 10 families [12–20]. For input, all images are resized to 512×512, and we use a unified prompt template per QA type. Models are instructed to return a parsable Python list encoding a permutation of indices. We apply the online verifier in Section 2.3 and report Task Accuracy and Pairwise Accuracy. (2) *Human evaluation setup.* We also recruit trained annotators to answer the benchmark under the same interface and instructions as the models. For inter-annotator agreement (IAA), we uniformly stratify 240 items over QA type and step length, collect independent labels from three annotators, and report Krippendorff's α with 95% bootstrap confidence intervals.

We visualize *Task Accuracy* for GPT-5 and human annotators in Figure 1. Since many models collapse at long horizons ($L = 8$–10, near-zero task success), we focus on the more informative *Pairwise Accuracy*. The main results are in Table 1.

Is inverse world modeling easier than forward? Across families and step lengths, inverse consistently outperforms forward, with the margin widening as L grows. For example, GPT-5 and Gemini 2.5 Pro maintain clear gaps at $L \geq 6$, and open-weight models such as GLM-4.5V and Qwen2.5-VL also show higher inverse scores than forward for most L (see Table 1).

How does performance scale with step length? Accuracy decreases monotonically with L for nearly every model, no matter proprietary or open-weight. Shorter tasks ($L \leq 4$) are manageable for several VLMs, while longer tasks ($L \geq 8$) are challenging even for the strongest models. Pairwise Accuracy slows down the performance drop compared to Task Accuracy, but follows the same trend.

Can SOTA VLMs achieve near-human performance? As we can see from the Table 1, human performance is far better than any tested VLM. SOTA VLMs like GPT-5 mini and Gemini-2.5 Pro achieve comparable performance with humans at step 3, but drop significantly when step length scales.

What is the performance comparison among VLMs? GPT-5 and Gemini 2.5 Pro are the strongest overall in both forward and inverse settings. Several open-weight VLMs are competitive: InternVL3.5-241B-A28B, GLM-4.5V, and Qwen2.5-VL often close much of the gap, and even surpass Claude 4 Sonnet in multiple settings (e.g., inverse at $L = 3$–6). Notably, GPT-5 mini is highly competitive, even achieving the best score in short and mid horizons (e.g., forward at $L = 3, 7, 8$).

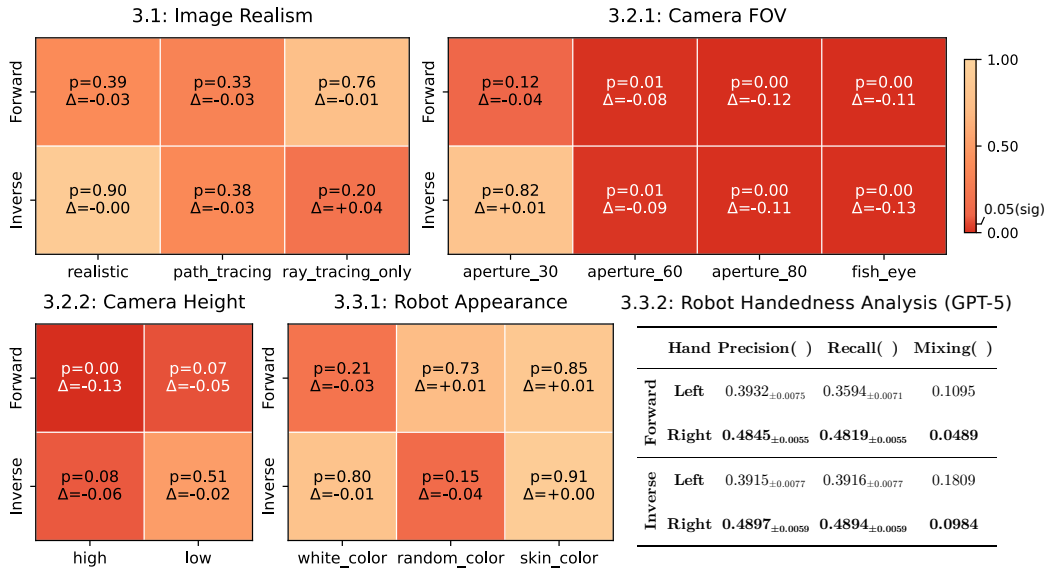

Figure 4: Probing studies' performance delta with the baseline and its significance (deeper the red, the smaller the p-value is).

Key Takeaways: World Modeling as a Proxy for Evaluating Embodied Cognition

- *Inverse consistently surpasses forward, and the margin grows as the horizon L increases.*
- *Accuracy declines steadily with step length L, and all VLMs drop sharper at long horizons.*
- *Humans achieve near-ceiling performance.*

3.2 Are VLMs Sensitive to Image Realism?

Experimental Setup. (1) *Probing configuration.* Motivated by GPT-5 mini's strong cost and performance balance in Section 3.1, we use it as the base model to represent SOTA VLMs. We evaluate step lengths $L \in \{3, 6, 9\}$. For each L and each QA type (forward, inverse), we sample 50 items, yielding 300 total QAs. Question text is held fixed, and we vary only the image source. Outputs are parsed as permutations and scored by the online verifier in Section 2.3 with Pairwise Accuracy. We report, for each setting, the Pairwise Accuracy difference $\Delta = \mathrm{PA}_{\text{baseline}} - \mathrm{PA}_{\text{variant}}$ and two-sided p-values versus the baseline. (2) *Image realism implementation.* Behavior uses Isaac Sim [21], our *baseline* uses Ray Tracing [22] with default global effects. We probe three alternatives on a realism spectrum: *Realistic* (segmented frames translated to a real-world style using GPT-image-1 [23] while we try to maintain the consistency), *Path Tracing* (higher-fidelity rendering, **(author?)** [24]), and *Ray Tracing Only* (Ray Tracing with global effects such as reflections and stage lights disabled). Results are summarized in Figure 4 (panel 3.1).

Does rendering realism change performance? We find no statistically significant degradation or improvement across the spectrum. All settings have $p \geq 0.2$ relative to the baseline, and observed deltas are small across both QA types and all step lengths (Figure 4, A). This suggests the model is not sensitive to image realism in our embodied tasks.

Key Takeaways: Image Realism

- *Robustness to realism variations reduces concern about simulator–real gaps for our tasks.*

3.3 How Do Camera Parameters Affect VLM Performance?

Experimental Setup. (1) *Probing configuration.* We reuse the setup from Section 3.2. We still use GPT-5 mini as the base VLM. (2) *Camera FOV.* The baseline is Aperture 40. We probe Aperture 30, 60, 80, and Fisheye. Rendering and all other parameters are held fixed. (3) *Camera Height.* The

baseline is $(1.75\,\mathrm{m})$ high for eye-level view used in Behavior replays. We probe High $(+0.5\,\mathrm{m})$ and Low $(-0.25\,\mathrm{m})$. We choose $(-0.25\,\mathrm{m})$ since a lower height will consistently make relevant objects invisible. Results are summarized in Figure 4 (panels 3.2.1 and 3.2.2).

Does field of view matter? Figure 4 (B.1) shows the results. A small change to Aperture 30 shows no significant difference from baseline $(p > 0.1)$. Larger deviations substantially hurt performance. Aperture $60, 80$, and Fisheye are consistently and significantly worse than baseline across QA types and step lengths $(p \leq 0.01)$. This suggests that the model performs better with a human-like FOV.

Does camera height matter? As shown in Figure 4 (B.2), raising the camera (*High*) degrades accuracy relative to baseline with statistical significance at the 10% level $(p < 0.10)$ and with negative Δ across settings. Lowering the camera (*Low*) yields mixed effects. Forward shows a mild drop $(p = 0.07,\ \Delta = -0.05)$. The inverse setting is statistically indistinguishable from the baseline $(p = 0.51,\ \Delta = -0.02)$. There are two likely reasons for this. First, the $-0.25\,\mathrm{m}$ shift falls within the typical variation of human eye-level, which keeps the image statistics similar to the pretraining data. Second, the inverse task itself may be less sensitive to small vertical changes because it relies on a fixed state sequence and coarse-grained relationships. Overall, the model performance peaks at a typical human eye level.

Key Takeaways: Camera Parameters

- *The accuracy of models declines as the perspective becomes overly wide or distorted.*
- *Performance is sensitive to camera height, peaking at a typical human eye-level.*

3.4 Do VLMs Have Embodied Biases?

To further understand the nature of VLM embodiment, we investigate two potential biases: **self-awareness** regarding the robot's own body and **handedness asymmetry**, a common trait in humans.

Experimental Setup. We probe these two aspects using distinct experimental setups. (1) *Robot Appearance*. To test for self-awareness, we assess whether VLMs can recognize their embodiment regardless of its appearance. We reuse the probing configuration from Section 3.2, with GPT-5 mini as the base model. The baseline is the default black-and-white robot appearance from the Behavior simulator. We test three variants: White Color (robot is entirely white), Random Color (robot color is randomized at each frame), and Skin Color (robot is rendered with a human-like skin tone). We hypothesize that a model with robust self-awareness will maintain consistent performance across these visual changes. (2) *Handedness Asymmetry.* Inspired by human motor control, where approximately 89% of the population is right-handed [25], we investigate if VLMs exhibit a similar "dominant hand". This analysis does not use the probing configuration but instead relies on a predicate-level error analysis of the main GPT-5 experiment results. We isolate all errors related to the `LeftGrasping` and `RightGrasping` predicates. Using the error analysis framework described in Section 3.5, we treat the ground-truth and model-predicted state differences for each hand as two distinct pools. This allows us to frame the problem in terms of precision and recall. For fair comparison between the hands, which may not appear equally in the data, we report Precision, Recall, and a Mixing rate. The mixing rate measures the proportion of ground-truth state differences for one hand that the model incorrectly attributes to the other. Higher precision and recall with lower mixing indicate greater proficiency with that hand.

Are VLMs aware of their own embodiment, and is this awareness robust to changes in their visual appearance? As shown in Figure 4 (panel 3.3.1), altering the robot's appearance has no statistically significant impact on performance. For all variants (White, Random, Skin Color), the performance deltas are small $(|\Delta| < 0.05)$ and the results are not significant (all $p > 0.10$). This suggests that the model's understanding of its interaction with the world is not tied to a specific visual representation of its own body, indicating a robust sense of self-embodiment within the task context.

Do VLMs exhibit a handedness asymmetry in their interactions with the world? Our analysis of hand-related errors, summarized in Figure 4 (panel 3.3.2), reveals a consistent and strong asymmetry. For both forward and inverse tasks, the right hand consistently outperforms the left hand across all metrics. Precision and recall are substantially higher for the right hand, while the mixing rate (i.e., misattributing a left-hand action to the right hand or vice versa) is significantly lower. For instance, in the forward task, 10.95% of true left-hand changes were incorrectly identified as right-hand changes, whereas only 4.89% of right-hand changes were misattributed to the left. This suggests VLMs are

more prone to making mistakes with their left hand, mirroring the right-hand dominance prevalent in humans.

Key Takeaways: Embodied Biases

- *GPT-5 mini demonstrates strong self-awareness, irrespective of its appearance.*
- *GPT-5 mini demonstrates a significant right-handed bias, which is similar to human handedness.*

3.5 Error Analysis

3.5.1 Preparation for Error Analysis

To gain a deeper insight into the reasoning failures of VLMs, we designed a systematic error analysis framework. Evaluating errors directly from output permutations (e.g., comparing predicted order [3, 2, 1] to ground truth [2, 3, 1] is difficult and often uninformative about the underlying cognitive mistakes. Our approach instead converts the model's output into a format that allows for a direct, fine-grained comparison with the ground truth. For the **forward world modeling** task, we take the model's predicted permutation of images $(o'_{\sigma(1)}, \ldots, o'_{\sigma(L-1)}) = (o_1, \ldots, o_{L-1})$ and compute the corresponding sequence of actions (i.e., visible state differences) that this ordering implies: $\hat{a}_k := \Delta_{\text{Vis}}(s'_{\sigma(k+1)}, s'_{\sigma(k)})$. This yields a predicted action sequence $(\hat{a}_0, \cdots, \hat{a}_{L-2})$. For the **inverse world modeling** task, the model already outputs a predicted action sequence.

With both a predicted and a ground-truth action sequence, we can perform a pairwise comparison at each step k. Each action a_k is a *set* of atomic state differences (e.g., `{add_Open(fridge), remove_Inside(basket, cabinet)}`). By comparing the predicted set $\hat{a}_k$ with the grounded-truth set a_k, we can categorize each atomic state difference. This comparison, akin to analyzing a Venn diagram, yields three primary outcomes for each ground-truth state difference: (1) *Correct*: The state difference is present in both the ground-truth and predicted sets. (2) *Omission*: The state difference is in the ground-truth set but missing from the prediction. (3) *Hallucination*: The state difference is in the predicted set but not in the ground truth.

We assume each state difference is an *independent event* and aggregate these counts across all actions and all questions in the dataset. Based on this framework, we classify errors into five main categories:

1. **Entity Substitution.** The model correctly identifies the state change predicate but applies it to the wrong object(s).
2. **Polarity Inversion.** The model correctly identifies both the object(s) and the predicate, but reverses the polarity of the change (e.g., 'remove' instead of 'add').
3. **Predicate Substitution.** The model correctly identifies the object(s) involved but describes the state change with an incorrect predicate.
4. **Hallucination.** The model predicts a state change that did not occur in the ground truth.
5. **Omission.** The model fails to predict a ground-truth state change that occurred.

3.5.2 Error Distribution Analysis

Our error analysis for GPT-5, shown in Figure 5, reveals that the vast majority of errors fall into two main categories: **Omission** and **Hallucination**. For the forward task, these two error types account for a combined 81% of all failures. This figure is even higher for the inverse task, where they make up nearly 84% of errors. This indicates that the model's primary challenge is not misinterpreting the specifics of a known state change, but rather correctly identifying which changes occurred and which did not. While Omission and Hallucination errors are dominant in both settings, their distribution shifts between tasks. In forward modeling, **Hallucination** is

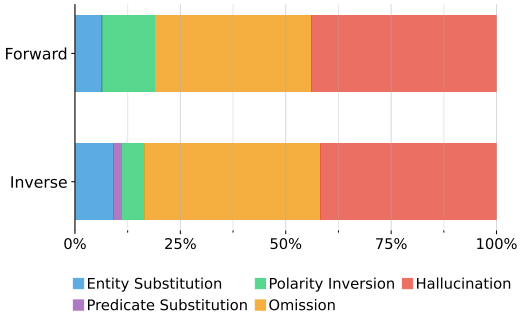

Figure 5: Error distribution (GPT-5) across ENACT, broken down by forward and inverse tasks.

the most common error at **43.9%**, followed by **Omission** at **37.1%**. Remarkably, in the inverse task, these two errors are perfectly balanced, each accounting for exactly **41.8%** of all failures. Other error types are far less frequent. **Polarity Inversion** is more common in the forward setting (12.4%) than the inverse (9.2%). Interestingly, **Entity Substitution** is also slightly more prevalent in the forward task (6.3% vs. 5.4%). Finally, **Predicate Substitution** remains the rarest error type, though it is more pronounced in the inverse setting (1.9%) compared to the forward task (0.3%).

4 Related Work

Embodied Cognition. Our work is grounded in classical *Embodied Cognition*, where cognition arises from brain–body–environment coupling, and meaning is grounded in an agent's sensorimotor repertoire [26–30]. This rich theoretical lineage, spanning ecological views, the extended mind, and sensorimotor accounts [31–35], is supported by empirical findings in psychology and has inspired innovations in robotics and active inference [36–41]. In modern AI, this has motivated a shift towards egocentric, interactive benchmarks (e.g., Ego4D, VLN) [42–45], though the term is sometimes used without theoretical grounding [46].

World Modeling. World Modeling is well aligned with this embodied perspective. It learns action-conditioned dynamics for imagination and planning [47, 48]. Despite achieving scalable imagination and policy gains from counterfactual rollouts [49–52], the grasp of embodied interaction in recent models remains limited. This limitation stems from two issues: many models lack real-world physics grounding due to their reliance on internet video or game data [50, 51], while others, focusing on short-horizon, low-level predictions, fail to maintain causal state progression [53, 54]. Correspondingly, benchmarks for embodied world modeling either score superficial qualities like outcome plausibility and action-video consistency, or remain coarse and non-interactive [55–58, 5, 59, 60, 10, 61]. Recent benchmarks like [62, 63], which emphasize sequencing, but do not verify the consequences of individual actions or the dynamics between states. As raised by [64], we posit that the ability to serve as a sandbox for reasoning and thought experiments is a core function of a world model. To address this, our benchmark is built to probe forward and inverse ordering with a clean action space and a scalable construction.

VLMs in Embodied AI. VLMs are increasingly central to embodied agents, serving as high-level planners that handle task decomposition and subgoal selection [65–70] or as end-to-end policies that directly map vision to action [71–74]. However, current applications and their corresponding benchmarks share critical limitations. Deployments are often confined to tabletop manipulation or simulated environments with limited real-world execution [75]. Similarly, benchmark evaluations tend to prioritize simple instruction-following, neglecting the *multi-step, consequence-aware reasoning* essential for complex interaction [45, 76–78, 7, 61]. Our work addresses this gap by introducing a benchmark focused on egocentric interaction that specifically probes an agent's understanding of *forward and inverse world modeling*.

5 Conclusion

ENACT offers a simple and scalable way to probe how perception and action cohere over time. In extensive simulations, we observe steady degradation with longer horizons and consistently higher accuracy on inverse than forward ordering, while sensitivity analyses suggest limited dependence on rendering realism or robot appearance, but noticeable effects from camera field of view and height. Future work should broaden tasks, diversify environments, and connect sequence ordering to real-robot control to test external validity.

References

[1] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005. 1

[2] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*, 2024. 1

[3] OpenAI. Gpt-5 system card. https://openai.com/index/gpt-5-system-card/, August 2025. Accessed: 2025-09-16. 1, 5

[4] Google DeepMind. Gemini 2.5 pro model card. https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf, June 2025. Updated: June 27, 2025; released to General Availability on June 17, 2025. 1, 5

[5] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 1, 9

[6] Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, et al. Do vision-language models have internal world models? towards an atomic evaluation. *arXiv preprint arXiv:2506.21876*, 2025. 1

[7] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2024. 1, 9

[8] Karl Johan Åström. Optimal control of markov processes with incomplete state information i. *Journal of mathematical analysis and applications*, 10:174–205, 1965. 2, 3

[9] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999. 3

[10] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024. 3, 4, 9

[11] Anthropic. Claude sonnet 4 model card. https://www.anthropic.com/claude/sonnet, May 2025. Release date: May 22, 2025. 5

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

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

[14] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025.

[15] MetaAI. Llama 4 model card, April 2025.

[16] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[17] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

[18] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[19] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

[20] Qwen Team. Qvq: To see the world with wisdom, December 2024. 5

[21] NVIDIA. Isaac Sim, 2025. Version 5.0.0, Apache-2.0 License. 6

[22] NVIDIA. Nvidia rtx ray tracing, 2021. Accessed: 2025-09-17. 6

[23] OpenAI. Introducing our latest image generation model in the api. OpenAI Blog, 2025. "gpt-image-1" model capabilities and API release. 6

[24] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 6

[25] Marietta Papadatou-Pastou, Eleni Ntolka, Judith Schmitz, Maryanne Martin, Marcus R Munafò, Sebastian Ocklenburg, and Silvia Paracchini. Human handedness: A meta-analysis. *Psychological bulletin*, 146(6):481, 2020. 7

[26] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014. 9

[27] Maurice Merleau-Ponty, Donald Landes, Taylor Carman, and Claude Lefort. *Phenomenology of perception*. Routledge, 2013.

[28] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.

[29] Francisco J Varela, Evan Thompson, and Eleanor Rosch. *The embodied mind, revised edition: Cognitive science and human experience*. MIT press, 2017.

[30] Andy Clark. *Being there: Putting brain, body, and world together again*. MIT press, 1998. 9

[31] Andy Clark and David Chalmers. The extended mind. *analysis*, 58(1):7–19, 1998. 9

[32] J Kevin O'regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–973, 2001.

[33] Lawrence W Barsalou. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660, 1999.

[34] Arthur M Glenberg. What memory is for. *Behavioral and brain sciences*, 20(1):1–19, 1997.

[35] Margaret Wilson. Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4):625–636, 2002. 9

[36] Richard Held and Alan Hein. Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology*, 56(5):872, 1963. 9

[37] Rodney A Brooks. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159, 1991.

[38] Rolf Pfeifer and Christian Scheier. *Understanding intelligence*. MIT press, 2001.

[39] Valentino Braitenberg. *Vehicles: Experiments in synthetic psychology*. MIT press, 1986.

[40] Giovanni Pezzulo and Paul Cisek. Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends in cognitive sciences*, 20(6):414–424, 2016.

[41] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010. 9

[42] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 9

[43] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[44] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[45] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018. 9

[46] Ronghao Dang, Yuqian Yuan, Wenqi Zhang, Yifei Xin, Boqiang Zhang, Long Li, Liuyi Wang, Qinyang Zeng, Xin Li, and Lidong Bing. Ecbench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24593–24602, 2025. 9

[47] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018. 9

[48] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019. 9

[49] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 9

[50] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 9

[51] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 9

[52] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022. 9

[53] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2786–2793. IEEE, 2017. 9

[54] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018. 9

[55] Stephen Tian, Chelsea Finn, and Jiajun Wu. A control-centric benchmark for video prediction. *arXiv preprint arXiv:2304.13723*, 2023. 9

[56] Xiaowei Chi, Chun-Kai Fan, Hengyuan Zhang, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chi-min Chan, Wei Xue, Qifeng Liu, Shanghang Zhang, et al. Eva: An embodied world model for future video anticipation. *arXiv preprint arXiv:2410.15461*, 2024.

[57] Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

[58] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019. 9

[59] Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021. 9

[60] Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 2023. 9

[61] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. 9

[62] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024. 9

[63] Delong Chen, Willy Chung, Yejin Bang, Ziwei Ji, and Pascale Fung. Worldprediction: A benchmark for high-level world modeling and long-horizon procedural planning. *arXiv preprint arXiv:2506.04363*, 2025. 9

[64] Eric Xing, Mingkai Deng, Jinyu Hou, and Zhiting Hu. Critiques of world models. *arXiv preprint arXiv:2507.05169*, 2025. 9

[65] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 9

[66] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[67] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[68] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2022.

[69] Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023.

[70] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. 9

[71] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 9

[72] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[73] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[74] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 9

[75] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023. 9

[76] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022. 9

[77] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[78] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022. 9