TRUE: A TRUSTWORTHY REASONING AND UNIFIED EMBODIED INTELLIGENCE BENCHMARK FOR MULTI-MODAL ROBOTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied vision-language-action (VLA) models are typically evaluated by whether their actions succeed, while the intermediate reasoning that connects perception, task progress, planning, and action remains largely unmeasured. We introduce **TRUE** (**T**rustworthy **R**easoning and **U**nified **E**mbodied Intelligence), a multimodal, multi-task dataset and benchmark for grounded and auditable reasoning in embodied intelligence. TRUE augments embodied trajectories with trustworthy reasoning-and-planning representations that describe task progress, intermediate sub-goals, scene-grounded evidence, and the consequences of recent actions while remaining constrained by the task specification and available observations. TRUE spans three tiers of increasing distance from standard robot learning: tabletop manipulation (CALVIN, LIBERO), CADRE-style lunar multi-rover navigation that we build in NVIDIA Isaac Sim, and real Perseverance Navcam/Hazcam imagery from NASA's Planetary Data System. The benchmark defines complementary two tracks under a fixed deployment interface: *control track* and *reasoning track*: the former measures whether reasoning supervision improves long-horizon action prediction, while the latter measures whether model-generated reasoning remains consistent with observable evidence across domains. We additionally provide a reference embodied model that combines visual observations, language instructions, explicit action history, and trustworthy reasoning within a shared representation. The benchmark reveals that (i) reasoning supervision and action history yield complementary gains in long-horizon control, improving CALVIN ABC→D average length by 38% over OpenVLA-OFT; (ii) gains vanish on near-saturated suites; and (iii) reasoning consistency degrades from simulated navigation (83%) to real planetary imagery (68%), exposing a substantial gap between simulated navigation and real planetary imagery. We release annotations, the Isaac Sim environments, evaluation code, and baselines. TRUE is designed to support joint study of embodied control, interpretability, and cross-task reasoning under a common benchmark.

1 INTRODUCTION

Vision-language-action (VLA) models increasingly connect visual perception, natural-language instructions, and action prediction within general embodied policies (Brohan et al., 2023; Brown et al., 2023; Kim et al., 2024). Progress is commonly measured through task success on manipulation benchmarks such as CALVIN and LIBERO (Mees et al., 2022a; Liu et al., 2023). These evaluations reveal whether a model acts successfully, but provide limited evidence about whether it can represent *what has already happened*, *how far the task has progressed*, and *why a particular action should follow, or whether its intermediate reasoning remains grounded in what is actually observed*. information that long-horizon control depends on, and that safety-critical deployments require to be auditable. In planetary exploration, where autonomous rovers such as NASA's CADRE team operate beyond real-time human supervision (Rabideau et al., 2025), a model's decisions must be accompanied by reasoning that operators can check against what the rovers actually see.

The gap becomes more important as embodied learning moves beyond short-horizon tabletop manipulation. Three gaps prevent the community from studying this systematically. **(1) No grounded**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reasoning labels. Existing robot datasets provide instructions and actions (Collaboration et al., 2025; Khazatsky et al., 2025) but not trajectory-level reasoning that is constrained to visual evidence. **(2) No evaluation of reasoning consistency.** Where models generate reasoning (Zawalski et al., 2025), it is rarely scored against observations. **(3) No path beyond the tabletop.** Benchmarks rarely connect manipulation to navigation, and none to real planetary imagery, where domain shift is extreme and data are scarce. Long-horizon tasks require models to connect current perception with previous actions and intermediate sub-goals, while navigation and planetary settings introduce larger changes in embodiment, visual domain, and planning horizon. Existing datasets rarely provide a common representation of task progress, action consequences, and intermediate planning across these settings, and existing benchmarks generally do not evaluate whether generated reasoning is consistent with the observations on which it is based. As a result, it remains difficult to study whether reasoning supervision improves control, whether interpretable reasoning generalizes across tasks, and where such reasoning fails under domain shift.

We introduce **TRUE** (**T**rustworthy **R**easoning and **U**nified **E**mbodied Intelligence), a multimodal and multi-task dataset and benchmark designed to study these questions that addresses these gaps. TRUE augments embodied trajectories with a shared trustworthy reasoning-and-planning representation that summarizes visible task progress, intermediate sub-goals, relevant scene evidence, and the consequences of recent behavior. We use "trustworthy" in an operational sense: reasoning must be constrained by the task specification, supported by available observations, and auditable against the evidence. This does not imply that generated reasoning is always correct; instead, grounding and consistency are explicitly measured as part of the benchmark.

TRUE spans a progression from standard robot learning to substantially different embodied domains. The manipulation tier uses CALVIN and LIBERO, preserving their original observations, actions, instructions, and official evaluation protocols. The navigation tier introduces CADRE-style multi-rover lunar environments in NVIDIA Isaac Sim, where the same reasoning schema describes terrain, navigation progress, and intermediate planning. The planetary tier uses Perseverance Navcam and Hazcam imagery from NASA's Planetary Data System (PDS), providing a real-world out-of-distribution setting in which reasoning must remain grounded despite a large visual and task-domain shift. Together, these tiers support a common evaluation of reasoning across manipulation, navigation, and planetary data rather than treating each setting as an unrelated application.

The benchmark contains two complementary evaluation tracks. The *control track* asks whether trustworthy reasoning supervision and explicit embodied history improve long-horizon action prediction without requiring an external reasoning model during deployment. The *reasoning track* asks whether a model can generate task-relevant reasoning and planning that remain consistent with the observations. To provide a controlled reference point, we also introduce a reference model that unifies visual observations, language instructions, recent executed actions, and trustworthy reasoning within a shared embodied representation. This model is not the benchmark itself; rather, it tests whether the dataset's reasoning signals are useful for downstream control and whether a unified representation can expose interpretable intermediate reasoning.

Contributions. TRUE makes four contributions.

- **The TRUE dataset.** First, it provides a multimodal, multi-task dataset of trustworthy reasoning-and-planning representations spanning manipulation, simulated lunar navigation, and real planetary observations for the trajectories across four sources and three tiers: manipulation (CALVIN, LIBERO), CADRE-style lunar navigation (new Isaac Sim environments), and planetary data (Perseverance imagery from NASA PDS). Each annotation records task progress, sub-goals, and action consequences, constrained to the instruction and visual evidence (Sec. 3).
- **A scalable, auditable annotation pipeline.** Second, with explicit grounding constraints, a model-agnostic generator, and human verification. It defines a shared reasoning schema and scalable generation pipeline (Sec. 3.3) with explicit grounding constraints and human verification, enabling the same intermediate representation to be studied across heterogeneous embodied settings.
- **The TRUE benchmark.** Third, it introduces complementary control and reasoning benchmark tracks that jointly measure action performance and reasoning consistency under a fixed deployment interface that forbids external reasoning at test time, and standardized protocols across tiers (Sec. 4).
- **Baselines and findings.** Finally, it provides reference models and controlled ablations that quantify when reasoning supervision and action history help, including a 38% improvement in

Resource	Manip.	Nav.	Planetary	Reas. Labels	Grounded	Reas. Eval.
CALVIN (Mees et al., 2022a)	✓					
LIBERO (Liu et al., 2023)	✓					
Open X-Embodiment (Collaboration et al., 2025)	✓	○				
DROID (Khazatsky et al., 2025)	✓					
ECoT (Zawalski et al., 2025)	✓			✓	○	
NASA PDS (NASA Planetary Data System, 2026)			✓			
TRUE (ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison with related robot-learning resources. *Reas. Labels*: trajectory-level reasoning or planning annotations. *Grounded*: labels explicitly restricted to visible evidence. *Reas. Eval.*: a protocol scoring generated reasoning against observations. ○: partial. Entries reflect our reading of each resource.

CALVIN ABC→D average completed length, while the cross-domain reasoning evaluation exposes a clear gap between simulated navigation and real planetary imagery. We evaluate OpenVLA, OpenVLA-OFT, and a unified reference model, and identify where reasoning supervision helps, where it saturates, and where reasoning breaks down under domain shift (Sec. 6).

2 Related Work

Robot learning datasets and benchmarks. Large-scale robot datasets have been the main driver of progress in generalist policies. Open X-Embodiment aggregates demonstrations from many robots and institutions to study cross-embodiment transfer (Collaboration et al., 2025), and DROID collects diverse real-world manipulation demonstrations across scenes and tasks (Khazatsky et al., 2025). Benchmarks provide the corresponding evaluation: CALVIN evaluates long-horizon chains of language-conditioned tasks in which each action changes the scene for the next (Mees et al., 2022a), and LIBERO probes generalization across spatial layouts, object identities, goals, and task horizons (Liu et al., 2023). These resources label instructions and actions and score a single quantity, task success. TRUE builds on CALVIN and LIBERO rather than replacing them: it adds a reasoning layer to their trajectories, keeps their official protocols so that numbers remain comparable to prior work, and adds a reasoning-consistency evaluation that they do not provide. Table 1 summarizes how TRUE relates to existing resources.

Reasoning in embodied models. RT-1 showed that transformer policies trained on diverse demonstrations generalize across tasks (Brohan et al., 2023); RT-2 and PaLM-E connected pretrained vision-language models to control (Brown et al., 2023; Driess et al., 2023); and OpenVLA, OpenVLA-OFT, π_0, $\pi_{0.5}$, and GR00T N1 scaled this recipe (Kim et al., 2024; 2025; Black et al., 2026; Physical Intelligence et al., 2025; NVIDIA et al., 2025). A parallel line equips policies with richer context: multimodal prompts (Jiang et al., 2023; Team et al., 2024), memory of past observations (Torne et al., 2026), intermediate embodied chain-of-thought (Zawalski et al., 2025), and richer task context for steerability (Physical Intelligence et al., 2026). Agent foundation models argue for unifying perception, memory, planning, and action in a single model (Durante et al., 2024; Huang et al., 2024). These works treat reasoning as a modeling device and evaluate it indirectly, through task success. TRUE supplies what is missing to evaluate it directly: reasoning labels constrained to visual evidence, and a protocol that scores generated reasoning against observations.

Planetary robotics data. NASA's Planetary Data System (PDS) archives mission imagery, including Perseverance Navcam and Hazcam frames (NASA Planetary Data System, 2026), but provides no reasoning or task annotations suited to embodied learning. NASA's CADRE mission demonstrates cooperative multi-rover autonomy on the lunar surface (Rabideau et al., 2025), yet no public learning benchmark reproduces its setting, because real planetary-robotics data require flight hardware, mission access, and costly field campaigns. TRUE contributes CADRE-style environments in NVIDIA Isaac Sim (NVIDIA) and annotated PDS image groups, and places both on the same evaluation scale as tabletop manipulation.

3 The TRUE Dataset

The design of TRUE is guided by three considerations and principles. **Complementing existing data:** the annotations are intended to supplement, rather than replace, the original observations, actions, instructions, and labels, which are kept unchanged and continue to provide task and action supervision. **Grounding in evidence:** we aim for each annotation statement to be supported by the

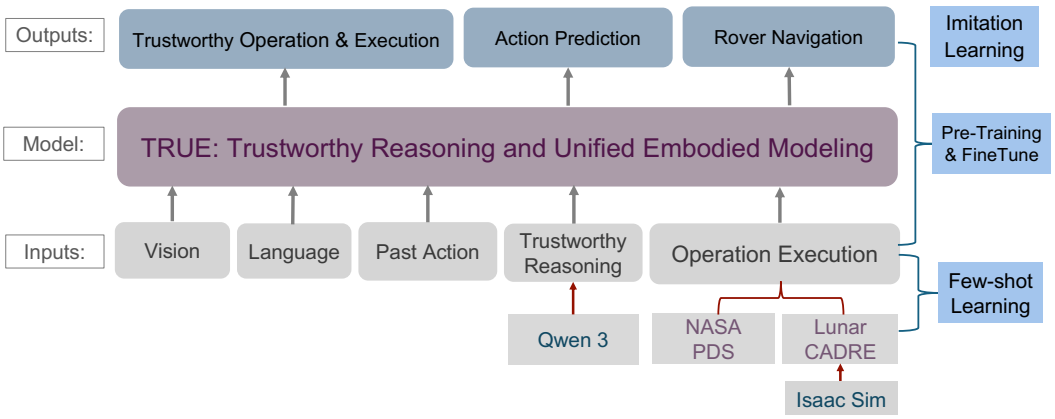


Figure 1: Overview of TRUE. Trajectories from three tiers (manipulation, navigation, planetary data) are paired with trustworthy reasoning-and-planning annotations r^{trp}. The benchmark evaluates models through a fixed interface (o_t, l, h_t) on a control track (action success) and a reasoning track (consistency of generated reasoning with observations).

Tier	Source	Episodes	Frames	Tasks	Avg. words	Track(s)
Manipulation	CALVIN	23k	1.38M	34	57.3	Control
Manipulation	LIBERO (4 suites)	1,693	273k	40	66.1	Control
Navigation	CADRE-style (Isaac Sim, new)	100	7,200	9	60.8	Control, Reasoning
Planetary	NASA PDS (Perseverance)	75 groups	234	4 sols	65.8	Reasoning
Total		25k	1.66M			

Table 2: TRUE dataset statistics.

frames and instruction from which it was generated, so that it can be checked against that evidence. **Gradual domain progression:** data sources are arranged from standard tabletop benchmarks toward real planetary imagery, which can help relate differences in performance to particular kinds of domain shift.

3.1 Sources and Tiers

TRUE is organized as a progression of increasing distance from standard robot learning (Table 2).

Tier 1: Manipulation. We annotate CALVIN and LIBERO trajectories, preserving their observations, actions, instructions, and official splits (Mees et al., 2022a; Liu et al., 2023). CALVIN contributes long-horizon, multi-step behavior across four table environments (A–D); LIBERO contributes controlled variation in spatial layout, object identity, goal, and horizon across four suites.

Tier 2: Navigation. We build CADRE-style lunar environments in NVIDIA Isaac Sim, with rover-mounted cameras, rocky regolith terrain, and navigation instructions, and record rover trajectories with observations and actions. This tier changes the embodiment (mobile rover instead of arm), the action semantics (traversal instead of grasping), and the visual domain (unstructured terrain instead of a tabletop), while retaining full trajectories and simulated ground truth. The action space includes rover driving commands for movement and rotation in the Shackleton Ridge lunar environment, where terrain is varied using hills, mountains, surface bumps and rocks; sampled episodes contained an average of 1.4 rovers, including multi-agent clustered CADRE driving scenarios as well as collisions.

Tier 3: Planetary data. From NASA PDS we collect Perseverance Navcam and Hazcam images. Because these are still images rather than continuous demonstrations, we group related frames and pair each group with the high-level mission objective for the corresponding Martian sol. This tier has no action labels and serves as a real-world, out-of-distribution test of whether reasoning learned elsewhere remains grounded under extreme domain shift.

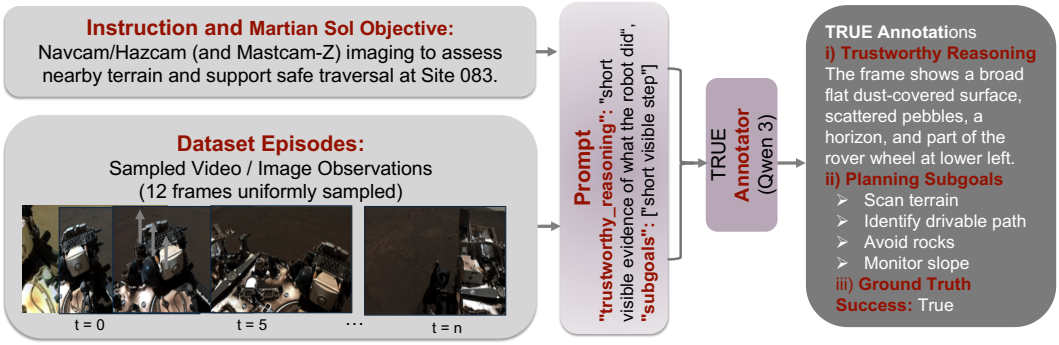


Figure 2: Example TRUE annotation for a PDS observations. The annotation summarizes task-relevant observations, ordered sub-goals, and the effect of the rovers actions.

3.2 Trustworthy Reasoning-and-Planning Schema

Each trajectory τ with instruction l receives one trustworthy reasoning-and-planning annotation r^{trp}. The annotation is associated with the *full trajectory* rather than with a single frame, because the information it encodes - what has been accomplished and in what order - is distributed across the episode. It has three fields:

- **Task progress**: which parts of the instruction are visibly accomplished;
- **Sub-goals (planning)**: one to four short, ordered, visually verifiable sub-goals that decompose l and describe what the robot does over the episode;
- **Action consequences**: a brief statement of the visible evidence of how the robot's behavior changed the scene and constrains what follows.

What "trustworthy" means. An annotation is trustworthy if every statement (i) is consistent with l, (ii) refers only to content visible in the provided frames, and (iii) avoids unsupported inference, including judgments of task success or outcomes absent from the data. The term denotes *auditability* - each statement can be checked against its evidence - not a guarantee of correctness; we measure correctness directly (Sec. 3.3). Annotations are deliberately concise, trajectory-grounded summaries rather than unrestricted chain-of-thought: the generator is not asked to expose hidden intermediate reasoning, only to state what the frames support.

3.3 Reasoning-Supervision Annotation Pipeline and Quality Control

Episode construction. We convert each source into episode-level trajectories using its native episode boundaries, preserving the visual observations, actions, instruction, task label, and all available metadata.

Temporal sampling. Rather than captioning every frame, we sample twelve temporally ordered frames spanning the beginning, middle, and end of each trajectory. This gives the generator the initial scene, intermediate progress, and final state - the minimum needed to describe progress and consequences - while keeping generation tractable across all sources.

Constrained generation. A vision-language model receives the instruction, the task label or metadata when available, and the ordered frames, and is prompted to produce the three schema fields under the trustworthiness constraints. Success labels are never provided to, or modified by, the generator, so annotations cannot leak outcome information into training.

Generator selection. The pipeline is model-agnostic. We compared three candidate VLMs with identical inputs and prompts: Gemini Robotics (Gemini Robotics Team et al., 2025), designed specifically for embodied reasoning; Gemma 3 (Gemma Team et al., 2025), a general-purpose multimodal model; and Qwen 3 (Bai et al., 2025). In pilot comparisons all three produced qualitatively comparable descriptions of progress, sub-goals, and scene content. We selected Qwen 3 because it offered comparable quality at higher generation throughput, which mattered for annotating all four sources. This is an efficiency choice, not a property of the dataset; the released pipeline allows annotations to be regenerated with any VLM.

Human verification. Annotators check a stratified sample of 100 annotations per source against the frames and instruction, marking each field as supported or unsupported. The annotation-level consistency rate is 98%, with inter-annotator agreement of 100% (Appendix B). Unsupported annotations were re-annotated and corrected.

Storage. Each annotation is stored with its trajectory identifier as an additional optional field in the source's native format (RLDS for manipulation), leaving all original fields untouched. Annotations are generated once, offline; they never need to be recomputed at evaluation time.

3.4 Tier-Specific Reasoning Construction

The schema is shared across tiers, but its content adapts to each embodiment. **Manipulation:** annotations describe the scene, the objects involved, the robot's visible progress, and the intermediate steps required by the instruction; demonstrated actions remain the action targets. **Navigation:** the generator receives the navigation or mission instruction and frames from the rover camera, and describes terrain and obstacles, rover position and progress, visible navigation sub-goals, and environmental factors that may affect subsequent motion. **Planetary data:** the generator receives an image group and the objective for its sol, and identifies visible terrain, objects, candidate traversable regions, and mission-relevant context. Because this tier lacks action labels, its annotations are evaluated only for consistency with the images and mission information.

3.5 Splits, Licensing, and Access

Manipulation splits follow the official CALVIN (D→D, ABC→D) and LIBERO protocols, so results remain directly comparable to prior work. For navigation and planetary data we define held-out test splits by training on 80 episodes and 2 martian sols and evaluating on 10 episodes and the 4th martian sol respectively. to prevent leakage between training and evaluation scenes. We release only annotations and pointers for CALVIN and LIBERO, which users download under their original licenses; PDS imagery is public NASA data; our Isaac Sim environments, annotations, repository, code and official project documentation state that the datasets are available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. A datasheet is provided in Appendix C.

4 The TRUE Benchmark

4.1 Deployment Interface

To make results comparable and to prevent leakage, all evaluated models receive only the current observation o_t, the instruction l, and optionally the executed-action history $h_t = [a_{t-H}, \ldots, a_{t-1}]$. Reasoning annotations may be used *only during training*. External reasoning models - including the annotation generator - may not be called at test time. This interface matches that of standard VLA policies, so any model evaluated on TRUE can be deployed without additional infrastructure, and any improvement must come from what the model learned rather than from information supplied at test time.

4.2 Control Track

Track	Tier / data	Metric
Control	CALVIN D→D, ABC→D	Avg. len. (0–5)
Control	LIBERO (4 suites)	Success (%)
Control	CADRE-style nav.	Nav. success (%)
Reasoning	CADRE-style nav.	Consistency (%)
Reasoning	NASA PDS	Consistency (%)

Table 3: Benchmark tracks and metrics.

The control track asks whether reasoning supervision improves action prediction. On CALVIN, models are evaluated on 1,000 chains of five consecutive language-conditioned tasks under the official protocol, reporting per-step success and average completed length; D→D tests in-distribution chaining and ABC→D tests transfer to a held-out environment. On LIBERO, models report success on the Spatial, Object, Goal, and Long suites under the standard protocol. On CADRE-style navigation, whose Isaac Sim trajectories provide rover driving commands as action labels, models predict rover actions from (o_t, l, h_t) on the held-out test episodes and report navigation success; this extends the control track from manipulation to mobile planetary robotics under the same deployment interface.

Reasoning-coverage axis. To measure how control scales with reasoning supervision, the track defines three training conditions: no annotations ($p = 0$), annotations for a random half of training samples ($p = 0.5$), and annotations for all samples ($p = 1$). All other training inputs, action

targets, and optimization settings are held fixed across conditions, which isolates the effect of adding trajectory-grounded reasoning from changes to the underlying policy or action objective.

Pathway ablations. Methods that use both action history and reasoning are expected to report the effect of removing each, under the same training budget, so that their gains can be attributed.

4.3 Reasoning Track

The reasoning track asks whether a model's *own* reasoning, generated at inference from (o_t, l, h_t) alone, is consistent with the observations and objective. A generated output is *consistent* if all of its statements about the scene, objects, task progress, sub-goals, and planned behavior are supported by the available evidence; a single unsupported statement makes the output inconsistent. The metric is the fraction of consistent outputs, judged by human raters (protocol in Appendix B). This deliberately strict criterion reflects the planetary-operations use case, where one hallucinated obstacle or traversable region is enough to mislead an operator. The navigation tier tests transfer from manipulation to mobile robotics; the planetary tier tests transfer from simulation to real mission imagery. On both tiers, this track measures reasoning consistency; action performance on the navigation tier is measured separately by the control track.

4.4 Cross-Task and Cross-Domain Evaluation

Beyond reporting each tier independently, TRUE uses the three-tier design to examine how reasoning quality changes as evaluation moves away from standard manipulation. CALVIN and LIBERO measure control in familiar robot-learning settings; CADRE-style Isaac Sim navigation changes the embodiment, action semantics, and visual environment while retaining simulated trajectories with action labels, and is therefore evaluated on both tracks; PDS introduces real planetary observations without conventional action labels and is evaluated on the reasoning track only. We therefore report tier-specific results rather than collapsing the tiers into a single score. This progression is intended to expose where a reasoning representation that is useful for embodied control ceases to remain grounded under increasingly different task and observation distributions.

5 Benchmark Baselines

We evaluate OpenVLA (Kim et al., 2024) and OpenVLA-OFT (Kim et al., 2025), and report literature results for HULC (Mees et al., 2022b), π_0 (Black et al., 2026), and SmolVLA. To establish how much the benchmark's reasoning supervision can be exploited, we also provide a **reference model**, TRUE-Ref, built on OpenVLA-OFT and unifying all four information sources. A detailed description of the reference architecture, input representation, training objective, and inference procedure is provided in Appendix J. Its input sequence is

$$X_t = \big[E_v(o_t);\ E_l(l);\ f_\phi(h_t);\ z \cdot E_l(r^{\mathrm{trp}})\big], \quad z \sim \mathrm{Bernoulli}(p), \tag{1}$$

where E_v is the fused DINOv2/SigLIP visual encoder, E_l the Llama 2 token embedding (Oquab et al., 2024; Zhai et al., 2023; Touvron et al., 2023), and f_ϕ a two-layer MLP mapping the $H = 4$ most recent executed actions (zero-padded at episode start) to a single token. A shared backbone produces a representation H_t, from which OpenVLA-OFT's parallel continuous head predicts the action chunk and the language head decodes reasoning. Training minimizes

$$\mathcal{L} = \lambda_{\mathrm{act}} \|a_t - \hat{a}_t\|_1 + \lambda_{\mathrm{reason}}\, \mathcal{L}_{\mathrm{reason}}, \tag{2}$$

where $\mathcal{L}_{\mathrm{reason}}$ is the causal next-token loss on r^{trp} when $z = 1$; under causal masking, each reasoning token is predicted only from (o_t, l, h_t) and the preceding reasoning tokens. At test time $z = 0$, complying with the deployment interface, and decoded reasoning is not fed back to the action head. We ablate the history pathway ("w/o history") and the reasoning pathway ("w/o reasoning", i.e., $p = 0$). All models are fine-tuned with LoRA (rank 32) on 4×A100 GPUs with identical hyperparameters and seeds within each benchmark (Appendix I).

6 Benchmark Results

Method	1	2	3	4	5	Avg. Len.
CALVIN D→D: reasoning coverage						
HULC† (Mees et al., 2022b)	82.7%	64.9%	50.4%	38.5%	28.3%	2.64
OpenVLA	64.3%	42.1%	27.4%	16.0%	9.2%	1.59
OpenVLA-OFT	70.2%	48.5%	28.7%	19.8%	9.9%	1.77
TRUE-Ref ($p = 0$)	76.7%	56.8%	41.0%	29.8%	20.5%	2.25
TRUE-Ref ($p = 0.5$)	85.9%	68.2%	50.7%	37.0%	25.9%	2.68
TRUE-Ref ($p = 1$)	**87.6%**	**75.7%**	**62.6%**	**52.6%**	**44.0%**	**3.26**
CALVIN ABC→D: pathway ablation						
HULC†	88.9%	73.3%	58.7%	47.5%	38.3%	3.06
OpenVLA-OFT	88.0%	71.0%	53.0%	30.0%	24.0%	2.66
TRUE-Ref w/o reasoning	91.7%	77.4%	63.5%	51.0%	39.6%	3.23
TRUE-Ref w/o history	92.2%	75.5%	57.6%	42.1%	28.9%	2.96
TRUE-Ref (full)	**94.1%**	**82.3%**	**76.4%**	**62.7%**	**50.9%**	**3.67**

Table 4: Control track, CALVIN: success after k consecutive tasks and average completed length (max 5). **Top:** reasoning-coverage axis in D→D (history enabled for all TRUE-Ref variants). **Bottom:** held-out ABC→D with pathway ablations. †Reported from the literature.

Finding 1: Reasoning supervision improves long-horizon control and scales with coverage. On CALVIN D→D (Table 4, top), action history alone ($p = 0$) raises average length from 1.77 to 2.25, and each increase in reasoning coverage yields further gains, reaching 3.26 at $p = 1$. The dose–response is monotonic: moving from $p = 0$ to $p = 0.5$ adds 0.43, and from $p = 0.5$ to $p = 1$ adds a further 0.58, with no sign of saturation at full coverage. This suggests that the dataset's reasoning supervision, rather than a one-time regularization effect, is what drives the improvement.

Model	CADRE nav.	PDS
OpenVLA	38	27
OpenVLA-OFT	46	36
TRUE-Ref	**83**	**68**

Table 5: Reasoning track: human-judged consistency of inference-time reasoning (%).

Finding 2: History and reasoning are complementary under environment shift. In the held-out ABC→D setting (Table 4, bottom), each pathway alone improves over OpenVLA-OFT (3.23 with history only; 2.96 with reasoning only), and combining them gives 3.67 (+38%). History contributes the larger individual gain, while reasoning adds most at later chain steps: at task 3, adding reasoning to history raises success from 63.5% to 76.4%. The two pathways therefore carry different information - history records what the robot just did, while reasoning encodes where the robot is in the task - and neither subsumes the other.

Finding 3: The benchmark's discriminative power lies in late chain steps. Figure 3 plots ABC→D success against chain length. At the first task, all methods lie within 6.1 points (88.0–94.1%); by the fifth task, the spread grows to 26.9 points (24.0–50.9%). The same pattern holds in D→D, where fifth-task success rises from 9.9% (OpenVLA-OFT) to 44.0% (TRUE-Ref, $p = 1$), a more than fourfold increase, while first-task success rises by only 17.4 points. Late chain steps are exactly where earlier actions have reshaped the scene and where progress tracking matters most. For benchmark users, this implies that average length and late-step success are the most informative control metrics, and that single-task success alone understates differences between methods.

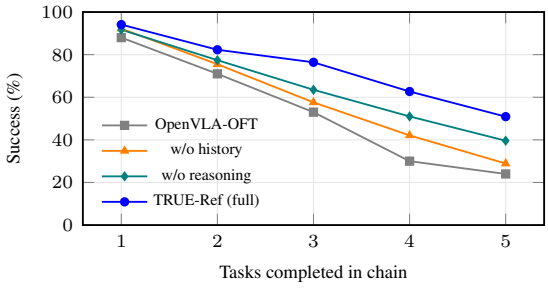


Figure 3: CALVIN ABC→D success versus chain length (data from Table 4). The gap between TRUE-Ref and OpenVLA-OFT widens with each step.

Finding 4: Gains vanish on saturated suites. On LIBERO (Table 6), TRUE-Ref is on par with OpenVLA-OFT (97.3% vs. 97.1% average), with small gains on Spatial, Object, and Long and a 1.0-point drop on Goal - differences within expected seed variance. The largest (still small) gain appears on LIBERO-Long, the suite closest to CALVIN's multi-step structure, consistent with Finding 3.

With the strongest baseline near ceiling, LIBERO no longer discriminates between models on this axis. We therefore retain LIBERO for comparability with prior work but recommend CALVIN's long-horizon chains and the reasoning track as the more informative parts of the benchmark.

Method	Spat.	Obj.	Goal	Long	Avg.
π_0 † (Black et al., 2026)	90.0	86.0	95.0	73.0	86.0
SmolVLA†	93.0	94.0	91.0	77.0	88.8
OpenVLA† (Kim et al., 2024)	83.5	86.9	80.1	53.4	75.9
OpenVLA-OFT	97.6	98.4	**97.9**	94.5	97.1
TRUE-Ref	**97.9**	**98.9**	96.9	**95.5**	**97.3**

Table 6: Control track, LIBERO success (%). †From the literature.

Finding 5: Reasoning consistency degrades from simulation to planetary imagery. On the reasoning track (Table 5), TRUE-Ref's reasoning is judged consistent in 83% of CADRE-style navigation cases and 68% of PDS cases, versus 46% and 36% for OpenVLA-OFT and 38% and 27% for OpenVLA. Training with reasoning supervision roughly doubles consistency on both tiers. Yet every model drops by 10–15 points from simulated navigation to real Perseverance imagery, and the ranking of models is preserved across tiers. This sim-to-planetary gap is the benchmark's most open challenge: even the best reference model produces at least one unsupported statement in about one third of real planetary cases, a rate that is far from acceptable for mission operations. For OpenVLA and OpenVLA-OFT, we get free-form text by decoding the shared language model.

Discussion. Taken together, the two tracks measure different things and should be read together. The control track shows that grounded reasoning supervision can be converted into better long-horizon actions, but only where the task structure requires tracking progress (Findings 1–4). The reasoning track shows that the same supervision makes a model's own reasoning substantially more faithful to what it observes, but that this faithfulness does not yet survive the shift to real planetary imagery (Finding 5). A model that improves on one track but not the other would reveal a trade-off between control and auditability that success-only benchmarks cannot detect.

7 Limitations

Annotation source. Annotations are VLM-generated under grounding constraints and verified on a sample; residual errors remain, and the choice of generator and prompt may bias style and content. **Evaluation cost.** The reasoning track currently relies on human raters, which limits leaderboard scalability; a validated automatic judge is future work. **Scope.** Navigation is simulated, the planetary tier has no action labels, and closed-loop success on physical rovers is not measured. Physical-robot, multi-embodiment, and larger planetary datasets would further strengthen the benchmark. **History length.** The reference model uses a fixed four-step history; longer or adaptive windows may help at higher cost. **Statistics.** Baselines use a single training seed; we will report multi-seed results with the release.

8 Conclusion

We introduced **TRUE**, a multimodal, multi-task dataset and benchmark for trustworthy reasoning and unified embodied intelligence. TRUE connects manipulation, simulated lunar navigation, and real planetary observations through a shared reasoning-and-planning schema that makes task progress, sub-goals, scene evidence, and action consequences explicit and auditable. Its complementary control and reasoning tracks evaluate both whether reasoning supervision improves embodied action prediction and whether model-generated reasoning remains grounded in observable evidence. A reference model further demonstrates how TRUE's reasoning signals and explicit action history can be incorporated into a unified embodied representation, yielding substantial gains on long-horizon CALVIN while exposing a persistent reasoning gap between simulated navigation and real PDS imagery. By releasing the reasoning representations, benchmark protocols, Isaac Sim environments, evaluation code, and reference baselines, TRUE provides a common testbed for studying control, interpretability, and cross-domain reasoning in multimodal embodied intelligence.

Ethics Statement

TRUE contains no personal data. Human raters were volunteers judging model-generated text against robot and rover imagery and were not exposed to harmful content. CALVIN and LIBERO are redistributed only as annotations and pointers under their original licenses; PDS imagery is public NASA data. We are not aware of direct negative societal impacts beyond those general to autonomous robotics.

Reproducibility Statement

The dataset construction is specified in Sec. 3 and Appendix A; benchmark protocols in Sec. 4; baselines and hyperparameters in Sec. 5 and Appendix I; and a datasheet in Appendix C. Annotations, prompts, Isaac Sim environments, evaluation scripts, and baseline checkpoints will be publicly released; an anonymized version is provided in the supplementary material.

AI Usage Statement

A vision-language model (Qwen 3) generated the reasoning-and-planning annotations that form part of the dataset (Sec. 3.3); a sample was verified by human annotators. Generative AI tools also assisted with portions of the code, which the authors inspected, tested, and verified. The authors take full responsibility for the content of this paper.

References

Shuai Bai et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Kevin Black et al. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan et al. Rt-1: Robotics transformer for real-world control at scale, 2023. URL https://arxiv.org/abs/2212.06817.

Noah Brown et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL https://arxiv.org/abs/2307.15818.

Embodiment Collaboration et al. Open x-embodiment: Robotic learning datasets and rt-x models, 2025. URL https://arxiv.org/abs/2310.08864.

Danny Driess et al. Palm-e: An embodied multimodal language model, 2023. URL https://arxiv.org/abs/2303.03378.

Zane Durante et al. An interactive agent foundation model, 2024. URL https://arxiv.org/abs/2402.05929.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723.

Gemini Robotics Team et al. Gemini robotics: Bringing ai into the physical world, 2025. URL https://arxiv.org/abs/2503.20020.

Gemma Team et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Qiuyuan Huang et al. Position paper: Agent ai towards a holistic intelligence, 2024. URL https://arxiv.org/abs/2403.00833.

Yunfan Jiang et al. Vima: General robot manipulation with multimodal prompts, 2023. URL https://arxiv.org/abs/2210.03094.

Alexander Khazatsky et al. Droid: A large-scale in-the-wild robot manipulation dataset, 2025. URL https://arxiv.org/abs/2403.12945.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success, 2025. URL https://arxiv.org/abs/2502.19645.

Moo Jin Kim et al. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

Bo Liu et al. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023. URL https://arxiv.org/abs/2306.03310.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks, 2022a. URL https://arxiv.org/abs/2112.03227.

Oier Mees et al. What matters in language conditioned robotic imitation learning over unstructured data, 2022b. URL https://arxiv.org/abs/2204.06252.

NASA Planetary Data System. Nasa planetary data system (pds). https://pds.nasa.gov/, 2026. Accessed: 2026-09-07.

NVIDIA. Isaac Sim. URL https://github.com/isaac-sim/IsaacSim.

NVIDIA et al. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

Maxime Oquab et al. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Physical Intelligence et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025.

Physical Intelligence et al. $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities, 2026. URL https://arxiv.org/abs/2604.15483.

Gregg Rabideau et al. Planning, scheduling, and execution on the moon: the cadre technology demonstration mission, 2025. URL https://arxiv.org/abs/2502.14803.

Octo Model Team et al. Octo: An open-source generalist robot policy, 2024. URL https://arxiv.org/abs/2405.12213.

Marcel Torne et al. Mem: Multi-scale embodied memory for vision language action models, 2026. URL https://arxiv.org/abs/2603.03596.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Michał Zawalski et al. Robotic control via embodied chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2407.08693.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

APPENDIX

A ANNOTATION PROMPT

The prompt requests a single trajectory-level description rather than a caption per frame. It instructs the generator to (i) identify the main task, (ii) list one to four short, visible sub-goals in order, (iii) state the visible evidence of task progress and of how the robot's actions changed the scene, (iv) refer only to content visible in the provided frames, and (v) not judge task success. The same prompt template is used across tiers, with the instruction field replaced by the navigation instruction (CADRE-style) or the sol objective (PDS).

Prompts A sample prompt is shown below; we name the dataset respectively and keep the same structure across sources:

```
You are a robotics trajectory annotator.
You are given RGB frames from one
episode (CALVIN, LIBERO, NASA Perseverance rover,
or Isaac Sim rover).

Use only visible evidence. Do not invent
objects, actions, or events.
If a dataset task string is provided, treat it as given metadata.

Return exactly one JSON object with the keys
listed in the schema.
No markdown. Keep the entire JSON under 80 words.

{
  "task_instruction": "string
  (given for Manipulation; guessed for Planetary Robotics)",
  "task_label": "string (optional)",
  "objects": ["string (Planetary Robotics)"],
  "subgoals": ["short visible step"],
  "trustworthy_reasoning": "short visible evidence ",
  "trustworthy_value_alignment": {
    "evidence_grounding": "string (Planetary Robotics)",
    "uncertainty_notes": "string (Planetary Robotics)"
  }
}
```

B HUMAN VERIFICATION AND EVALUATION PROTOCOL

Annotation verification. Annotators compare each sampled annotation with its frames and instruction and mark each field as supported or unsupported. 20 annotators independently reviewed 100 sampled instances each, for a total of 2,000 verification instances: 60 from CALVIN, 30 from LIBERO, 5 from PDS, and 5 from CADRE-style navigation. The annotators achieved 100% agreement; any unsupported annotation was re-annotated and independently rechecked.

Reasoning-track evaluation. Raters compare each model's inference-time reasoning with the objective and images supplied to the model; an output is consistent if all its statements are supported by the evidence. For PDS, 24 raters compared outputs against Perseverance imagery and mission information; for CADRE-style navigation, raters used the simulated episodes and instructions.For the reasoning-track evaluation, 24 raters, blinded to model identity, evaluated 10 samples from each of the four datasets. Agreement was achieved by 23 of the 24 raters across datasets, corresponding to 95.8% inter-rater agreement.

Table 7 shows example outputs.

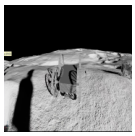
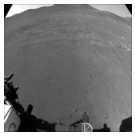
Setting	Image	Task Instruction	Model Output
LIBERO		put both the alphabet soup and the cream cheese box in the basket	`the basket is`
CALVIN		place the grasped object in the sliding cabinet	`to the left of the red one`
CADRE-style Navigation		Drive the rover across the rocky terrain.	`group across the rocky terrain.`
PDS (Perse-verance)		Assess traversable terrain ahead and maintain safe rover path.	`terrain behind`

Table 7: Example inference-time reasoning outputs of TRUE-Ref across the three tiers.

C DATASHEET

Following the datasheet framework of Gebru et al. (2021), we document the motivation, composition, construction process, preprocessing, recommended uses, limitations, distribution, and maintenance plan for the TRUE benchmark. TRUE combines existing embodied-learning resources with newly constructed reasoning-and-planning supervision and simulated navigation trajectories. We therefore distinguish between data inherited from the original sources and data or reasoning representations introduced by TRUE.

Motivation. TRUE is designed to support the study of *trustworthy reasoning in embodied intelligence*. Existing embodied-learning benchmarks primarily evaluate whether a policy predicts appropriate actions or successfully completes a task, while the intermediate reasoning that connects observations, task progress, prior actions, planning, and subsequent behavior is often implicit. This makes it difficult to evaluate whether an embodied model's decisions are supported by observable evidence or whether its intermediate interpretation of a trajectory is consistent with what has actually occurred.

The benchmark addresses this gap by introducing a common reasoning-and-planning representation across heterogeneous embodied settings. In particular, TRUE is intended to support three related research questions: whether explicit reasoning supervision can improve downstream embodied control, whether generated reasoning remains grounded in observations and task specifications, and whether a common reasoning representation can transfer across substantially different tasks and visual domains. The benchmark therefore spans manipulation, simulated planetary navigation, and real planetary imagery rather than focusing on a single robot embodiment or environment.

Composition. TRUE consists of three complementary data tiers. The first tier is constructed from CALVIN and LIBERO manipulation trajectories and augments the original observations, language instructions, and actions with trajectory-level trustworthy reasoning-and-planning representations. These representations describe observable task progress, intermediate sub-goals, relevant scene state, and the consequences of recently executed actions.

The second tier contains CADRE-style multi-rover navigation trajectories generated in NVIDIA Isaac Sim. These trajectories provide a substantially different embodied setting in which the reasoning representation describes navigation progress, terrain and scene context, intermediate navigation

objectives, and planning decisions. The CADRE-style environments are intended as research simulations inspired by multi-rover planetary exploration rather than reproductions of flight operations.

The third tier contains grouped Perseverance Navcam and Hazcam observations obtained from NASA's Planetary Data System (PDS). These observations are associated with available mission context and trustworthy reasoning-and-planning representations grounded in visible planetary evidence. Unlike the manipulation and simulated-navigation tiers, this tier is not treated as a conventional action-prediction dataset because the PDS observations do not provide the same form of robot action labels. Instead, it serves as a real-world planetary reasoning and grounding benchmark and introduces a substantial domain shift relative to the simulated environments.

Table 2 summarizes the composition of the three tiers. TRUE contains no intentionally collected personal, biometric, demographic, or other sensitive human data. The benchmark focuses on robot trajectories, simulated environments, and publicly available planetary observations.

Data instances. A TRUE instance contains a task or mission context together with a temporally ordered set of visual observations and a structured reasoning-and-planning target. When available, the instance additionally includes the original action sequence and other trajectory metadata inherited from the source benchmark. For CALVIN and LIBERO, the original action labels and task definitions are retained. For the CADRE-style tier, actions and navigation states are obtained from the generated Isaac Sim trajectories. For the PDS tier, the primary inputs are temporally or contextually grouped rover observations and associated mission context rather than low-level control actions.

The reasoning target is designed to capture information that can be related back to the embodied evidence, including current task progress, observable scene state, intermediate objectives, consequences of recent actions when such actions are available, and the next planning-relevant considerations. The schema is described in Sec. 3.2. Importantly, TRUE does not treat unconstrained chain-of-thought text as ground truth. The released reasoning representations are task-oriented benchmark targets constructed under explicit grounding constraints.

Reasoning-supervision generation. Reasoning-and-planning representations were generated using Qwen 3 under the task and grounding constraints defined in Sec. 3.2. The generation procedure receives sampled observations together with the task specification and, where available, trajectory context. Prompts explicitly instruct the model to base its descriptions on observable evidence and provided task information rather than introducing unsupported scene details.

The same high-level reasoning schema is used across all three tiers, while the information available to the generation process differs by domain. Manipulation reasoning focuses on object state, task progress, recent actions, and intermediate manipulation goals. CADRE-style reasoning emphasizes terrain, navigation state, rover progress, and intermediate planning. PDS reasoning is restricted to evidence that can be supported by the available rover imagery and supplied mission context. This common schema is intended to make cross-task evaluation possible without assuming that the three domains share identical observations or action spaces.

Human verification and quality control. Automatically generated reasoning should not be interpreted as intrinsically correct simply because it was produced under grounding constraints. TRUE therefore treats trustworthiness as an empirical property to be evaluated rather than an assumed property of the generation model. A stratified sample of generated reasoning representations is reviewed by human annotators for consistency with the corresponding observations and task context.

Verification focuses on whether the reasoning is supported by visible evidence, whether it correctly describes task or navigation progress, whether it introduces unsupported objects or events, and whether proposed intermediate planning is compatible with the available context. The sampling procedure is stratified across data tiers to avoid evaluating quality only on the comparatively familiar manipulation domain. We report tier-specific human-consistency results rather than combining all domains into a single aggregate score, because the difficulty and available evidence differ substantially between manipulation, simulation, and real planetary imagery.

The final release will document the number of independently reviewed instances, number of reviewers, agreement procedure, disagreement-resolution policy, and correction criteria. For quality

verification, 20 reviewers examined 100 instances each, for a total of 2,000 verification instances. The reviewers achieved 100% agreement on the evaluated annotations. When an annotation was identified as unsupported, it was reannotated and rechecked before inclusion in the final release.

Preprocessing. For trajectory-based data, twelve frames are sampled from each trajectory to provide a compact representation of temporal progression while maintaining coverage of the beginning, intermediate states, and end of the episode. The sampling procedure is applied consistently within each trajectory-based tier. TRUE does not modify the original CALVIN or LIBERO observations, action labels, or task labels. Instead, the benchmark adds reasoning-and-planning supervision and the metadata necessary to connect the generated representation to the corresponding trajectory.

Image resizing, normalization, tokenization, and other model-specific transformations are not considered part of the dataset itself and are performed by individual reference models during training or evaluation. This separation is intended to avoid tying the benchmark to a particular VLA architecture or visual encoder.

For the PDS tier, rover images are organized into contextually related groups before reasoning construction. Because these data originate from real planetary operations rather than a benchmark designed for action learning, we do not infer or fabricate missing low-level actions. The PDS tier is consequently evaluated through reasoning and grounding metrics rather than conventional robot-control success.

Data splits and leakage prevention. TRUE preserves the official evaluation structure of its source benchmarks whenever applicable. CALVIN experiments follow the established ABC→D generalization protocol, and LIBERO experiments follow the corresponding benchmark task suites and evaluation procedures. TRUE reasoning representations are generated separately for the trajectories assigned to each benchmark partition so that evaluation examples are not intentionally introduced into the training set through the added reasoning supervision.

CADRE-style navigation and PDS data are partitioned at the trajectory, scenario, or observation-group level rather than by individual frame wherever possible, preventing temporally adjacent observations from being distributed across training and evaluation partitions. We split CADRE by episode: 80 for training, 10 for validation, and 10 for testing, using a fixed random seed. We split PDS by Martian day (sol): 2 sols for training and 1 sol held out for validation/test respectively. We will release the split identifiers with the benchmark to enable deterministic reproduction and direct comparison between methods.

Benchmark tasks. TRUE supports two primary evaluation tracks. The *control track* evaluates whether reasoning supervision and explicit action history improve downstream embodied action prediction and long-horizon task completion. This track is applicable to tiers for which appropriate action or control supervision is available, particularly CALVIN, LIBERO, and the CADRE-style simulation.

The *reasoning track* evaluates whether a model can produce task-relevant intermediate reasoning that remains consistent with the corresponding observations. This track can be evaluated across all three tiers and is particularly important for the PDS tier, where real planetary imagery provides an out-of-distribution test of visual grounding. The benchmark intentionally reports control and reasoning separately: a model that performs well at action prediction is not automatically assumed to reason faithfully, and fluent reasoning is not treated as evidence of successful embodied control.

Recommended uses. TRUE is intended for research on multimodal embodied learning, VLA models, temporal representation learning, reasoning supervision, interpretable embodied inference, long-horizon planning, cross-task generalization, and evaluation of grounded reasoning. Researchers may use the dataset to train models with explicit reasoning supervision, study the contribution of action history and intermediate reasoning through ablation, evaluate generated reasoning against observations, and investigate transfer across manipulation, navigation, and planetary domains.

The benchmark may also be useful for studying whether improvements in reasoning consistency correlate with improvements in downstream control. Because the reasoning and control tracks are

separable, TRUE supports models that generate explicit reasoning as well as models that use reasoning only as training-time supervision.

Uses that are not recommended. TRUE is a research benchmark and is *not* intended to certify the safety, reliability, or flight readiness of autonomous robotic systems. In particular, performance on the CADRE-style simulation or PDS reasoning tier should not be interpreted as evidence that a model is suitable for autonomous deployment on a planetary mission. The CADRE-style tier is simulated, while the PDS tier contains observational data without the complete state, operational constraints, action histories, and mission-assurance processes required for real flight autonomy.

The generated reasoning representations should likewise not be treated as infallible ground truth. They are supervision and evaluation resources whose grounding quality is measured and documented, but individual instances may contain errors or ambiguities. Researchers using TRUE for safety-critical applications should perform independent domain-specific verification.

Known limitations and biases. The three tiers differ in embodiment, visual distribution, available metadata, and supervision. Consequently, benchmark scores across tiers are not directly interchangeable, and TRUE does not define a single scalar score intended to collapse all forms of embodied reasoning into one measure. CALVIN and LIBERO are established simulated manipulation environments, CADRE-style trajectories are generated in simulation, and PDS contains real planetary imagery but lacks the low-level action supervision available in the other tiers.

Reasoning supervision may also inherit limitations from the model used during its construction. A language model can produce plausible but unsupported statements, overgeneralize from visual cues, or express planning decisions that are not uniquely determined by the observations. The grounding constraints and human-verification protocol are designed to quantify and reduce these issues rather than to claim that they have been eliminated. Differences in visual complexity and available context may also contribute to the observed gap in reasoning consistency between simulated CADRE-style navigation and PDS imagery.

Source data and ownership. TRUE does not claim ownership of the original CALVIN, LIBERO, or NASA PDS source data. Users should obtain CALVIN and LIBERO observations from their respective official distributions and comply with their original licenses and terms of use. PDS observations remain subject to the applicable NASA/PDS data policies. The TRUE release will provide the newly introduced reasoning-and-planning representations, benchmark metadata, split definitions, evaluation protocols, and the components of the CADRE-style simulated dataset that the authors are permitted to redistribute.

Where redistribution of original source observations is restricted or unnecessary, TRUE will provide identifiers or mappings that associate the released reasoning representations with the corresponding source instances rather than repackaging the original data.

Distribution and accessibility. The benchmark release will include reasoning-and-planning representations, dataset metadata, deterministic split definitions, benchmark evaluation code, reference-model configurations, and documentation describing how to obtain or reconstruct the corresponding source observations. We additionally plan to release the CADRE-style Isaac Sim benchmark assets and trajectory-generation configuration needed to reproduce the simulated navigation tier, subject to the applicable software and asset licenses.

The archival release, persistent identifier, and license for the TRUE-specific components will be provided with the final dataset release. The code will be released upon acceptance.

Reproducibility. To facilitate reproducible evaluation, the release will include the prompts and reasoning schema used to construct the supervision, preprocessing and temporal-sampling procedures, train/validation/test identifiers, evaluation scripts, and reference-model hyperparameters. Benchmark results should be reported separately for each tier and evaluation track. We encourage future work to report both task performance and reasoning-grounding metrics when applicable, rather than using task success alone as a proxy for trustworthy reasoning.

For model-generated reasoning, evaluation should use the released human-verification criteria and should distinguish unsupported statements from task-consistent but non-unique planning interpre-

tations. This distinction is particularly important in open-ended navigation and planetary settings, where multiple valid plans may exist.

Maintenance and versioning. The authors will maintain TRUE as a versioned benchmark. Reported reasoning errors, broken source mappings, split inconsistencies, or evaluation-code issues will be reviewed and corrected in subsequent releases. Changes that alter benchmark instances, reasoning targets, or evaluation procedures will be documented through version numbers and release notes so that previously reported results remain traceable to the benchmark version on which they were obtained.

We also plan to maintain a benchmark results page containing results submitted under the standardized evaluation protocol. Future additions to TRUE may introduce new embodiments or domains, but existing evaluation partitions will remain available to preserve comparability with earlier work. A project website hosted on GitHub Pages will be made publicly available upon acceptance; it is omitted here to preserve anonymity during review.

Ethical and societal considerations. TRUE contains no intentionally collected personal or sensitive human information. Nevertheless, the benchmark concerns autonomous robotic reasoning, and improvements in embodied planning can have broader implications beyond the research settings considered here. We therefore emphasize that TRUE evaluates research models under benchmark and simulation conditions and does not establish operational safety.

The term "trustworthy" refers specifically to the benchmark's operational criteria of task constraint, observational grounding, and auditability. It should not be interpreted as a general certification of model reliability, factual correctness, robustness, or safety. We encourage users of the benchmark to preserve this distinction when reporting results and to document failure cases in addition to aggregate performance.

D Episode Construction and Temporal Sampling

We convert each dataset into episode-level trajectories using the episode boundaries provided by the original dataset. For each episode, we preserve the visual observations, robot actions, task instruction, task label, and available dataset metadata.

The visual observations remain in temporal order so that the annotator can capture changes in the scene and the robot's behavior throughout the episode. Rather than annotating every frame independently, we uniformly sample 12 observations from each trajectory and provide them to the robotics trajectory annotator.

The twelve observations are selected from the beginning, middle, and end of each episode. This sampling strategy provides the annotator with information about the initial scene, intermediate task progress, and final state while keeping the annotation input computationally manageable.

The annotator receives the original task instruction, the dataset-provided task label when available, and the ordered visual observations. The success labels remain unchanged from the original datasets. They are not provided to or modified by the VLM annotator; the reasoning annotations are generated independently while remaining grounded in the observed visual trajectory.

E Reasoning-Supervision Generation Objective

The annotation prompt asks the VLM to produce a concise trajectory-level description rather than a separate caption for every frame. The generated text is expected to identify the main task, describe relevant scene information, summarize task progress, and mention visible intermediate sub-goals.

Our prompt is designed so that generated outputs are a concise, trajectory-grounded representation rather than unrestricted chain-of-thought. The VLM is not asked to expose hidden intermediate reasoning or judge task success. Instead, it produces a short summary based only on visible evidence in the sampled frames, making relevant task progress available as additional context during VLA training.

The annotation contains two types of information. First, it lists one to four short, visible subgoals that describe the robot's actions throughout the episode. Second, it provides a brief trustworthy reasoning statement summarizing visible evidence of what the robot did. The task instruction and task label provide additional task context, while the generated output remains grounded in the observed trajectory and is stored as a supplementary field associated with the original data.

F Selection of the Reasoning-Supervision Model

TRUE is designed to be independent of any particular vision-language model. To select the annotator used in our current implementation, we compared several candidate VLMs using the same task information, visual observations, and annotation prompt. The candidate models included Gemma 3 Gemma Team et al. (2025), Gemini Robotics Gemini Robotics Team et al. (2025), and Qwen 3 Bai et al. (2025).

Gemini Robotics was considered because it is specifically designed for embodied reasoning and robot-task understanding. Gemma 3 was also considered as a general-purpose multi-modal model. In pilot comparisons using the same sampled observations and prompt, the candidate models produced qualitatively comparable descriptions of task progress, visible sub-goals, and relevant scene information.

However we selected Qwen 3 for the current annotation pipeline because it provided comparable annotation quality with higher token-generation throughput in our setup. This difference was important for processing a large number of robot demonstrations across the four datasets. The selection of Qwen 3 is therefore an implementation choice motivated by annotation efficiency and does not define the TRUE architecture. Other VLMs can be used to generate the same type of trajectory-grounded context.

G Dataset-Specific Reasoning-Supervision Procedures

CALVIN and LIBERO. For CALVIN and LIBERO, the annotator receives the dataset-provided task instruction, the task label or task metadata when available, and twelve visual observations sampled from the demonstration. The generated text describes the manipulation scene, the objects involved in the task, the robot's visible progress, and the intermediate steps required to complete the instruction. The original demonstrated actions remain the action targets used for VLA training.

CADRE-style navigation in Isaac Sim. For the CADRE-style experiments, we generate demonstrations in NVIDIA Isaac Sim using a simulated rover environment. A camera attached to a rover provides the visual observations, and the annotator receives the corresponding navigation or mission instruction together with observations sampled throughout the episode. The generated text describes relevant terrain and objects, rover position and progress, visible navigation sub-goals, and environmental information that may affect subsequent rover behavior. This procedure tests whether the same annotation framework can describe navigation and multi-agent robotic behavior rather than only tabletop manipulation.

NASA Planetary Data System observations. For the PDS experiments, we use still images from Perseverance rover observations, including Navcam and Hazcam data. Because these observations are not always provided as continuous robot demonstrations, we group related images and provide the annotator with the corresponding high-level mission objective or objective associated with the relevant Martian sol. The generated descriptions identify visible terrain, objects, possible traversable regions, and mission-relevant context. These annotations are evaluated for consistency with the provided images and mission information rather than with a conventional manipulation-policy success metric.

H Reasoning-Supervision Coverage Ablation

To examine the effect of annotation availability, we train models using different levels of annotation coverage. The baseline setting uses no trustworthy reasoning annotations. Additional settings enable

the generated annotations for either a subset or the full set of training demonstrations. All other training inputs, action targets, and optimization settings remain unchanged across these conditions.

This comparison separates the effect of adding trajectory-grounded language from changes to the underlying VLA policy or action objective. The full TRUE configuration uses the trustworthy reasoning representation together with recent executed action history, while the corresponding ablations remove one of these inputs. The results of these coverage and component comparisons are reported in the main experimental section.

I Implementation Details

Reference-model training. The reference model uses the same optimization settings across component ablations so that the effects of reasoning supervision and past-action conditioning can be isolated. Training uses LoRA with rank 32, AdamW optimization, an L1 action-regression objective, and four NVIDIA A100 GPUs. The action-history window contains the four most recently executed actions, with unavailable entries zero-padded at the beginning of an episode. CALVIN configurations use the same random seed and differ only in the enabled TRUE components; LIBERO follows the corresponding benchmark training recipe, with suite-specific training length where required.

Hyperparameter	Value
Hardware	4× NVIDIA A100
Optimizer	AdamW
Action loss	L1 regression
LoRA rank	32
Global batch size	64
Learning rate (steps 0–100k)	5×10^{-4}
Learning rate (steps 100k–150k)	5×10^{-5}
Max training steps	150k (LIBERO-Goal: 50k)
Checkpoint frequency	every 25k steps
History length H	4
Action history	on / off (ablation)
Reasoning coverage p	0 / 0.5 / 1.0
$\lambda_{\text{act}}, \lambda_{\text{reason}}$	1.0,1.0
History MLP hidden dim	4096

Table 8: Training hyperparameters for all baselines.

All models are fine-tuned on 4×NVIDIA A100 GPUs with LoRA (rank 32), AdamW, and the L1 action loss; Table 8 lists the configuration. On CALVIN, all configurations share one random seed, and ablations vary only the history flag and the reasoning coverage p. On LIBERO we use the same recipe, except that LIBERO-Goal is trained for 50k rather than 150k steps, following OpenVLA-OFT (Kim et al., 2025); seeds and hyperparameters are shared across suites.

J Reference Model

This section details the reference model TRUE-Ref summarized in Sec. 5. TRUE-Ref is not part of the benchmark definition; it is a controlled reference point for measuring how much of TRUE's reasoning supervision can be converted into better control and more consistent reasoning under the deployment interface of Sec. 4.

J.1 Problem Formulation

Consider a language-conditioned trajectory $\tau = \{(o_t, a_t)\}_{t=1}^{T}$ with instruction l and metadata m_τ, where o_t is the visual observation at time t and $a_t \in \mathbb{R}^{d_a}$ the corresponding action or action chunk. A standard VLA policy predicts

$$\hat{a}_t = \pi_\theta(o_t, l). \tag{3}$$

TRUE-Ref adds two sources of information: the executed-action history h_t (Sec. J.4) and, during training only, the trustworthy reasoning-and-planning representation r^{trp} of the trajectory (Sec. 3.2), gated by $z \sim \text{Bernoulli}(p)$:

$$\hat{a}_t = \begin{cases} \pi_\theta(o_t, l, h_t, z \cdot r^{\text{trp}}) & \text{training,} \\ \pi_\theta(o_t, l, h_t) & \text{deployment,} \end{cases} \tag{4}$$

where $z \cdot r^{\text{trp}}$ denotes that r^{trp} is included in the input when $z = 1$ and omitted otherwise. The instruction l is always retained as the authoritative task specification; r^{trp} supplements it and never replaces it.

J.2 ARCHITECTURE

TRUE-Ref is built on OpenVLA-OFT (Kim et al., 2025), whose backbone combines a Llama 2 language model with fused DINOv2 and SigLIP visual features (Touvron et al., 2023; Oquab et al., 2024; Zhai et al., 2023). The four information sources are mapped into a single input sequence (Eq. 1),

$$X_t = \big[E_v(o_t);\ E_l(l);\ f_\phi(h_t);\ z \cdot E_l(r^{\mathrm{trp}})\big],$$

where E_v is the visual encoder, E_l the token embedding, and f_ϕ the history encoder. The shared backbone produces a representation $H_t = F_\theta(X_t)$, from which two heads decode

$$\hat{a}_t = g_{\mathrm{act}}(H_t), \qquad \hat{r}_t \sim p_\theta(r \mid o_t, l, h_t)\,, \tag{5}$$

where g_{act} is OpenVLA-OFT's parallel continuous action head and p_θ is the backbone's autoregressive language head. Because actions and reasoning are decoded from the same backbone state, $\hat{r}_t$ is a readout of information the model uses for control rather than a post-hoc explanation produced by a separate system.

J.3 TRUSTWORTHY REASONING-AND-PLANNING PATHWAY

The reasoning pathway gives the model an explicit representation of task progress, intermediate sub-goals, and the consequences of recent actions. Its training targets are the TRUE annotations r^{trp}, which are generated offline under the grounding constraints of Sec. 3.2: they must be consistent with l, refer only to content visible in the sampled observations, and avoid unsupported inference, including judgments of task success. "Trustworthy" therefore describes the intended grounding and auditability of the targets, not a guarantee that every statement is correct. The annotation generator is used only to construct these targets and is never called at deployment. For CALVIN and LIBERO, the recorded instructions and demonstrated actions remain the sources of task and action supervision; for CADRE-style navigation, the rover driving commands serve as action targets; for PDS, which has no action labels, r^{trp} is used only for the reasoning pathway.

J.4 PAST-ACTION CONDITIONING

The history pathway represents recent robot behavior explicitly. Let

$$h_t = [a_{t-H}, \ldots, a_{t-1}] \tag{6}$$

denote the H actions executed immediately before the current prediction, with $H = 4$. Unavailable entries at the start of an episode are zero-padded, and the target a_t is excluded. A two-layer MLP maps the flattened history to the backbone embedding dimension,

$$x_t^{\mathrm{hist}} = f_\phi(h_t), \tag{7}$$

which enters X_t as a single token. Because h_t contains only already-executed actions, it is available at deployment and adds one token of inference cost, without a separate recurrent memory.

J.5 TRAINING OBJECTIVE AND INFERENCE

TRUE-Ref is trained with the joint objective of Eq. 2,

$$\mathcal{L} = \lambda_{\mathrm{act}}\, \|a_t - \hat{a}_t\|_1 + \lambda_{\mathrm{reason}}\, \mathcal{L}_{\mathrm{reason}},$$

with $\lambda_{\mathrm{act}} = \lambda_{\mathrm{reason}} = 1.0$. The first term is OpenVLA-OFT's L1 action loss. $\mathcal{L}_{\mathrm{reason}}$ is the causal next-token loss over the reasoning tokens, applied when $z = 1$. Under causal masking, each reasoning token is predicted only from (o_t, l, h_t) and the preceding reasoning tokens, so the pathway learns to infer progress, sub-goals, and action consequences from inputs that are available at test time, and no token is visible to itself as input. The probability p sets the reasoning coverage: $p = 0$ recovers history-only training ("w/o reasoning"), and the "w/o history" ablation removes $f_\phi(h_t)$ from X_t. All ablations share the same objective, optimizer, and training budget (Table 8).

At deployment, TRUE-Ref receives only (o_t, l, h_t), in compliance with the benchmark interface. It predicts the next action chunk with the parallel action head and can optionally decode reasoning $\hat{r}_t$ from the same representation, which is the output evaluated on the reasoning track. Decoded reasoning is not fed back into the action head, so it adds no latency to control, and the external annotation generator is never called.