

COLOSSEUM V2: Appendix

APPENDIX

The following contains additional details on the hardware and simulation results from the proposed benchmark. Note that throughout this report, **Blue** indicates a visual variation, **green** indicates an action variation, and **red** indicates a language variation.

Note that in figs. 4 to 7 only tasks for which the `None` success rate is above 10% are included in the calculation for ‘Mean % Change from None’. This reduces the noise of the statistic.

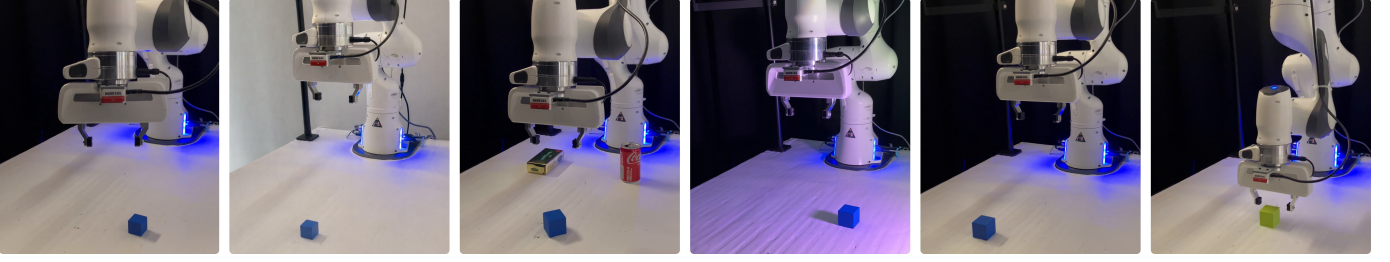


Fig. 1. Hardware setup with 5 variations for the RAISECUBE task. In order from left to right, the variations are None, Background-Color, Distractor-Objects, Light-Color, MO-Size, and MO-Color

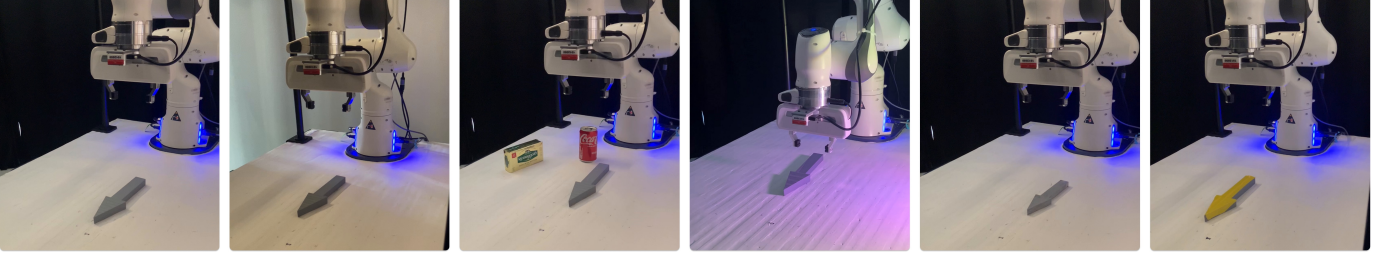


Fig. 2. Hardware setup with 5 variations for the ROTATEARROW task. In order from left to right, the variations are None, Background-Color, Distractor-Objects, Light-Color, MO-Size, and MO-Color

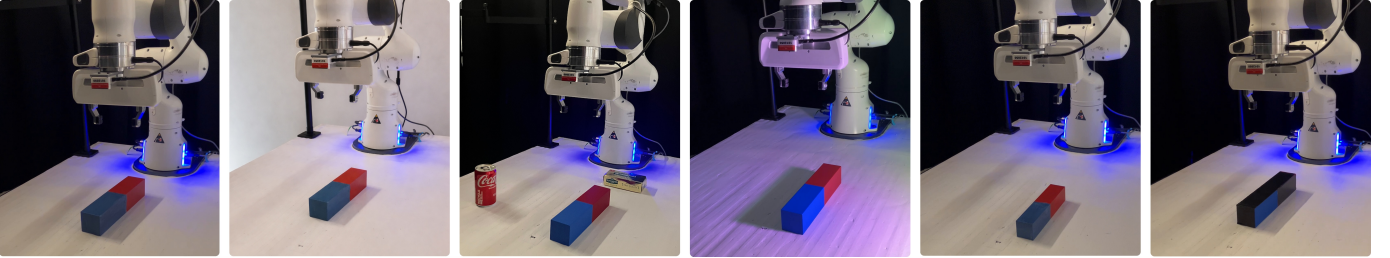


Fig. 3. Hardware setup with 5 variations for the LIFTPEGUPRIGHT task. In order from left to right, the variations are None, Background-Color, Distractor-Objects, Light-Color, MO-Size, and MO-Color

TASK	None	All															Language
			Mo color	Mo texture	Ro color	Ro texture	Table color	Light color	Table texture	Distractor object	Background texture	Background color	Camera pose	Pose randomization	Mo size	Ro size	
RaiseCube	100.0	0.0	41.5	35.5	-	-	0.0	81.5	10.5	85.0	100.0	100.0	96.5	79.5	99.0	-	100.0
PickDishFromRack	100.0	1.5	91.5	66.5	57.5	58.5	0.0	47.5	17.0	89.0	100.0	100.0	100.0	-	-	-	100.0
RotateArrow	97.0	0.0	45.5	12.5	-	-	0.0	14.5	1.5	58.5	59.5	94.5	96.5	74.5	86.0	-	98.5
LiftPegUpright	53.5	0.0	36.5	8.0	-	-	0.0	28.0	5.0	37.5	32.0	25.0	47.5	36.5	24.0	-	53.5
HammerNail	52.0	2.5	5.0	2.0	61.0	33.5	0.0	0.5	1.0	27.0	13.5	21.0	16.5	36.0	27.5	-	41.0
PickSodaFromCabinet	24.5	0.0	1.0	3.5	5.5	1.5	9.0	0.0	29.0	23.0	25.0	46.5	5.0	-	-	-	31.5
ScoopBanana	18.5	7.5	5.5	6.5	18.0	18.5	3.0	19.5	4.0	11.5	11.0	9.0	13.0	18.0	20.5	13.0	19.0
PlaceDishInRack	9.5	0.0	10.5	2.5	7.5	15.5	0.0	9.5	1.5	6.5	14.5	24.0	10.0	11.0	-	-	8.0
StackCube	8.5	0.0	1.5	0.5	-	-	0.0	2.5	0.0	2.5	12.5	6.5	2.5	4.0	6.0	8.0	10.5
OpenDrawer	3.0	6.0	-	-	-	-	1.0	0.0	0.0	8.0	0.0	0.0	0.0	2.0	-	-	2.5
PegInsertionSide	1.5	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	-	-	0.5
PlaceCubeInDrawer	0.5	0.0	-	-	-	-	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	-	0.5
PlaceBookInShelf	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	1.0	0.0	0.0	-	-	0.0
PlugCharger	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	0.0
OpenCabinet	0.0	4.5	-	-	-	-	0.0	0.0	0.0	6.0	0.0	0.0	0.0	0.0	-	-	0.0
CookItemInPan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Mean % Change from None	0.0	-93.3	-58.3	-73.8	-26.4	-42.7	-92.4	-56.8	-74.2	-26.8	-27.3	-11.0	-27.5	-21.8	-20.8	-29.7	1.7

Fig. 4. Raw results for ACT on the Single-Arm task suite

TASK	None	All	Mo color	Mo texture	Ro color	Ro texture	Table color	Light color	Table texture	Distractor object	Background texture	Background color	Camera pose	Pose randomization	Mo size	Ro size	Language
DualArmLiftPot	93.0	3.0	63.0	36.5	-	-	76.5	30.5	23.0	85.0	86.0	87.0	81.0	88.5	47.0	-	95.5
DualArmCubeHandover	86.0	1.5	42.5	36.5	-	-	23.0	34.5	7.0	9.5	0.0	0.0	0.0	51.5	85.0	-	89.0
DualArmLiftTray	85.5	0.5	89.0	51.5	-	-	73.5	76.0	28.0	68.0	74.5	82.0	45.0	58.5	20.0	-	82.5
DualArmPushBox	85.0	0.5	75.5	67.0	85.0	69.0	57.5	75.0	28.0	75.5	83.0	87.5	52.0	51.0	75.5	88.5	84.0
DualArmThreading	76.0	0.5	49.5	29.0	53.5	47.5	11.0	33.0	1.0	19.0	44.0	54.5	47.0	37.5	-	-	74.0
DualArmPourPot	62.0	2.5	66.0	26.0	55.5	60.0	54.5	32.5	61.5	48.0	58.5	58.5	47.5	21.5	25.0	60.5	65.0
DualArmStackCube	46.5	0.0	7.5	5.5	41.5	50.5	7.0	24.5	7.0	15.0	30.0	39.5	2.5	21.5	38.0	51.5	48.0
DualArmPenCap	4.5	0.5	6.0	2.5	2.5	3.5	5.5	2.5	2.5	4.0	4.5	2.0	3.5	5.5	-	-	3.5
DualArmStackTwoCubes	1.5	0.0	0.0	0.5	1.5	2.0	1.0	1.5	0.0	1.0	1.5	2.0	0.0	0.5	3.5	1.0	3.0
DualArmBottleHandover	0.0	0.0	0.0	0.0	-	-	0.0	0.0	0.0	1.0	1.0	0.5	0.0	0.0	0.0	-	0.5
DualArmDrawerPlace	0.0	0.0	0.0	0.0	-	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0
DualArmDrawerOpen	0.0	0.0	-	-	-	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	0.0
Mean % Change from None	0.0	-98.5	-28.9	-55.3	-12.7	-12.7	-45.7	-43.1	-69.4	-42.1	-29.4	-22.4	-50.8	-40.9	-36.1	4.2	1.0

Fig. 5. Raw results for ACT on the Bimanual task suite

TASK	None	All	Mo color	Mo texture	Ro color	Ro texture	Table color	Light color	Table texture	Distractor object	Background texture	Background color	Camera pose	Pose randomization	Mo size	Ro size	Language
PickDishFromRack	100.0	-	71.0	79.0	93.0	85.0	88.0	95.0	59.0	89.0	95.0	87.0	82.0	-	-	-	77.0
RotateArrow	86.0	6.0	59.0	62.0	-	-	68.0	76.0	44.0	65.0	78.0	85.0	73.0	62.0	80.0	-	22.0
RaiseCube	74.0	14.0	58.0	59.0	-	-	52.0	67.0	50.0	70.0	83.0	73.0	74.0	53.0	75.0	-	61.0
OpenDrawer	25.0	15.0	-	-	-	-	22.0	21.0	29.0	20.0	29.0	29.0	32.0	25.0	-	-	22.0
ScoopBanana	21.0	14.0	18.0	17.0	19.0	24.0	16.0	21.0	19.0	27.0	21.0	20.0	21.0	22.0	22.0	21.0	8.0
PickSodaFromCabinet	11.0	4.0	9.0	14.0	6.0	4.0	12.0	5.0	8.0	14.0	9.0	2.0	10.0	-	-	-	12.0
HammerNail	7.0	2.0	14.0	16.0	14.0	14.0	16.0	16.0	6.0	14.0	13.0	12.0	5.0	8.0	6.0	-	8.0
StackCube	6.0	0.0	1.0	1.0	-	-	3.0	3.0	1.0	6.0	10.0	10.0	8.0	6.0	7.0	6.0	10.0
LiftPegUpright	4.0	0.0	2.0	0.0	-	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	-	17.0
PlaceDishInRack	3.0	1.0	7.0	4.0	7.0	7.0	8.0	8.0	5.0	7.0	8.0	10.0	8.0	6.0	-	-	4.0
PlaceBookInShelf	2.0	1.0	1.0	3.0	3.0	1.0	3.0	5.0	4.0	3.0	3.0	2.0	3.0	1.0	-	-	3.0
PlaceCubeInDrawer	1.0	0.0	-	-	-	-	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	-	0.0
PegInsertionSide	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	0.0
PlugCharger	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	0.0
OpenCabinet	0.0	1.0	-	-	-	-	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	-	-	0.0
CookItemInPan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
Mean % Change from None	0.0	-62.2	-22.9	-12.2	-20.7	-21.5	-14.9	-16.1	-23.8	-0.8	-0.7	-14.3	-2.4	-12.9	-0.3	0.0	-30.0

Fig. 6. Raw results for $\pi_{0.5}$ on the Single-Arm task suite

TASK	None	All	Mo color	Mo texture	Ro color	Ro texture	Table color	Light color	Table texture	Distractor object	Background texture	Background color	Camera pose	Pose randomization	Mo size	Ro size	Language
DualArmPushBox	56.0	0.0	51.0	40.0	38.0	33.0	37.0	55.0	13.0	46.0	57.0	60.0	46.0	29.0	58.0	56.0	5.0
DualArmPourPot	41.0	0.0	23.0	26.0	51.0	38.0	21.0	34.0	52.0	23.0	33.0	35.0	42.0	12.0	17.0	41.0	34.0
DualArmCubeHandover	30.0	0.0	10.0	11.0	-	-	12.0	31.0	5.0	8.0	31.0	34.0	21.0	16.0	17.0	-	1.0
DualArmLiftTray	15.0	0.0	4.0	3.0	-	-	19.0	7.0	2.0	4.0	0.0	1.0	15.0	9.0	5.0	-	6.0
DualArmLiftPot	5.0	4.0	9.0	6.0	-	-	7.0	4.0	6.0	6.0	14.0	18.0	5.0	6.0	6.0	-	11.0
DualArmPenCap	4.0	3.0	1.0	1.0	3.0	3.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	6.0	-	-	1.0
DualArmStackCube	3.0	0.0	0.0	0.0	1.0	2.0	1.0	4.0	0.0	0.0	0.0	0.0	9.0	1.0	0.0	3.0	0.0
DualArmBottleHandover	0.0	0.0	3.0	3.0	-	-	1.0	3.0	1.0	1.0	1.0	1.0	2.0	2.0	3.0	-	4.0
DualArmThreading	0.0	0.0	0.0	0.0	1.0	1.0	4.0	1.0	0.0	2.0	0.0	2.0	1.0	0.0	-	-	0.0
DualArmDrawerPlace	0.0	0.0	0.0	0.0	-	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0
DualArmDrawerOpen	0.0	0.0	-	-	-	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	0.0
DualArmStackTwoCubes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean % Change from None	0.0	-100.0	-48.2	-52.1	-3.9	-24.2	-29.0	-17.2	-55.0	-52.1	-28.6	-21.9	-11.4	-51.4	-41.2	0.0	-66.2

Fig. 7. Raw results for $\pi_{0.5}$ on the Bimanual task suite

Model	Included tasks	Excluded tasks
ACT - Single-Arm	HAMMERNAIL, LIFTPEGUPRIGHT, PICKDISHFROMRACK, PICKSODAFROMCABINET, RAISECUBE, ROTATEARROW, SCOOPBANANA	COOKITEMINPAN, OPENCABINET, OPENDRAWER, PEGINSERTIONSIDE, PLACEBOOKINSHELF, PLACECUBEINDRAWER, PLACEDISHINRACK, PLUGCHARGER, STACKCUBE
ACT - Bimanual	DUALARMCUBEHANDOVER, DUALARMLIFTPOT, DUALARMLIFTTRAY, DUALARMPOURPOT, DUALARMPUSHBOX, DUALARMSTACKCUBE, DUALARMTHREADING	DUALARMBOTTLEHANDOVER, DUALARMDRAWEROPEN, DUALARMDRAWERPLACE, DUALARMPENCAP, DUALARMSTACKTWOUBES
Pi0.5 - Single-Arm	OPENDRAWER, PICKDISHFROMRACK, PICKSODAFROMCABINET, RAISECUBE, ROTATEARROW, SCOOPBANANA	COOKITEMINPAN, HAMMERNAIL, LIFTPEGUPRIGHT, OPENCABINET, PEGINSERTIONSIDE, PLACEBOOKINSHELF, PLACECUBEINDRAWER, PLACEDISHINRACK, PLUGCHARGER, STACKCUBE
Pi0.5 - Bimanual	DUALARMCUBEHANDOVER, DUALARMLIFTTRAY, DUALARMPOURPOT, DUALARMPUSHBOX	DUALARMBOTTLEHANDOVER, DUALARMDRAWEROPEN, DUALARMDRAWERPLACE, DUALARMLIFTPOT, DUALARMPENCAP, DUALARMSTACKCUBE, DUALARMSTACKTWOUBES, DUALARMTHREADING

TABLE I

TASKS INCLUDED AND EXCLUDED WHEN CALCULATING THE AVERAGE CHANGE IN SUCCESS RATE FOR EACH PERTURBATION.