

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 CHAIN OF INTERACTION BENCHMARK (COIN): WHEN REASONING MEETS EMBODIED INTERACTION

Anonymous authors

Paper under double-blind review

ABSTRACT

Generalist embodied agents must perform interactive, causally-dependent reasoning, continually interacting with the environment, acquiring information, and updating plans to solve long-horizon tasks before they could be adopted in real-life scenarios. For instance, retrieving an apple from a cabinet may require opening multiple doors and drawers before the apple becomes visible and reachable—demanding sequential interaction under partial observability. However, existing benchmarks fail to systematically evaluate this essential capability. We introduce **COIN**, a benchmark designed to assess interactive reasoning in realistic robotic manipulation through three key contributions. First, we construct **COIN-50**: 50 interactive tasks in daily scenarios, and create **COIN-Primitive** required by causally-dependent tasks, and **COIN-Composition** with mid-term complexity for skill learning and generalization evaluation. Second, we develop a low-cost mobile AR teleoperation system and collect the COIN-Primitive Dataset with 50 demonstrations per primitive task (1,000 in total). Third, we develop systematic evaluation metrics about execution stability and generalization robustness to evaluate **CodeAsPolicy**, **VLA**, and language-conditioned **H-VLA** approaches. Our comprehensive evaluation reveals critical limitations in current methods: models struggle with interactive reasoning tasks due to significant gaps between visual understanding and motor execution. We provide fine-grained analysis of these limitations.

1 INTRODUCTION

Recent advances in large-scale pretraining NVIDIA et al. (2025b); Black et al.; Brohan et al. (b) and the creation of diverse datasets O’Neill et al. (2024); Khazatsky et al. (2024) and benchmarks Zhang et al.; Li et al. (2024); Liu et al. have significantly advanced robotic manipulation capabilities. However, current benchmarks primarily focus on simplified tasks that fail to capture the complexity of real-world manipulation challenges, particularly those requiring interaction and causal reasoning over long time horizons in partially observable environments.

Consider a robot tasked with "open a locked door". This seemingly simple instruction requires a sequence of interdependent actions: locating the keyhole, inserting and turning the key, and then rotating the handle with trials for the right directions. Such tasks demand what we term **interactive reasoning**—the ability to continually interact with the environment, gather information, update beliefs, and adapt plans accordingly. This requires multiple capabilities: perceiving partial environmental states, reasoning about causal dependencies between actions, maintaining memory of previous interactions, and dynamically adjusting strategies based on feedback. This capability remains beyond the reach of most current Vision-Language-Action (VLA) models and VLM-based planning approaches.

To address this gap, we introduce **COIN** (Chain Of INteraction) Benchmark, consisting of three complementary components: **COIN-50**, featuring 50 interactive reasoning tasks grounded in everyday activities (with one demonstration per task); **COIN-Primitive**, containing 20 fundamental manipulation skills that serve as building blocks (with approximately 50 trajectories per task); and **COIN-Composition**, bridging **COIN-Primitive** and **COIN-50** for evaluating the robustness of VLA learning across visual understanding and instruction variations. Unlike previous benchmarks that primarily test perception or simple manipulation, our tasks are systematically organized according to a taxonomy of reasoning capabilities required in partially observable environments. Based on

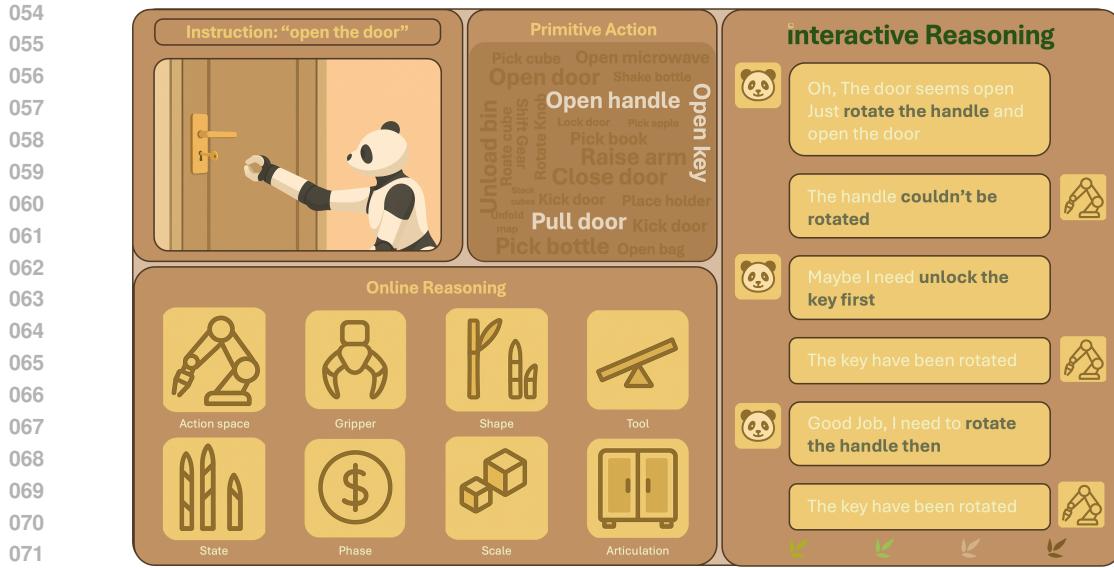


Figure 1: An illustration of **COIN**. Our benchmark focuses on evaluating the crucial **interactive reasoning** ability of Vision-Language-Action (VLA) models and VLM-based robotic planning systems, covering both rich reasoning knowledge and diverse primitive actions.

our analysis, we categorize these capabilities into three principal domains: (1) **Object-Centric Reasoning**, encompassing physical property inference, spatial reasoning, mechanism understanding, and visual reasoning; (2) **Robot-Centric Reasoning**, covering control optimization and embodiment awareness (such as collision handling); and (3) **Compositional Reasoning**, including tool-mediated problem solving, failure-driven adaptation, hierarchical planning, and experience utilization. These capabilities, essential for robots to function effectively in human environments, remain underexplored in existing benchmarks.

To support algorithm development and evaluation, we created a low-cost, phone-based teleoperation system (hardware cost under \$20 according to second-hand websites in China) inspired by Rayyan (2024). Using this system, we collected the COIN-Primitive Dataset—over 1,000 expert demonstrations across 20 fundamental manipulation skills recorded from multiple viewpoints. These primitives serve as essential building blocks for VLA model fine-tuning and compositional task solving. **Our contributions include:**

1. **COIN Benchmark:** We construct **COIN-50** with 50 interactive tasks in daily scenarios, **COIN-Primitive** with 20 fundamental manipulation skills required by causally-dependent tasks, and **COIN-Composition** with mid-term complexity for skill learning and generalization evaluation, systematically organized according to a principled taxonomy of interactive reasoning capabilities.
2. **Low-Cost Mobile AR Teleoperation System and Dataset:** We develop a smartphone-based teleoperation system (hardware cost under \$20 according to second-hand websites in China) and collect the COIN-Primitive Dataset with 50 demonstrations per primitive task (1,000 in total), enabling accessible data collection for the robotics community.
3. **Systematic Evaluation Metrics and Analysis:** We develop comprehensive evaluation metrics about execution stability and generalization robustness to evaluate **CodeAsPolicy**, **VLA**, and **H-VLA** approaches, revealing critical limitations including significant gaps between visual understanding and motor execution, and provide fine-grained analysis for these limitations.

2 RELATED WORKS

2.1 VISION-LANGUAGE-ACTION MODELS AND APPROACHES

CodeAsPolicy approaches Liang et al. (2022) combine VLMs with predefined skills to orchestrate perception modules Kirillov et al. (2023; 2024); Yang et al. (2023) and low-level controllers in a

Benchmark	Tasks	Demos	Avg. Steps	Cont. Action	Caus. Dep.	Visual Comp.	Inter. Reas.	Visual Unobs.	Mech. Unobs.
Robot Manipulation Benchmarks									
CALVIN Mees et al.	34	200K	30	✓	✗	✓	✗	✗	✗
Arnold Gong et al. (2023)	8	40	125.8	✓	✗	✗	✗	✗	✗
SimplerEnv Zhu et al.	10	N/A	52.3	✓	✗	✗	✗	✗	✗
Libero Liu et al.	130	50/task	77.3	✓	✓	✓	✗	✗	✗
VLABench Zhang et al.	100	163	157.2	✓	✗	✓	✗	✗	✗
RoboCASA Zheng et al.	100	100/task	371.9	✓	✓	✓	✗	✓	✓
EmbodiedBench Yang et al.	100	N/A	N/A	✗	✗	✓	✗	✓	✗
RoboVerse Geng et al. (2025)	1000	9331	N/A	✓	✓	✓	✗	✓	✓
Vision-Language Reasoning Benchmarks									
VLMbench Li et al. (b)	100	N/A	N/A	✗	✓	✓	✗	✗	✗
ClevrSkills Haresh et al.	12	N/A	N/A	✗	✓	✓	✗	✗	✗
ReflectVLM Feng et al.	50	N/A	N/A	✗	✓	✓	✓	✗	✗
COIN (Ours)	90	1000+	988.9	✓	✓	✓	✓	✓	✓

Table 1: Comprehensive benchmark comparison including quantitative metrics and reasoning capabilities. COIN demonstrates the longest average trajectory length (988.9 steps) and uniquely combines all critical reasoning capabilities, particularly interactive reasoning. Our systematic evaluation framework with 1000+ demonstrations across 90 tasks provides unprecedented depth for analyzing interactive manipulation.

modular, zero-shot framework. Works like Huang et al. (a;c;b) excel in generalization but struggle with online adaptation due to their “plan-then-execute” paradigm, where VLMs disengage after initial planning. While recent work Duan et al. introduces replanning mechanisms, significant challenges remain in dynamic, partially observable scenarios requiring continuous interactive reasoning.

End-to-End VLA models Brohan et al. (b;a); NVIDIA et al. (2025b); Li et al. (a) directly map visual observations and language to robotic actions via token prediction, learning policies through imitation. Their unified architecture enables emergent reasoning through large-scale pretraining. Despite success in basic manipulation tasks, these models struggle with long-horizon scenarios requiring state maintenance and adaptive planning over extended interactions.

Hierarchical VLA (H-VLA) architectures Figure AI (2025); Team et al. (2025) bridge planning and execution by decomposing high-level instructions into subtasks coordinated with low-level executors. This approach combines explicit reasoning with learned behaviors, showing promise in complex manipulation tasks and advancing toward more generalist robotic systems.

2.2 ROBOT MANIPULATION BENCHMARKS

Robot manipulation benchmarks excel in physical interaction and control capabilities. Works like Arnold Gong et al. (2023) and SimplerEnv Zhu et al. offer photorealistic simulation but lack reasoning components. Libero Liu et al. and RoboCASA Zheng et al. incorporate partial observability, but most lack the combination of dynamic interaction, failure recovery, and interactive reasoning needed for realistic scenarios. Table 1 shows that these benchmarks do not cover interactive reasoning well, while our benchmark emphasizes this crucial capability of embodied AI.

Vision-language embodied reasoning benchmarks prioritize reasoning over physical realism. VLMbench Li et al. (b) and ClevrSkills Haresh et al. support causal reasoning in simplified environments, while ReflectVLM Feng et al. offers failure recovery but limited physical interaction. COIN uniquely bridges this gap by combining all eight critical dimensions shown in Table 1, enabling evaluation of true interactive reasoning in realistic, partially observable environments.

3 COIN: CHAIN OF INTERACTION BENCHMARK

In this section, we introduce: the formulation of tasks in COIN (3.1), how we built such tasks in COIN (3.2), how we collected datasets with human-in-the-loop teleoperation (3.3), the statistics of COIN (3.4), and the evaluation metrics (3.5).

162 3.1 TASKS FORMULATION
163

164 We formalize interactive reasoning tasks as a Partially Observable Markov Decision Process
165 (POMDP): $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{O}, \mathcal{Z} \rangle$. The state space \mathcal{S} encompasses robot configuration,
166 object states and physical properties. We implement two action spaces, the same as ManiSkill3: for
167 VLA models, $A_{VLA} = \{\Delta p, \Delta R, g\} \in \mathbb{R}^3 \times SO(3) \times \{0, 1\}$ using delta end-effector poses; for
168 CodeAsPolicy approaches, $A_{VLM} = \{q_1 \dots q_7, g\} \in [q_{\min}, q_{\max}]^7 \times \{0, 1\}$ using absolute joint positions.
169 The transition function \mathcal{T} models state dynamics while the reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \{0, 1\}$
170 provides sparse binary success feedback. Observations \mathcal{O} include five camera views (front, left/right
171 back, left front, and wrist-mounted) with depth and segmentation maps, language instructions for
172 the task, and robot proprioceptive data, enabling agents to infer occluded state information through
173 interaction. More details can be found in Appendix C.

174 3.2 TASK BUILDING
175

176 **COIN comprises 3 categories and 90 total tasks.** We design a hierarchical task structure that
177 systematically evaluates interactive reasoning capabilities across different complexity levels:

- 179 • **COIN-Primitive (20 tasks):** Fundamental manipulation skills extracted from interactive reasoning
180 tasks by identifying commonly recurring behavioral patterns and essential manipulation primitives
181 (open-close, pick-place, push-pull, rotation).
- 182 • **COIN-Composition (20 tasks):** Mid-term complexity tasks that bridge the gap between primitives
183 and full interactive reasoning, introducing controlled increases in complexity through small visual
184 differences or instruction variations.
- 185 • **COIN-50 (50 tasks):** Full interactive reasoning tasks requiring multi-step causal reasoning under
186 partial observability, where agents must continually interact with the environment to gather
187 information and adapt their strategies.

188 As shown in Figure 2 and Appendix H.2, we categorized these interactive tasks into three main
189 domains: (1) object information perception and manipulation, (2) robot-understanding and control,
190 and (3) compositional reasoning. This taxonomy helps systematically evaluate different aspects of
191 interactive reasoning capabilities in embodied agents.

192 **Technical implementation.** We built all environments on the ManiSkill3 platform Tao et al., using
193 the Franka Emika Panda robot in tabletop manipulation settings. Environmental assets include
194 articulated objects from PartNet-Mobility Xiang et al. (2020) and additional assets from Zeng et al.;
195 Li et al. (b); Sketchfab. All 90 tasks provide language instructions and corresponding reward.

196 **Subtasks and VQA** For each interactive task, we provide: (1) expert demonstrations, (2) ground truth
197 planning in the form of decomposed subtask sequences (*oracle manipulation flow*). As illustrated in
198 Figure 2, the average subtask length is 2.83, highlighting the multi-stage nature of these tasks. (3)
199 VQA (Visual Question-Answering) evaluations. The VQA is similar to ERQA protocol NVIDIA
200 et al. (2025a); Embodied Reasoning (2024) queries VLMs with task-specific questions about success
201 conditions or interaction history, serving as an embodied reasoning probe for future research. We
202 formulated them as multiple-choice problems, where VLMs answer these questions by selecting the
203 right answer.

204 3.3 TELEOPERATION AND DATA COLLECTION
205

206 **Low-cost mobile AR teleoperation system.** We introduce COIN-teleoperation, a smartphone-
207 based teleoperation system with a total hardware cost under \$20. Built on ARKit Inc. (2023) and
208 ARCore LLC (2023) with the help of Rayyan (2024), this system captures 6-DoF pose data from
209 mobile devices and achieves stable 20Hz control frequency even on older phones (e.g., iPhone 7 Plus),
210 making robotic data collection broadly accessible. Our comprehensive validation G demonstrates
211 90% data replay success and cross-device compatibility, confirming the reliability of our collection
212 approach.

213 **COIN-Primitive Dataset.** We collected a comprehensive dataset of 20 **COIN-Primitive** tasks as
214 mentioned in with 50 demonstrations per task captured from 5 camera viewpoints, totaling 1,000
215 trajectories using COIN-teleoperation. This dataset serves as the primary training resource for VLA

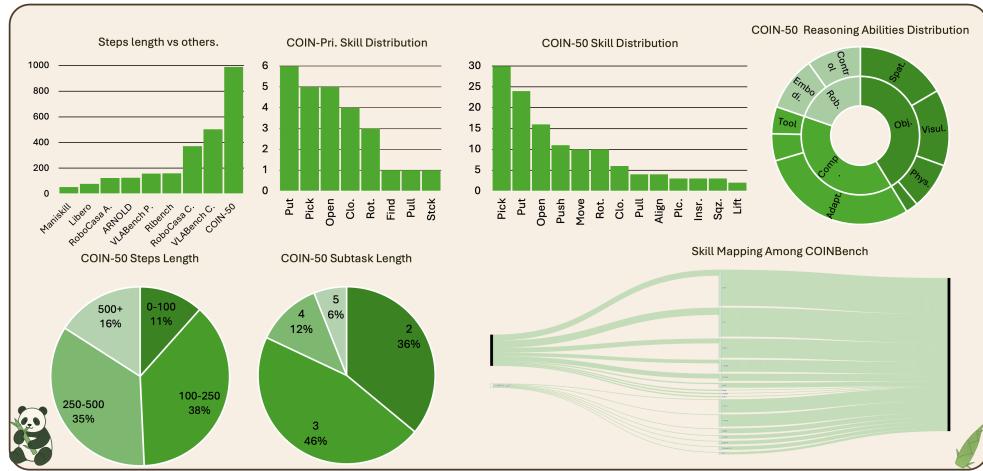


Figure 2: Tasks in COIN: we provide diverse tasks with feasible primitive tasks, and provide GT planning for these tasks, which could be used to guide the planning

model fine-tuning, providing diverse manipulation primitives that form the building blocks for more complex interactive reasoning tasks.

3.4 COIN STATISTICS

Statistics. Figure 2 presents a comprehensive overview of COIN’s benchmark structure. COIN-50 features an average task length of approximately 990 steps, substantially longer than existing benchmarks. More critically, each task requires an average of 2.83 subtasks with frequent interactive reasoning cycles, where 36% of tasks contain 2 subtasks, 46% contain 3, and 12% contain 4. This reveals that our benchmark poses greater challenges not merely through temporal extension, but through the density of reasoning interactions required—necessitating iterative “interaction-reasoning-interaction” loops rather than simple sequential execution, fundamentally distinguishing interactive reasoning from purely long-horizon tasks.

The benchmark’s reasoning taxonomy spans object-centric, robot-centric, and compositional reasoning. This focus addresses the under-representation of interactive reasoning in prior benchmarks and supports the modeling of complex “interaction-reasoning-interaction” loops. Overall, COIN offers a comprehensive and realistic testbed for assessing manipulation skills and reasoning capabilities under partial observability and task complexity.

3.5 EVALUATION METRICS

We introduce a comprehensive evaluation framework with six complementary metrics that assess different aspects of interactive reasoning and manipulation performance across all COIN tasks:

Task Performance Metrics:

- **Success Rate (SR)** measures the proportion of successfully completed trajectories across all evaluated tasks.
- **Class Success Rate (CSR)** measures category-specific performance across reasoning domains (object-centric, robot-centric, compositional).

Reasoning Assessment Metrics:

- **Visual Question Answering Score (VS)** assesses perceptual and reasoning capabilities by evaluating whether models correctly answer questions about environmental states and interactive consequences.

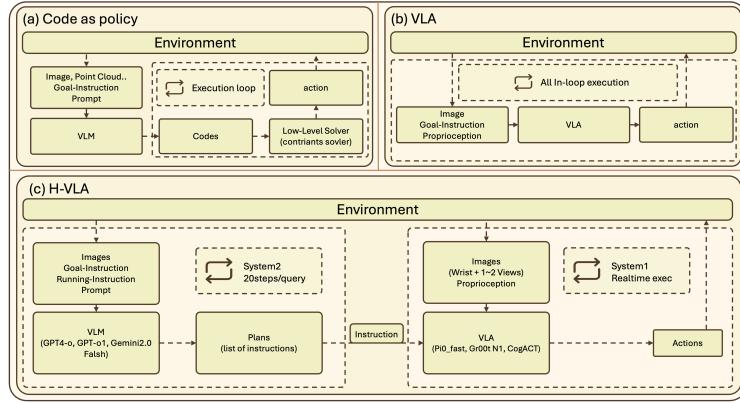


Figure 3: Model Architecture Comparison: **(a)** CodeAsPolicy uses VLMs for planning, with execution handled separately by low-level code and constraint optimizers. **(b)** End-to-End VLA performs in-loop perception and action directly from the environment. **(c)** Hierarchical VLA (H-VLA) combines high-level planning (System 2) with low-level VLA execution (System 1), connected via language instructions.

Fine-grained Execution Quality Metrics:

- **Trajectory Stability Score (TS)** measures action quality and smoothness to identify erratic VLA behaviors:

$$TS = 0.3 \cdot S_{vel} + 0.3 \cdot S_{acc} + 0.4 \cdot S_{jerk}$$

where each component uses $\text{Smooth}(x) = \exp(-CV_x)$ with $CV_x = \frac{\sigma_x}{\mu_x}$ (coefficient of variation) applied to velocity, acceleration, jerk (3rd derivative), and position respectively. Higher scores indicate better trajectory stability.

- **Gripper Control Stability (GS)** assesses manipulation quality through coordination analysis:

$$GS = 0.4 \cdot S_{smooth} + 0.3 \cdot S_{freq} + 0.3 \cdot S_{coord}$$

where $S_{smooth} = \exp(-\text{abrupt changes})$ penalizes sudden gripper state transitions, $S_{freq} = \exp(-\frac{N_{changes}}{N_{expected}})$ evaluates action frequency appropriateness, and S_{coord} analyzes arm-gripper coordination timing. Higher scores indicate better gripper control quality.

- **Generalization Capability Score (GCS)** evaluates model adaptability through controlled task variations:

$$GCS = \frac{SR_{composition}}{SR_{primitive}}$$

where success rates are averaged across all tasks in each category. Scores close to 1.0 indicate good generalization; lower scores reveal generalization failures.

These metrics provide comprehensive evaluation across task completion, reasoning understanding, execution quality, and generalization capability, enabling detailed analysis of model performance across all COIN benchmark components.

3.6 HIERARCHICAL VLA (H-VLA) ARCHITECTURE FOR COIN

Similar to Helix Figure AI (2025), we propose a two-layered VLA framework that decomposes complex reasoning tasks into manageable skill sequences, illustrated in Figure 3(c).

- **System 2 (High-level planner)**, a VLM that processes multi-view images and task instructions to generate a sequence of sub-tasks. Operating at fixed intervals, it monitors execution progress by periodically evaluating current observations and adjusting the instruction queue accordingly.
- **System 1 (Low-level executor)**, a VLA model that converts individual skill instructions into robot actions. Taking images, proprioception data, and the current instruction as input, it generates actions in real-time without knowledge of the overall task plan.

324 **4 EXPERIMENTS**

325

326 In this section, we evaluate model performance across different task sets. Section 4.1 introduces
 327 our experiment setup and the models tested on COIN (see Figure 3 for an overview). We first
 328 analyze how H-VLA and CodeAsPolicy perform on the complex COIN-50 tasks in Section 4.2. Since
 329 most models struggle with COIN-50, we then examine their abilities on basic manipulation tasks in
 330 COIN-Primitive and COIN-Composition to better understand the causes of failure (Section 4.3).

331

332 **4.1 EXPERIMENTAL SETUP**

333

334 **Models for COIN-50.** COIN-50’s complex interactive reasoning tasks require models capable of
 335 adaptive planning and execution. We evaluate:

336 **• H-VLA models.** As described in Section 3.6, this two-tier architecture combines VLMs for high-
 337 level planning with VLAs for execution. We evaluate six configurations pairing two high-level
 338 planners (**GPT-4o** and **Gemini 2.0 Flash**) with three VLA models (**Gr00t N1**, **Pi0**, and **CogACT**).
 339 Unlike end-to-end VLAs, H-VLA can update plans during execution as new information becomes
 340 available through interaction.

341 **• CodeAsPolicy approaches.** We implement two code-based planning systems: **Voxposer** and
 342 **Rekep**, both using `gpt-4o-2024-11-20` for task decomposition and execution planning. Each
 343 system reconstructs the environment from three camera views, with Voxposer additionally utilizing
 344 ground truth object lists to enhance scene understanding. These approaches separate perception
 345 and planning from execution through programmatic interfaces.

346

347 **Models for COIN-Primitive and COIN-Composition.** We only consider the “low-level controller”
 348 of the two families of models above for COIN-50. Effectively, these are end-to-end VLA models as
 349 in H-VLA models (see Section 3.6) and CodeAsPolicy itself, which is the same model on different
 350 benchmarks.

351 **• End-to-end VLA models.** We evaluate 3 cutting-edge vision-language-action models as adopted
 352 in H-VLA above: **Gr00t N1** NVIDIA et al. (2025b), **Pi0** Black et al., and **CogACT** Li et al. (a).
 353 Both Gr00t N1 and Pi0 process multi-view observations from three cameras (base-front, left-front,
 354 and wrist-mounted), while CogACT processes only the left-front view per its design requirements.
 355 We fine-tune all VLA models on the COIN-Primitive dataset until convergence or for a maximum
 356 of three days (see Appendix F for details). Following Liu et al., we select checkpoints based on
 357 validation success rates.

358 **• CodeAsPolicy approaches.** We evaluate the same Voxposer and Rekep implementations on
 359 COIN-Primitive tasks to assess their performance on fundamental manipulation skills.

360

361 **Evaluation Details.** We report **SR** averaged over 10 trials. **CSR** is generated from the **SR**, and the
 362 VQA score is generated by querying the VLM with expert demonstrations for about 50 steps per query.
 363 For **TS** and **GS**, we report the scores according to the recorded trajectories during evaluation. For
 364 **GCS**, we evaluate the score according to the **SR** between **COIN-Primitive** and **COIN-Composition**.
 365 All tasks and environment specifications can be found in Appendix C and H.2.

366

367 **4.2 MAIN RESULTS FOR COIN-50**

368

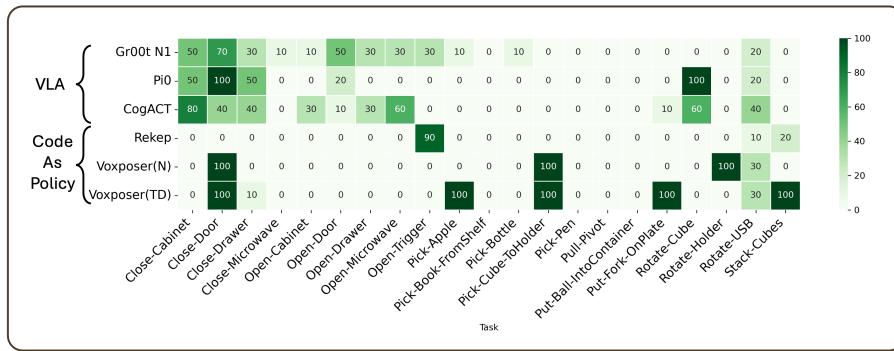
369 **Overview: Interactive reasoning remains a fundamental challenge for current AI approaches.**
 370 Our evaluation reveals a stark capability gap in all tested systems when faced with tasks requiring
 371 interactive reasoning. As shown in Table 4, all models fail to solve complex interactive reasoning
 372 tasks, with success rates rarely exceeding 3%. Our analysis on COIN-50 reveals fundamental
 373 limitations in both major approach categories:

374 **CodeAsPolicy approaches face two critical issues:** (1) *Non-interactive planning architecture*:
 375 These methods cannot update plans based on environmental feedback, making them fundamentally
 376 unsuited for partially observable environments requiring iterative interaction. For example, in “Pick
 377 the cube” task, if the cube was not picked, it only repeat “back-home” and “pick the cube” loop,
 378 without any new strategies. (2) *Planning-execution gap*: Significant disconnects exist as shown in

378

379
380
381
Table 2: Trajectory and gripper stability analysis across different task types. Values show mean \pm
standard deviation, revealing execution quality patterns across model architectures. **Bold values**
indicate performance exceeding human baseline.

Model	Task Type	Trajectory Stability	Gripper Stability
CogACT	Primitive	0.150 \pm 0.055	0.872 \pm 0.134
CogACT	Composition	0.138 \pm 0.039	0.796 \pm 0.136
CogACT	Interactive	0.146 \pm 0.041	0.782 \pm 0.141
Gr00t N1	Primitive	0.082 \pm 0.015	0.318 \pm 0.116
Gr00t N1	Composition	0.086 \pm 0.002	0.327 \pm 0.058
Gr00t N1	Interactive	0.084 \pm 0.002	0.294 \pm 0.050
Pi0	Primitive	0.084 \pm 0.067	0.440 \pm 0.198
Pi0	Composition	0.035 \pm 0.043	0.465 \pm 0.253
Pi0	Interactive	0.061 \pm 0.050	0.440 \pm 0.219
Human Dataset	Primitive	0.134 \pm 0.035	0.684 \pm 0.297

405
406
407
408
Figure 4: Performance heatmap for different models on COIN-Primitive tasks. The visualization
reveals that VLA models achieve broader task coverage than CodeAsPolicy approaches, though with
different strengths across task types. Color intensity indicates success rate.409
410
411
Huang et al. (b) between high-level plans and low-level execution capabilities, which we analyze in
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1188
1189
1190
1191
1192
1193
1194
1195
1195
1196
1197
1198
1199
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1288
1289
1290
1291
1292
1293
1294
1295
1295
1296
1297
1298
1299
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1488
1489
1490
1491
1492
1493
1494
1495
1495
1496
1497
1498
1499
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1588
1589
1590
1591
1592
1593
1594
1595
1595
1596
1597
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1688
1689
1690
1691
1692
1693
1694
1695
1695
1696
1697
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1788
1789
1790
1791
1792
1793
1794
1795
1795
1796
1797
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1888
1889
1890
1891
1892
1893
1894
1895
1895
1896
1897
1898
1899
189

432
 433 Table 3: Generalization capability evaluation using COIN-Composition tasks. Models demonstrate
 434 severe generalization failures when faced with minor visual or instruction variations from primitive
 435 tasks.

436 Model	437 Primitive SR	438 Composition SR	439 Finished Tasks	440 GCS
437 CogACT	438 19.0%	439 1.5%	440 3/20	441 0.079
437 Pi0	438 16.1%	439 6.5%	440 4/20	441 0.404
437 Gr00t N1	438 16.7%	439 0.0%	440 0/20	441 0.000

442 and Rekep perform poorly even on basic manipulation tasks, indicating fundamental misalignment
 443 between planning and execution. (2) Articulation Manipulation problems: These two models are not
 444 able to handle the articulated objects, such as cabinet, doors and switchs. This is mainly caused by
 445 the structure is not feasible for key-points based representation.

446 **H-VLA approaches confirm the three limitations identified in COIN-50.** Our detailed analysis
 447 reveals specific manifestations of the issues identified earlier: (1) *Poor trajectory and gripper control*:
 448 VLAs exhibit severe control precision issues, particularly struggling with gripper timing NVIDIA
 449 et al. (2025b). As shown in Table 2, CogACT demonstrates relatively stable trajectories compared to
 450 other VLAs, potentially due to its temporal ensemble mechanism, which is left for discussion in future
 451 work. However, all VLAs show significant jerky movements and high discontinuity. (2) *Catastrophic
 452 generalization failures*: As shown in Table 3, models achieving reasonable success on primitive tasks
 453 (16-19%) experience complete failure when faced with composition tasks. Even adding a single new
 454 object or switching instructions causes task failure. (3) *Weak VLM-VLA integration*: Despite broader
 455 task coverage when overfitted to primitive tasks, the coordination between high-level planning and
 456 low-level execution remains fundamentally problematic. For example, while VLAs can successfully
 457 execute "open the door" commands, changing the instruction to "pull the door" for the same physical
 458 action results in dramatically reduced success rates and moves the gripper to unreasonable locations,
 459 demonstrating that the natural language interface fails to capture the underlying action semantics.

461 5 CONCLUSIONS

462 We present COIN, a systematic evaluation benchmark for interactive reasoning in embodied AI
 463 that encompasses three hierarchical evaluation levels: fundamental skill learning (COIN-Primitive),
 464 intermediate capability testing (COIN-Composition), and critical interactive reasoning assessment
 465 (COIN-50) providing 50 interactive tasks in partially observable settings. Additionally, our COIN-
 466 teleoperation pipeline contributes a dataset of 1,000 demonstration trajectories for model training.

467 Through multi-layered evaluation metrics, our comprehensive analysis reveals fundamental limitations
 468 in current EAI approaches, particularly in generalization and adaptability. We provide in-depth
 469 analysis of these critical issues across both CodeAsPolicy and H-VLA paradigms. While achieving
 470 better interactive reasoning capabilities remains a significant challenge, our findings highlight several
 471 promising research directions worthy of deeper investigation:

472 **Promising Research Directions.** Based on our comprehensive analysis, we identify four critical
 473 research directions:

474 (1) **Improve VLA Trajectory Smoothness:** Our findings suggest that adaptive ensemble mecha-
 475 nisms, as potentially employed in CogACT, may contribute to more stable trajectory control compared
 476 to Pi0 and Gr00t.(Table 2). (2) **Enhance Multimodal Perception Ability:** Improving VLA visual
 477 generalization and instruction-following capabilities through better multimodal perception could
 478 enable more effective VLM-VLA coordination.(Section E.1) (3) **VLM-VLA Integration Mech-
 479 anisms:** The result showed that H-VLA with natura language is facing great problems according
 480 to (Table 3) and Li (2025), showing the necessity to improve the integration between high-level
 481 planning and low-level execution, models like Figure AI (2025) might bring more help. (4) **Adaptive
 482 CodeAsPolicy Frameworks:** CodeAsPolicy approaches now perform poorly on interactive and
 483 primitive tasks, we should adopt adaptive feedback mechanisms to achieve more robust control in
 484 dynamic environments (Section 4.2), moving beyond static "plan-then-execute" paradigms.

486

6 ETHICS STATEMENT

488 Our research presents no significant ethical risks and does not negatively impact human welfare. On
 489 the contrary, COIN contributes to the advancement of robotic applications that can assist humans in
 490 daily tasks and industrial automation. The benchmark focuses on fundamental manipulation skills in
 491 controlled environments, promoting safer and more reliable robotic systems. All experimental data
 492 was collected in laboratory settings without involving human subjects or sensitive information.

494

7 REPRODUCIBILITY STATEMENT

496 To ensure full reproducibility of our results, we commit to open-sourcing all components of our
 497 work: (1) Complete source code for COIN benchmark implementation, evaluation metrics, and
 498 baseline models will be made publicly available at [<https://anonymous.4open.science/r/coin-EB1B/>];
 499 (2) The full COIN dataset including 1,000+ demonstration trajectories, task specifications, and
 500 environment configurations will be released alongside the code; (3) Detailed experimental protocols,
 501 hyperparameter settings, and computational requirements are documented in the appendix to facilitate
 502 replication of our findings.

504

REFERENCES

506 Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo
 507 Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming
 508 Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang
 509 Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. II0: A
 510 Vision-Language-Action Flow Model for General Robot Control.

511 Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski,
 512 Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gon-
 513 zalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jas-
 514 mine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang,
 515 Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor
 516 Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag
 517 Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricu, Huong Tran, Vincent Van-
 518 houcke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao,
 519 Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models
 520 transfer web knowledge to robotic control, a. URL <http://arxiv.org/abs/2307.15818>.

521 Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn,
 522 Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian
 523 Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalash-
 524 nikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deek-
 525 sha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez,
 526 Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi,
 527 Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vin-
 528 cent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu,
 529 and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale, b. URL
 530 <http://arxiv.org/abs/2212.06817>.

531 Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay
 532 Krishna. Manipulate-anything: Automating real-world robots using vision-language models. URL
 533 <http://arxiv.org/abs/2406.18915>.

535 Embodied Reasoning. Erqa: Embodied reasoning question answering benchmark. <https://github.com/embodiedreasoning/ERQA>, 2024. Accessed: 2025-05-07.

538 Yunhai Feng, Jiaming Han, Zhuoran Yang, Xiangyu Yue, Sergey Levine, and Jianlan Luo. Reflective
 539 planning: Vision-language models for multi-stage long-horizon robotic manipulation. URL
<http://arxiv.org/abs/2502.16707>.

540 Figure AI. Helix: A vision-language-action model for generalist humanoid control. <https://www.figure.ai/news/helix>, February 2025. Accessed: 2025-04-27.
541
542

543 Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue
544 Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, Yutong Liang, Dylan Goetting, Chaoyi Xu,
545 Haozhe Chen, Yuxi Qian, Yiran Geng, Jiageng Mao, Weikang Wan, Mingtong Zhang, Jiangran
546 Lyu, Siheng Zhao, Jiazhao Zhang, Jialiang Zhang, Chengyang Zhao, Haoran Lu, Yufei Ding,
547 Ran Gong, Yuran Wang, Yuxuan Kuang, Ruihai Wu, Baoxiong Jia, Carlo Sferrazza, Hao Dong,
548 Siyuan Huang, Yue Wang, Jitendra Malik, and Pieter Abbeel. Roboverse: Towards a unified
549 platform, dataset and benchmark for scalable and generalizable robot learning, 2025. URL
550 <https://arxiv.org/abs/2504.18904>.
551

551 Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai,
552 Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. Arnold: A
553 benchmark for language-grounded task learning with continuous states in realistic 3d scenes, 2023.
554 URL <https://arxiv.org/abs/2304.04321>.
555

555 Sanjay Haresh, Daniel Dijkman, Apratim Bhattacharyya, and Roland Memisevic. ClevrSkills:
556 Compositional language and visual reasoning in robotics. URL <https://openreview.net/forum?id=64sZtFSOh6#discussion>.
557

558 Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. CoPa: General robotic
559 manipulation through spatial constraints of parts with foundation models, a. URL <http://arxiv.org/abs/2403.08248>.
560

561 Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. ReKep: Spatio-temporal
562 reasoning of relational keypoint constraints for robotic manipulation, b. URL <http://arxiv.org/abs/2409.01652>.
563

564 Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser:
565 Composable 3d value maps for robotic manipulation with language models, c. URL <http://arxiv.org/abs/2307.05973>.
566

566 Apple Inc. Arkit - apple developer. 2023. URL <https://developer.apple.com/augmented-reality/arkit/>. Augmented reality framework for iOS.
567

567

568 Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth
569 Karamchetti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis,
570 et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*,
571 2024.
572

572 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
573 Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. 2023. URL
574 <https://arxiv.org/abs/2304.02643>.
575

575 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
576 Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything model 2.
577 2024. URL <https://arxiv.org/abs/2404.14192>.
578

578 Chengshu Li et al. Evaluating real-world robot manipulation policies in simulation. 2024. URL
579 <https://arxiv.org/abs/2405.05941>.
580

580 Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng,
581 Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen,
582 Yuanchun Shi, Jiaolong Yang, and Baining Guo. CogACT: A foundational vision-language-
583 action model for synergizing cognition and action in robotic manipulation, a. URL <http://arxiv.org/abs/2411.19650>.
584

584 Quanyi Li. Task Reconstruction and Extrapolation for π_0 using Text Latent, August 2025.
585

585 Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa
586 Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan
587 Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation, b. URL
588 <http://arxiv.org/abs/2405.05941>.
589

594 Percy Liang, Rishi Chen, Po-Hsun Huang, Nikhil Vatsal, Pieter Xu, Tete Zhou, Yuhui Zhou, Fei Xia,
 595 Karol Hausman, Brian Ichter, et al. Code as policies: Language model programs for embodied
 596 control. 2022. URL <https://arxiv.org/abs/2209.07753>.

597

598 Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO:
 599 Benchmarking knowledge transfer for lifelong robot learning. URL <http://arxiv.org/abs/2306.03310>.

600

601 Google LLC. Arcore - google developers. 2023. URL <https://developers.google.com/ar>. Augmented reality platform for Android.

602

603 Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark
 604 for language-conditioned policy learning for long-horizon robot manipulation tasks. URL <http://arxiv.org/abs/2112.03227>.

605

606 NVIDIA, :, Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu,
 607 Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, Jinwei Gu, Siddharth
 608 Gururani, Imad El Hanafi, Zekun Hao, Jacob Huffman, Jingyi Jin, Brendan Johnson, Rizwan Khan,
 609 George Kurian, Elena Lantz, Nayeon Lee, Zhaoshuo Li, Xuan Li, Tsung-Yi Lin, Yen-Chen Lin,
 610 Ming-Yu Liu, Alice Luo, Andrew Mathau, Yun Ni, Lindsey Pavao, Wei Ping, David W. Romero,
 611 Misha Smelyanskiy, Shuran Song, Lyne Tchapmi, Andrew Z. Wang, Boxin Wang, Haoxiang
 612 Wang, Fangyin Wei, Jiashu Xu, Yao Xu, Xiaodong Yang, Zhuolin Yang, Xiaohui Zeng, and Zhe
 613 Zhang. Cosmos-reason1: From physical common sense to embodied reasoning, 2025a. URL
 614 <https://arxiv.org/abs/2503.15558>.

615

616 NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding,
 617 Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang,
 618 Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith
 619 Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang
 620 Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzen Xu, Zhenjia
 621 Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke
 622 Zhu. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots, March 2025b.

623

624 Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham
 625 Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment:
 626 Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE
 627 International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

628

629 Omar Rayyan. MuJoCo AR: Phone Teleoperation for Robots, 2024. URL <https://github.com/omarrayyann/mujocoar>.

630

631 Sketchfab. Sketchfab - the leading platform for 3d & ar on the web. <https://sketchfab.com/>.
 Accessed: 2025-05-07.

632

633 Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, and Lin.
 634 MANISKILL3: GPU PARALLELIZED ROBOTICS SIMULATION AND RENDERING FOR
 635 GENERALIZABLE EMBODIED AI.

636

637 Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat
 638 Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza,
 639 Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann,
 640 Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique
 641 Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep
 642 Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny
 643 Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey,
 644 Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever,
 645 Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen,
 646 Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler,
 647 Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine
 Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath
 Maddineni, Anirudha Majumdar, Assaf Hurwitz Michael, Robert Moreno, Michael Neunert,
 648 Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao,

648 Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah,
 649 Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost To-
 650 bias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent
 651 Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan
 652 Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu,
 653 Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang,
 654 Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing ai into the physical
 655 world, 2025. URL <https://arxiv.org/abs/2503.20020>.

656 Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao
 657 Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment.
 658 In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.
 659 11097–11107, 2020.

660 Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark
 661 prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*,
 662 2023.

663 Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang,
 664 Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang.
 665 EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-
 666 driven embodied agents. (arXiv:2502.09560). doi: 10.48550/arXiv.2502.09560. URL [http://arxiv.org/abs/2502.09560](https://arxiv.org/abs/2502.09560).

667 Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian,
 668 Travis Armstrong, Ivan Krasin, Dan Duong, Ayzaan Wahid, Vikas Sindhwani, and Johnny Lee.
 669 Transporter networks: Rearranging the visual world for robotic manipulation. (arXiv:2010.14406).
 670 doi: 10.48550/arXiv.2010.14406. URL [http://arxiv.org/abs/2010.14406](https://arxiv.org/abs/2010.14406).

671 Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue
 672 Yin, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. VLABench: A large-scale benchmark for
 673 language-conditioned robotics manipulation with long-horizon reasoning tasks.

674 Liming Zheng, Feng Yan, Fanfan Liu, Chengjian Feng, Zhuoliang Kang, and Lin Ma. RoboCAS: A
 675 benchmark for robotic manipulation in complex object arrangement scenarios. (arXiv:2407.06951).
 676 doi: 10.48550/arXiv.2407.06951. URL [http://arxiv.org/abs/2407.06951](https://arxiv.org/abs/2407.06951).

677 Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martin-Martin, Abhishek Joshi, Soroush Nasiriany,
 678 and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning.

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

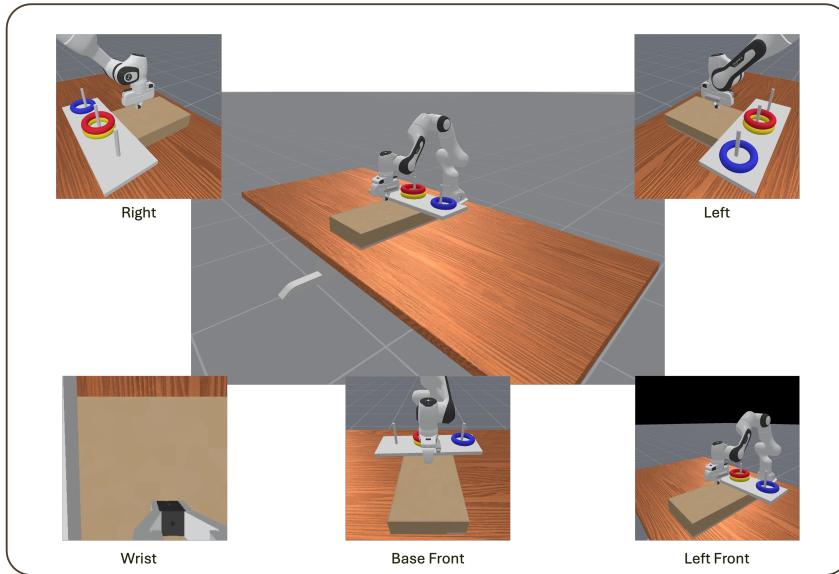


Figure 5: Environment Setup

A THE USAGE OF LLM

We acknowledge the use of Large Language Models (LLMs) in the preparation of this work in the following capacities:

Writing Assistance and Polishing: LLMs were employed to aid in refining the clarity and coherence of our manuscript. This includes improving sentence structure, enhancing readability, and ensuring consistent academic writing style throughout the paper. All technical content, experimental results, and scientific contributions remain entirely our own work.

B LIMITATIONS AND FUTURE WORK

Despite COIN’s comprehensive design, several limitations must be acknowledged: (1) our focus on a single robotic platform in static environments fails to capture the full complexity of dynamic real-world scenarios; (2) the absence of dual-arm manipulation tasks that could reveal additional coordination challenges in interactive reasoning.

For future work, we plan to pursue the promising research directions identified in our conclusions. Specifically, we aim to investigate: (1) trajectory smoothness mechanisms inspired by CogACT’s temporal ensemble approach to improve VLA execution stability; (2) enhanced multimodal perception architectures that better integrate visual understanding with instruction following; (3) novel VLM-VLA integration paradigms, comparing latent vector bridges against natural language interfaces; and (4) adaptive CodeAsPolicy frameworks incorporating closed-loop feedback for dynamic replanning. Additionally, we anticipate significant advancement in human-inspired learning approaches that enable iterative "think-execute-think" cycles, allowing models to formulate hypotheses, design informative tests, and recursively update their world models based on interactive outcomes.

C ENVIRONMENT SETUP

We use a 7-DoF Franka Emika Panda robotic arm equipped with a parallel gripper as our standard platform. As described in Section 3.1, we use different action spaces for VLA-based controllers and VLM-based planners. Specifically, for VLA models using Panda inverse kinematics, we define the action space as $\Delta p \in [-0.3, 0.3]$ for positional deltas and $\Delta R \in [-0.5, 0.5]$ for orientation deltas. The robot is observed from five camera perspectives providing comprehensive spatial and task

756
 757 Table 4: Performance comparison on COIN-50 interactive reasoning tasks. Human participants
 758 achieve 40% success rate via teleoperation (100% in real-world settings), while current AI approaches
 759 rarely exceed 3% success rates, revealing a substantial capability gap that highlights the significant
 760 challenges in achieving interactive reasoning.

761 Model/Human	762 Object-centric	763 Robot-centric	764 Compositional	765 Overall
766 Human (Sim/ 10)		N/A		40%
767 Human (Real/ 10)		N/A		100%
H-VLA				
768 Pi0 + Gemini 2.0	1.88%	2.14%	1.97%	1.99%
769 Pi0 + GPT-4o	1.96%	2.50%	2.05%	2.17%
770 Gr00t N1 + Gemini 2.0	1.74%	2.50%	1.82%	2.02%
771 Gr00t N1 + GPT-4o	1.52%	1.79%	1.59%	1.63%
772 CogACT + Gemini 2.0	2.14%	1.37%	2.24%	1.92%
773 CogACT + GPT-4o	1.74%	1.07%	1.82%	1.54%
CodeAsPolicy				
774 Voxposer(TD)	0.43%	0.00%	0.45%	0.29%
775 Voxposer(Normal)	2.17%	3.57%	2.27%	2.67%
776 Rekep	3.04%	3.57%	3.18%	3.26%

777 context for both VLA and VLM-based planning models. All environments are built on the ManiSkill3
 778 platform using physics-based simulation.

781 D PERFORMANCE COMPARISON: MODELS VS. HUMAN BASELINE ON 782 COIN-50

783 To contextualize the difficulty level of COIN tasks and establish a performance baseline, we evaluated
 784 both current AI systems and human performance on COIN-50 interactive reasoning tasks. For human
 785 evaluation, we recruited 3 participants with B.S. degrees who had no prior exposure to our tasks, with
 786 each participant attempting a representative subset of 10 tasks twice via teleoperation.

787 The results reveal a dramatic performance gap between human capabilities and current AI systems.
 788 While humans achieve 40% success rates in simulation (100% in real-world settings, confirming task
 789 feasibility), the best-performing AI model (Rekep) achieves only 3.26% overall success rate. This
 790 12-fold performance difference demonstrates that current approaches are fundamentally limited in
 791 their interactive reasoning capabilities, with substantial gaps that must be addressed before these
 792 systems can effectively operate in partially observable environments requiring adaptive manipulation
 793 strategies.

794 E VLM COMPARISON

795 E.1 VLM PLANNING COMPARISON

800 We evaluated VLM planners on COIN-50 tasks across execution time steps. GPT-4o consistently
 801 outperforms Gemini 2.0 by approximately 1.5 points on our reasoning scale, maintaining this
 802 advantage throughout task execution. While they could not improve the performance along the time
 803 goes on, indicating there might be some problem on models' historical information utilization.

804 E.2 VQA EVALUATION DETAILS

805 While the end-to-end task execution success rates are low, we evaluate the reasoning capabilities of
 806 different VLMs through our embodied VQA protocol. The results reveal that GPT-4o consistently
 807 outperforms Gemini-2.0-Flash, and all models show increased accuracy in the middle phases of

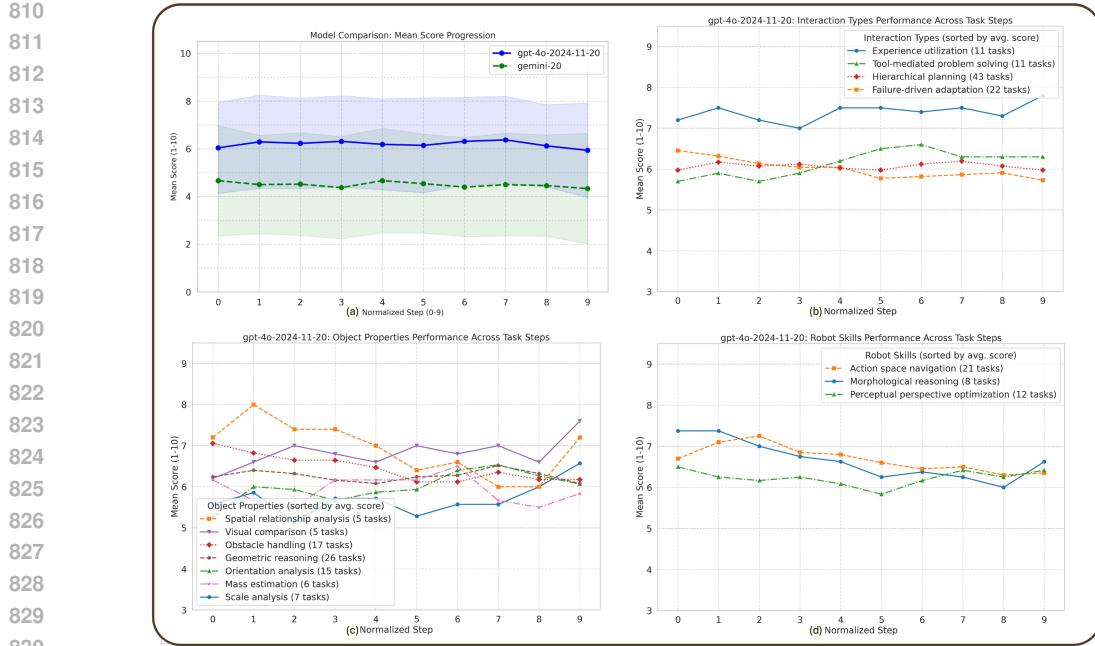


Figure 6: Comparison of VLM reasoning abilities on COIN-50 tasks evaluated along expert demonstration videos. GPT-4o consistently outperforms Gemini 2.0 across different reasoning categories and time steps.

task execution before slightly declining in the final phase. This pattern suggests models gradually accumulate task-relevant information through observation.

F MODEL TRAINING CONFIGURATIONS

We use the following training configurations: CogACT-Base trained on $4 \times$ A800 GPUs with device batch size 32 for 30K steps; Gr00t N1 2B trained on $4 \times$ A800 GPUs with device batch size 16 for 120K steps; Pi0-Fast trained on $3 \times$ A800 GPUs with device batch size 2 for 470K steps.

G COIN-TELEOPERATION ALGORITHM AND COMPARISON

Our AR teleoperation system demonstrates robust performance across multiple validation metrics:

Data Quality Validation: 90% of collected trajectories can be successfully replayed in ManiSkill3, indicating high fidelity data capture. Models trained on our data achieve consistent task performance, validating data quality.

Cross-Device Compatibility: The system works reliably across Android and iOS devices released after 2016, with stable 20Hz control frequency maintained even on older hardware (iPhone 7 Plus tested).

Comparison with Traditional Methods: Our approach offers significant advantages in accessibility (no specialized hardware required), scalability (easy deployment across environments), and cost-effectiveness (hardware cost under \$20 vs. thousands for traditional systems).

H MORE DETAILS ABOUT BENCHMARK

H.1 TASK DIVERSITY AND TEMPORAL ANALYSIS

COIN demonstrates unprecedented temporal complexity compared to existing benchmarks:

864 **Algorithm 1** COIN-teleoperation Pipeline

865 **Require:** Mobile device with sensors and AR framework

866 **Ensure:** Robot control commands

867 1: **function** MOBILEPHONEPROCESSING

868 2: $X_t^{IMU} \leftarrow$ IMU sensor readings at time t

869 3: $X_t^{gyro} \leftarrow$ Gyroscope readings at time t

870 4: $I_t \leftarrow$ Camera image at time t

871 5: $(p_t, R_t) \leftarrow$ AR framework (ARKit/ARCore) processing of $(X_t^{IMU}, X_t^{gyro}, I_t)$

872 6: Establish Web socket connection with PC

873 7: Transmit (p_t, R_t) to PC

874 8: **end function**

875 9: **function** PCPROCESSINGANDCONTROL

876 10: Receive (p_t, R_t) from mobile device

877 11: $\Delta p_t = p_t - p_{t-1}$, $\Delta R_t = R_t \cdot R_{t-1}^{-1}$

878 12: $\Delta \hat{p}_t = \text{MedianFilter}(\{\Delta p_{t-9}, \dots, \Delta p_t\})$

879 13: $\Delta \hat{R}_t = \text{MedianFilter}(\{\Delta R_{t-9}, \dots, \Delta R_t\})$

880 14: $\Delta q_t = \text{InverseKinematics}(\Delta \hat{p}_t, \Delta \hat{R}_t)$

881 15: Send joint position commands Δq_t to robot

882 16: **end function**

883

Benchmark	Average Length
ManiSkill	52.3
CALVIN	30
Libero	77.3
ARNOLD	125.8
VLABench	157.2
RLBench	180.2
RoboCASA Composition	371.9
COIN-50	988.9

893
894 Table 5: Trajectory length comparison across benchmarks. COIN features substantially longer
895 temporal horizons, requiring extended reasoning and planning capabilities.
896

897 **Multi-Solution Task Diversity:** Over 50% of COIN tasks exhibit substantial procedural diversity
898 with multiple valid solution paths. For example, in Tabletop-Find-Dice, agents can either systemati-
899 cally examine all faces or directly place dice on markers. This diversity prevents simple memorization
900 and requires genuine reasoning capabilities.
901

902 **Temporal Dependencies:** At least 40 tasks exhibit strong temporal dependencies where earlier
903 information and actions are essential for later stages, ensuring models must reason over extended
904 horizons beyond local cues.
905

906 **H.2 TASK CLASSIFICATION DETAILS**

907 COIN evaluates interactive reasoning across three principal domains—object-centric, robot-centric,
908 and compositional—each capturing distinct yet interdependent aspects of embodied intelligence
909 required for manipulation under partial observability.
910

911 **H.3 OBJECT-CENTRIC REASONING**

912 Object-centric reasoning encompasses an agent’s capacity to infer and utilize knowledge about
913 environmental entities through strategic interaction:
914

915 • **Physical Property Inference (MAS, FRI, SCA, MOV):** This category examines the agent’s
916 ability to:

918 – *Mass Estimation (MAS)*: Infer object mass distributions and leverage this information to
 919 modulate manipulation forces appropriately for optimal handling.
 920 – *Friction Coefficient Assessment (FRI)*: Deduce surface characteristics and their implications
 921 for stable grasping and precise manipulation without slippage.
 922 – *Scale Analysis (SCA)*: Assess dimensional compatibility between objects and end-effectors to
 923 determine feasible manipulation strategies in space-constrained environments.
 924 – *Moveable Analysis (MOV)*: Determine which objects or object parts can be manipulated and
 925 the constraints on their movement, distinguishing between fixed, partially constrained, and
 926 freely movable elements.
 927 • **Spatial Reasoning (OBS, GEO, ORI, SRA)**: This domain evaluates the agent’s capacity to:
 928 – *Obstacle Handling (OBS)*: Identify which objects constitute impediments in the manipulation
 929 space, distinguish between movable and fixed obstacles, and determine whether to relocate
 930 obstacles or navigate around them based on task constraints and efficiency.
 931 – *Orientation Analysis (ORI)*: Determine optimal object reorientation for task completion based
 932 on geometric and functional constraints in three-dimensional space.
 933 – *Spatial Relationship Analysis (SRA)*: Identify and reason about relative positions between
 934 objects, including containment relationships (objects inside other objects), proximity rela-
 935 tionships (nearest/farthest objects), and spatial arrangements crucial for task execution.
 936 • **Mechanism Understanding (LOC, KIN, SEQ)**: This category evaluates the agent’s ability to:
 937 – *Locking System Comprehension (LOC)*: Deduce the operational principles of locking mech-
 938 anisms and develop appropriate manipulation sequences to engage or disengage them.
 939 – *Kinematic Constraint Inference (KIN)*: Identify axes of motion in articulated objects to
 940 facilitate effective manipulation within their degrees of freedom.
 941 – *Sequential Mechanism Navigation (SEQ)*: Comprehend multi-stage mechanisms and their
 942 state-dependent behavior patterns to achieve desired functional outcomes.
 943 • **Visual Reasoning (GEO, VCP, SEM, OCC)**: This category evaluates the agent’s ability to
 944 process and interpret visual information for manipulation:
 945 – *Geometric Reasoning (GEO)*: Infer shape-based affordances and constraints that influence
 946 grasp planning and execution for objects with complex geometries and non-standard forms.
 947 – *Visual Comparison (VCP)*: Compare visual properties across object states or between ob-
 948 served objects and internal representations to detect changes, identify matching features, or
 949 recognize anomalies.
 950 – *Semantic Segmentation (SEM)*: Distinguish between different objects or object parts based on
 951 visual features, enabling precise targeting of specific components during manipulation tasks.
 952 – *Occlusion Handling (OCC)*: Reason about partially visible or temporarily hidden objects by
 953 maintaining object permanence and inferring obscured geometries from limited viewpoints.
 954
 955

H.4 ROBOT-CENTRIC REASONING

956 Robot-centric reasoning evaluates an agent’s capacity for self-awareness and adaptation within the
 957 manipulation context, addressing the embodied nature of interaction:

958 • **Embodiment Awareness (MOR, PPO, KCA)**: This domain assesses the agent’s ability to:
 959 – *Morphological Reasoning (MOR)*: Account for the robot’s physical dimensions when plan-
 960 ning trajectories and interactions to avoid self-collisions and ensure manipulability.
 961 – *Perceptual Perspective Optimization (PPO)*: Strategically adjust sensor positioning to max-
 962 imize information gain during task execution, particularly in partially observable environ-
 963 ments.
 964 – *Kinematic Constraint Awareness (KCA)*: Reason about joint limitations and workspace
 965 boundaries during motion planning to ensure executable action sequences.
 966 • **Control Optimization (DYN, ACT, SKL)**: This category evaluates the agent’s capacity to:
 967 – *Dynamic Response Tuning (DYN)*: Adapt control parameters based on task requirements and
 968 environmental conditions to achieve desired manipulation outcomes.
 969
 970

972
 973 – *Action Space Navigation (ACT)*: Select appropriate actions from the available repertoire based
 974 on current state and task objectives within continuous control spaces.
 975 – *Skill Adaptation (SKL)*: Identify and modify learned manipulation skills to meet specific task
 976 requirements and environmental conditions.

977 **H.5 COMPOSITIONAL REASONING CAPABILITIES**

979 Compositional reasoning encompasses higher-order cognitive functions that integrate multiple reasoning
 980 modalities to address complex, interactive challenges requiring adaptive strategies:

982 • **Tool-Mediated Problem Solving (TOO)**: Identify, create, or adapt tools to overcome manipulation
 983 constraints and extend interaction capabilities beyond the robot's native end-effector, enabling
 984 solutions to otherwise infeasible tasks.

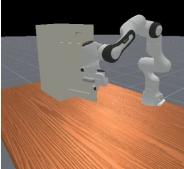
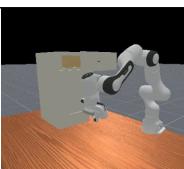
985 • **Failure-Driven Adaptation (FDA)**: Actively interact with the environment to gather information
 986 about failure modes, then systematically refine strategies based on observed outcomes to develop
 987 more robust manipulation approaches through iterative testing.

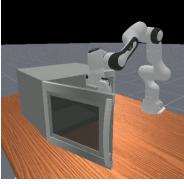
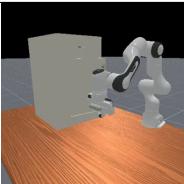
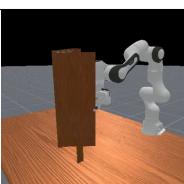
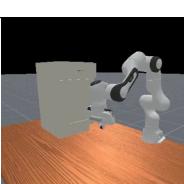
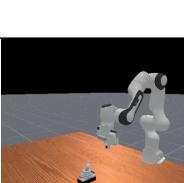
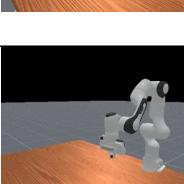
988 • **Hierarchical Planning (PLA)**: Decompose complex tasks into coherent sequences of subtasks
 989 with appropriate dependencies, adjusting the plan hierarchy in response to changing environmental
 990 conditions or task requirements during execution.

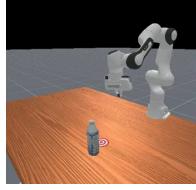
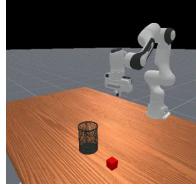
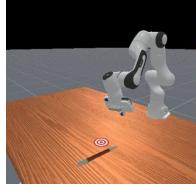
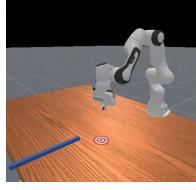
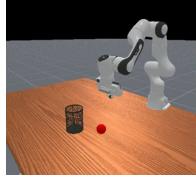
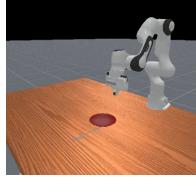
991 • **Experience Utilization (EXP)**: Incorporate historical interaction data into current decision-making
 992 processes, applying lessons from previous manipulation attempts to enhance performance in novel
 993 but related scenarios through transfer learning.

995 **I TASKS TABLE**

997 **I.1 PRIMITIVE TASK**

Task ID	Task Image	Description
Tabletop-Close-Cabinet-v1		Close the cabinet door
Tabletop-Close-Door-v1		Close the door
Tabletop-Close-Drawer-v1		Close the drawer

1026	Task ID	Task Image	Description
1027	Tabletop-Close-Microwave-v1		Close the microwave
1028	Tabletop-Open-Cabinet-v1		Open the cabinet door
1029	Tabletop-Open-Door-v1		Open the door
1030	Tabletop-Open-Drawer-v1		Open the drawer
1031	Tabletop-Open-Microwave-v1		Open the microwave
1032	Tabletop-Open-Trigger-v1		Turn on the trigger
1033	Tabletop-Pick-Apple-v1		Pick the apple to the marker
1034			
1035			
1036			
1037			
1038			
1039			
1040			
1041			
1042			
1043			
1044			
1045			
1046			
1047			
1048			
1049			
1050			
1051			
1052			
1053			
1054			
1055			
1056			
1057			
1058			
1059			
1060			
1061			
1062			
1063			
1064			
1065			
1066			
1067			
1068			
1069			
1070			
1071			
1072			
1073			
1074			
1075			
1076			
1077			
1078			
1079			

1080	Task ID	Task Image	Description
1081	Tabletop-Pick-Book-FromShelf-v1		Find and pick the book from the bookshelf
1082	Tabletop-Pick-Bottle-v1		Pick up the bottle and put it on the marker
1083	Tabletop-Pick-Cube-ToHolder-v1		Pick up the cube, put it in the holder
1084	Tabletop-Pick-Pen-v1		Pick up the pen and put it to the marker
1085	Tabletop-Pull-Pivot-v1		Pull the pivot to the target area
1086	Tabletop-Put-Ball-IntoContainer-v1		Put the ball into the container
1087	Tabletop-Put-Fork-OnPlate-v1		Put the fork on the plate
1088			
1089			
1090			
1091			
1092			
1093			
1094			
1095			
1096			
1097			
1098			
1099			
1100			
1101			
1102			
1103			
1104			
1105			
1106			
1107			
1108			
1109			
1110			
1111			
1112			
1113			
1114			
1115			
1116			
1117			
1118			
1119			
1120			
1121			
1122			
1123			
1124			
1125			
1126			
1127			
1128			
1129			
1130			
1131			
1132			
1133			

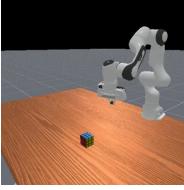
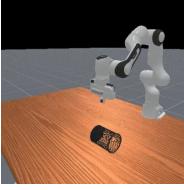
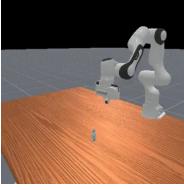
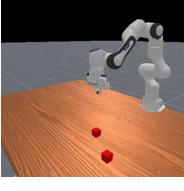
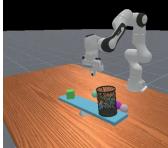
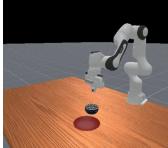
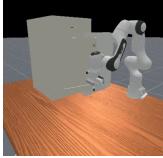
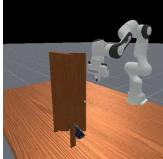
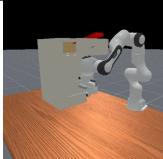
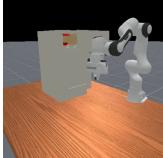
Task ID	Task Image	Description
Tabletop-Rotate-Cube-v1		Rotate the cube till the white face upward
Tabletop-Rotate-Holder-v1		Rotate the holder till the hole upward
Tabletop-Rotate-USB-v1		Rotate the USB body for 90 degrees
Tabletop-Stack-Cubes-v1		Stack all the cubes

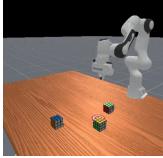
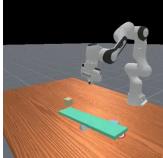
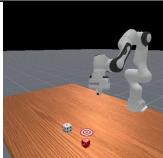
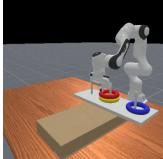
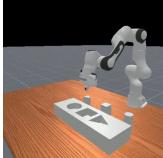
Table 6: Complete COIN-Primitive task specifications with visual examples (20 tasks)

I.2 INTERACTIVE TASK

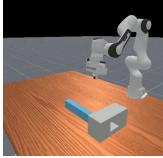
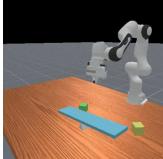
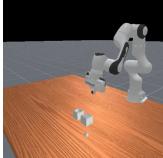
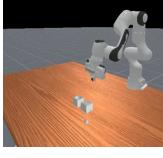
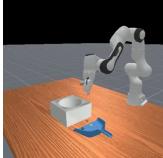
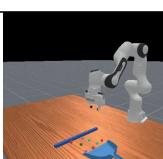
Task ID	Task Image	Description	Obj.	Rob.	Comp.
Tabletop-Balance-Pivot-WithBalls-v1		Put the balls in to the holder to balance the long board on the triangular prism	MAS SCA	no	TOO LPE FDA PLA
Tabletop-Clean-For-Dinner-v1		Arrange the bowl, fork onto the plate, clean for dinner	no	no	no

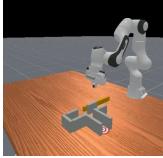
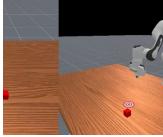
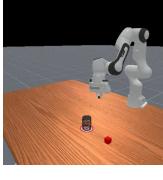
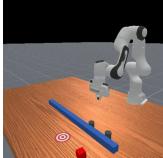
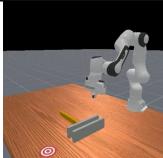
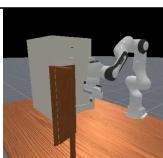
1188	Task ID	Task Image	Description	Obj.	Rob.	Comp.
1189	Tabletop-Close-Cabinet-WithObstacle-v1		close the cabinet door	OBS	no	PLA
1190	Tabletop-Close-Door-WithObstacle-v1		close the door	no	no	no
1191	Tabletop-Close-Drawer-WithLongObstacle-v1		close the drawer	OBS GEO	no	PLA FDA
1192	Tabletop-Close-Drawer-WithObstacle-v1		close the drawer	OBS GEO	ACT PPO	PLA FDA
1193	Tabletop-Find-Book-Black-v1		Find and pick the black book from the bookshelf and put it on the marker	GEO OBS	PPO	EXP
1194	Tabletop-Find-Book-FromShelf-v1		Find and pick the highest book from the bookshelf and put it on the marker	GEO	PPO	EXP

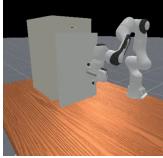
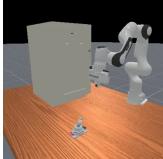
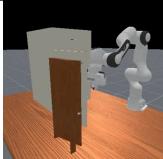
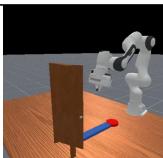
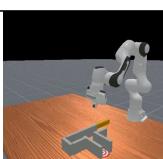
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

Task ID	Task Image	Description	Obj.	Rob.	Comp.
1242 1243 1244 1245 1246 1247 1248 1249 1250 1251		find the cube which have red face downward, and put it on the marker with red face upward	ORI	PPO ACT	EXP PLA
1252 1253 1254 1255 1256 1257 1258		Move the heavy cube to the goal region	MAS	no	TOO PLA
1259 1260 1261 1262 1263 1264 1265 1266 1267		find the dice which have 2 and 4 point in the corresponding face and put it on the marker	GEO ORI	PPO	EXP PLA
1268 1269 1270 1271 1272 1273 1274		Place all the hanobi in big to small from bottom to up	GEO SEQ	ACT	PLA EXP
1275 1276 1277 1278 1279 1280 1281		insert the conical to the container	GEO ORI	no	no
1282 1283 1284 1285 1286 1287 1288 1289		insert all the stick on the table into corresponding holes	GEO	no	PLA

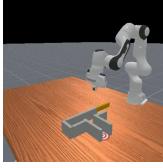
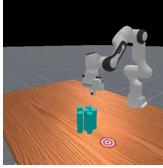
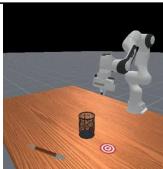
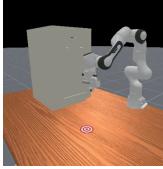
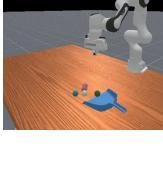
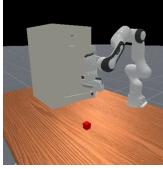
1290
1291
1292
1293
1294
1295

1296	Task ID	Task Image	Description	Obj.	Rob.	Comp.
1297	Tabletop-Insert-WithOrientation-v1		insert the board on the wall	GEO ORI	no	PLA FDA
1298	Tabletop-Keep-Pivot-Balance-v1		Balance the long board on the triangular prism and place the cubes to maintain balance	MAS	no	TOO LPE FDA
1299	Tabletop-Lift-Book-v1		lift the book up to the higher platform	GEO ORI SCA	MOR	PLA
1300	Tabletop-Merge-Box-v1		Merge ball and boxs up	GEO ORI	no	no
1301	Tabletop-Merge-USB-v1		Pick up the USB body and insert it into the USB hub	GEO	no	PLA
1302	Tabletop-Move-Balls-WithDustpan-v1		move all the balls into the holder with dustpan	MAS SCA GEO	no	TOO LPE
1303	Tabletop-Move-Balls-WithPivot-v1		move all the balls into the dustpan as fast as you can	SCA GEO	no	TOO LPE PLA
1304						
1305						
1306						
1307						
1308						
1309						
1310						
1311						
1312						
1313						
1314						
1315						
1316						
1317						
1318						
1319						
1320						
1321						
1322						
1323						
1324						
1325						
1326						
1327						
1328						
1329						
1330						
1331						
1332						
1333						
1334						
1335						
1336						
1337						
1338						
1339						
1340						
1341						
1342						
1343						
1344						
1345						
1346						
1347						
1348						
1349						

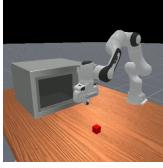
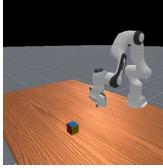
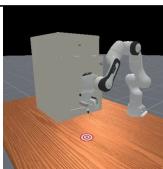
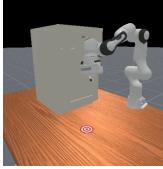
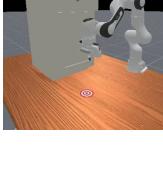
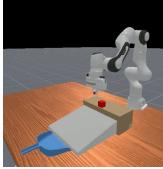
1350	Task ID	Task Image	Description	Obj.	Rob.	Comp.
1351	Tabletop-Move-Cross-WithStick-v1		Use the stick to move the small cube along the cross-shaped path to the target position	no	no	no
1352	Tabletop-Move-Cube-DynamicFriction-v1		move the cube to the marker	FRI MAS	no	PLA LPE FDA
1353	Tabletop-Move-Cube-WithHolder-v1		move the cube to the marker and put the holder on the cube	SCA GEO SEQ	no	PLA
1354	Tabletop-Move-Cube-WithPivot-v1		move the cube with the pivot to the marker	MAS	ACT DYN	PLA TOO LPE FDA
1355	Tabletop-Move-Line-WithStick-v1		Use the stick to move the small cube along the straight line path to the target position	GEO ORI	ACT	PLA TOO
1356	Tabletop-Open-Cabinet-WithDoor-v1		open the cabinet door	OBS	ACT	PLA
1357						
1358						
1359						
1360						
1361						
1362						
1363						
1364						
1365						
1366						
1367						
1368						
1369						
1370						
1371						
1372						
1373						
1374						
1375						
1376						
1377						
1378						
1379						
1380						
1381						
1382						
1383						
1384						
1385						
1386						
1387						
1388						
1389						
1390						
1391						
1392						
1393						
1394						
1395						
1396						
1397						
1398						
1399						
1400						
1401						
1402						
1403						

1404	Task ID	Task Image	Description	Obj.	Rob.	Comp.
1405	Tabletop-Open-Cabinet-WithObstacle-v1		open the cabinet door	OBS	no	PLA
1406	Tabletop-Open-Cabinet-WithSwitch-v1		open the door, notice the switch will control the state of the door	LOC	no	PLA FDA
1407	Tabletop-Open-Door-WithCabinet-v1		open the door	OBS	no	PLA
1408	Tabletop-Open-Door-WithObstacle-v1		open the door	OBS	no	PLA FDA
1409	Tabletop-Pick-Cube-Slippery-v1		Pick the slippery cube	FRI	ACT	PLA TOO LPE FDA
1410	Tabletop-Pick-Cube-WithDoor-v1		put the cube to the marker	OBS	KIN ACT	PLA FDA

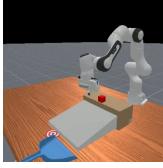
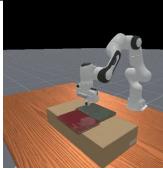
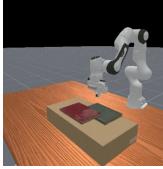
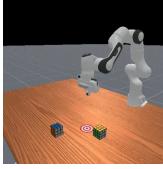
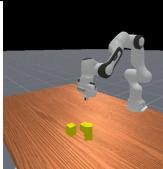
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

1458	Task ID	Task Image	Description	Obj.	Rob.	Comp.
1459	Tabletop-Pick-Cube-WithStick-v1		Use the stick to move the small cube along the T-shaped path to the target position	GEO ORI	ACT	PLA TOO
1460	Tabletop-Pick-Cylinder-WithObstacle-v1		pick up the center cylinder	LOC	KIN	PLA FDA EXP
1461	Tabletop-Pick-Eraser-FromHolder-v1		Pick up the eraser in the holder and place it to the marker	GEO ORI	MOR	PLA EXP
1462	Tabletop-Pick-Object-FromCabinet-v1		pick up the object from the cabinet	OBS GEO	PPO MOR ACT	PLA FDA
1463	Tabletop-Put-Balls-IntoContainer-v1		move all the balls into the dustpan as fast as you can	GEO	ACT	TOO LPE PLA FDA
1464	Tabletop-Put-Cube-IntoCabinetWithObstacle-v1		put the object into the cabinet	OBS GEO	PPO MOR ACT	PLA FDA

1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

1512	Task ID	Task Image	Description	Obj.	Rob.	Comp.
1513	Tabletop-Put-Cube-IntoMicrowave-v1		put the cube into the microwave	OBS GEO	ACT MOR	PLA FDA
1514	Tabletop-Rotate-Cube-Twice-v1		rotate the cube till the green face upward	ORI	ACT	PLA FDA
1515	Tabletop-Seek-Holder-InCabinet-v1		Find the holder containing an eraser the cabinet, put it to the marker	OBS GEO SEQ	ACT MOR PPO	PLA EXP
1516	Tabletop-Seek-Objects-InCabinet-v1		Find the apple and the plate in the cabinet, put them on the table	OBS	ACT MOR PPO	FDA PLA EXP
1517	Tabletop-Seek-Objects-WithObstacle-v1		find the cube in the cabinet and pick it up	OBS	MOR ACT PPO	FDA PLA
1518	Tabletop-Slide-Cube-Into-Container-v1		Slide a cube down a slope into a container	no	no	no

1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

1566	Task ID	Task Image	Description	Obj.	Rob.	Comp.
1567	Tabletop-Slide-Cube-WithPath-v1		Slide a cube down a slope to the marker	FRI ORI GEO	ACT	PLA
1568	Tabletop-Stack-Books-OnBox-v1		Stack all things on the table	ORI GEO	PPO	PLA
1569	Tabletop-Stack-Books-v1		Stack all things on the table	SCA ORI	PPO	PLA
1570	Tabletop-Stack-Cube-WithColor-v1		Stack the cube with same color	ORI	ACT	PLA FDA
1571	Tabletop-Stack-LongObjects-v1		stack all the objects to make it most high	SCA ORI GEO OBS	ACT	FDA PLA
1572						
1573						
1574						
1575						
1576						
1577						
1578						
1579						
1580						
1581						
1582						
1583						
1584						
1585						
1586						
1587						
1588						
1589						
1590						
1591						
1592						
1593						
1594						
1595						
1596						
1597						
1598						
1599						
1600						
1601						
1602						
1603						
1604	Table 7: COIN-50 interactive reasoning task specifications with visual examples (50 tasks)					
1605						
1606						
1607						
1608						
1609						
1610						
1611						
1612						
1613						
1614						
1615						
1616						
1617						
1618						
1619						