

491 6 Appendix

492 6.1 RL Bench Tasks

493 We provide a brief summary of the RL Bench tasks in Tab. 3. There are 18 tasks with 249 variations.
494 For more detailed description of each task, please refer to PerAct [6], Appendix A.

Task	Language Template	# of Variations
open drawer	“open the __ drawer”	3
slide block	“slide the __ block to target”	4
sweep to dustpan	“sweep dirt to the __ dustpan”	2
meat off grill	“take the __ off the grill”	2
turn tap	“turn __ tap”	2
put in drawer	“put the item in the __ drawer”	3
close jar	“close the __ jar”	20
drag stick	“use the stick to drag the cube onto the __ target”	20
stack blocks	“stack __ __ blocks”	60
screw bulb	“screw in the __ light bulb”	20
put in safe	“put the money away in the safe on the __ shelf”	3
place wine	“stack the wine bottle to the __ of the rack”	3
put in cupboard	“put the __ in the cupboard”	9
sort shape	“put the __ in the shape sorter”	5
push buttons	“push the __ button, [then the __ button]”	50
insert peg	“put the __ peg in the spoke”	20
stack cups	“stack the other cups on top of the __ cup”	20
place cups	“place __ cups on the cup holder”	3

Table 3: **Tasks in RL Bench** We evaluate on 18 RL Bench tasks which are same as those used in PerAct [6]. For more details, check see PerAct [6], Appendix A. For videos, visit <https://corlrvt.github.io/>

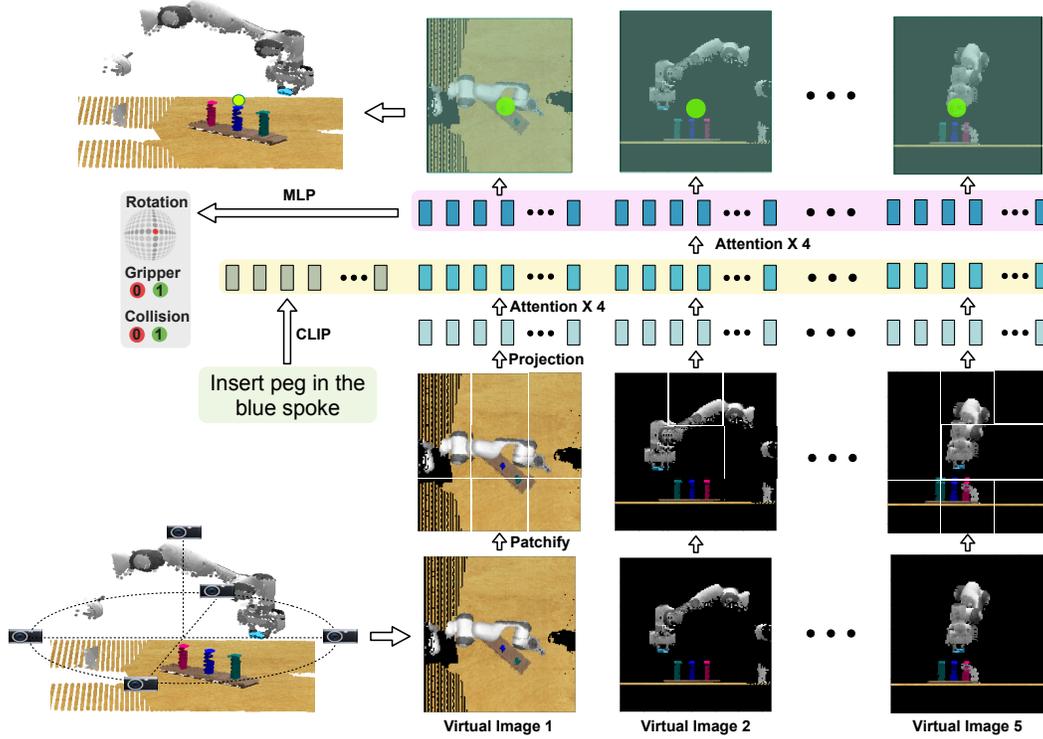


Figure 5: **Overview of the transformer used in RVT.** The input to the transformer is a language description of the task and virtual images of the scene point cloud. The text is converted into token embeddings using the pretrained CLIP [56] model, while the virtual images are converted into token embeddings via patchify and projection operations. For each virtual image, tokens belonging to the same image are processed via four attention layers. Finally, the processed image tokens as well as the language tokens are jointly processed using four attention layers. The 3D action is inferred using the resulting image tokens.

496 **6.3 Ablations**

497 We report the ablations mentioned in Tab. 2, along with the mean and standard deviations for each task Tab. 4.

Im. Res.	View Corr.	Dep. Ch.	Bi-Lev.	Proj. Type	Rot. Aug.	Cam Loc.	# of View	Avg. Succ.	Close Jar	Drag Stick	Insert Peg	Meat off Grill	Open Drawer	Place Cups
220	✓	✓	✓	Orth.	✓	Cube	5	62.9	52 ± 2.5	99.2 ± 1.6	11.2 ± 3	88 ± 2.5	71.2 ± 6.9	4 ± 2.5
100	✓	✓	✓	Orth.	✓	Cube	5	51.1	60 ± 0	83 ± 1.7	4 ± 2.8	91 ± 3.3	67 ± 5.2	1 ± 1.7
220	✗	✓	✓	Orth.	✓	Cube	5	59.7	44 ± 0	100 ± 0	17 ± 4.4	90 ± 6	71 ± 9.1	7 ± 5.9
220	✓	✗	✓	Orth.	✓	Cube	5	60.3	37 ± 3.3	96 ± 0	11 ± 3.3	97 ± 1.7	57 ± 8.2	3 ± 3.3
220	✓	✓	✗	Orth.	✓	Cube	5	58.4	32 ± 7.5	96 ± 0	11 ± 3.3	90 ± 2	68 ± 2.8	2 ± 2
220	✓	✓	✓	Pers.	✓	Cube	5	40.2	20 ± 2.5	90.4 ± 2	4 ± 0	84.8 ± 4.7	13.6 ± 4.8	2.4 ± 2
220	✓	✓	✓	Orth.	✗	Cube	5	60.4	52 ± 0	92 ± 0	12.8 ± 1.6	97.6 ± 4.8	85.6 ± 5.4	0 ± 0
220	✓	✓	✓	Orth.	✓	Cube	3	60.2	44.8 ± 1.6	75.2 ± 4.7	15 ± 3.3	89.6 ± 4.1	68.8 ± 9.3	3.2 ± 1.6
220	✓	✓	✓	Orth.	✓	Front	1	35.8	36 ± 4.9	87 ± 1.7	2 ± 2	90 ± 6	58 ± 6.6	0 ± 0
220	✓	✓	✓	Orth.	✓	Rot. 15	5	59.9	48.8 ± 1.6	99.2 ± 1.6	12 ± 4.4	80 ± 2.5	71.2 ± 9.3	0 ± 0
220	✓	✓	✓	Pers.	✗	Real	4	10.4	14.4 ± 6.5	14.4 ± 5.4	0 ± 0	0 ± 0	22.4 ± 5.4	0 ± 0
220	✓	✓	✓	Ortho.	✗	Real	4	22.9	43.2 ± 4.7	54.4 ± 3.2	0 ± 0	0 ± 0	15.2 ± 5.3	0.8 ± 1.6
Im. Res.	View Corr.	Dep. Ch.	Bi-Lev.	Proj. Type	Rot. Aug.	Cam Loc.	# of View	Avg. Succ.	Place Wine	Push Buttons	Put in Cupboard	Put in Drawer	Put in Safe	Screw Bulb
220	✓	✓	✓	Orth.	✓	Cube	5	62.9	91 ± 5.2	100 ± 0	49.6 ± 3.2	88 ± 5.7	91.2 ± 3	48 ± 5.7
100	✓	✓	✓	Orth.	✓	Cube	5	51.1	38 ± 8.7	100 ± 0	49 ± 4.4	86 ± 2	77 ± 1.7	22 ± 4.5
220	✗	✓	✓	Orth.	✓	Cube	5	59.7	96 ± 2.8	99 ± 1.7	48 ± 6.9	50 ± 6	79 ± 5.9	36 ± 0
220	✓	✗	✓	Orth.	✓	Cube	5	60.3	71 ± 1.7	99 ± 1.7	56 ± 0	92 ± 4.9	77 ± 3.3	39 ± 4.4
220	✓	✓	✗	Orth.	✓	Cube	5	58.4	65 ± 5.2	100 ± 0	54 ± 2	94 ± 4.5	78 ± 3.5	48 ± 6.3
220	✓	✓	✓	Pers.	✓	Cube	5	40.2	28 ± 5.7	91.2 ± 1.6	26.4 ± 2	64.8 ± 3	51.2 ± 3.9	20 ± 4.4
220	✓	✓	✓	Orth.	✗	Cube	5	60.4	84 ± 3.6	96 ± 2.5	40 ± 2.5	88 ± 7.2	90.4 ± 4.1	48 ± 8.4
220	✓	✓	✓	Orth.	✓	Cube	3	60.2	84.8 ± 8.9	97.6 ± 2	40.8 ± 4.7	94.4 ± 4.1	82.4 ± 7.8	43.2 ± 3.9
220	✓	✓	✓	Orth.	✓	Front	1	35.8	82 ± 4.5	46 ± 2	14 ± 4.5	29 ± 7.1	57 ± 5.9	6 ± 2
220	✓	✓	✓	Orth.	✓	Rot. 15	5	59.9	74.4 ± 5.4	99.2 ± 1.6	46.4 ± 4.1	81.6 ± 2	80.8 ± 4.7	45.6 ± 4.8
220	✓	✓	✓	Pers.	✗	Real	4	10.4	11.2 ± 3.9	26.4 ± 4.1	0 ± 0	0 ± 0	0 ± 0	0 ± 0
220	✓	✓	✓	Ortho.	✗	Real	4	22.9	67.2 ± 5.9	76 ± 5.7	0 ± 0	0 ± 0	0 ± 0	0 ± 0
Im. Res.	View Corr.	Dep. Ch.	Bi-Lev.	Proj. Type	Rot. Aug.	Cam Loc.	# of View	Avg. Succ.	Slide Block	Sort Shape	Stack Blocks	Stack Cups	Sweep to Dustpan	Turn Tap
220	✓	✓	✓	Orth.	✓	Cube	5	62.9	81.6 ± 5.4	36 ± 2.5	28.8 ± 3.9	26.4 ± 8.2	72 ± 0	93.6 ± 4.1
100	✓	✓	✓	Orth.	✓	Cube	5	51.1	93 ± 3.3	18 ± 2	17 ± 5.2	1 ± 1.7	36 ± 0	76 ± 2.8
220	✗	✓	✓	Orth.	✓	Cube	5	59.7	83 ± 1.7	41 ± 4.4	26.7 ± 5	20 ± 4.9	72 ± 0	95 ± 4.4
220	✓	✗	✓	Orth.	✓	Cube	5	60.3	72 ± 4	37 ± 5.2	23 ± 3.3	33 ± 5.9	92 ± 0	95 ± 4.4
220	✓	✓	✗	Orth.	✓	Cube	5	58.4	66 ± 6	31 ± 6.6	25 ± 3.3	29 ± 5.2	72 ± 0	91 ± 3.3
220	✓	✓	✓	Pers.	✓	Cube	5	40.2	88 ± 4.4	19.2 ± 4.7	22.4 ± 9	1.6 ± 2	16 ± 0	80.8 ± 3
220	✓	✓	✓	Orth.	✗	Cube	5	60.4	72.8 ± 1.6	25.6 ± 2	18.4 ± 6	8.8 ± 5.3	84 ± 0	92 ± 2.5
220	✓	✓	✓	Orth.	✓	Cube	3	60.2	95.2 ± 1.6	37.6 ± 4.1	29.6 ± 3.2	8.8 ± 4.7	80 ± 0	92.8 ± 3
220	✓	✓	✓	Orth.	✓	Front	1	35.8	42 ± 2	2 ± 2	0 ± 0	0 ± 0	0 ± 0	93 ± 5.2
220	✓	✓	✓	Orth.	✓	Rot. 15	5	59.9	83 ± 1.7	30.4 ± 5.4	46.4 ± 9.3	20.8 ± 4.7	64 ± 0	94.4 ± 3.2
220	✓	✓	✓	Pers.	✗	Real	4	10.4	37.6 ± 10.6	2.4 ± 3.2	0.8 ± 1.6	0 ± 0	0 ± 0	56.8 ± 6.9
220	✓	✓	✓	Ortho.	✗	Real	4	22.9	72.8 ± 3	7.2 ± 1.6	11.2 ± 4.7	0 ± 0	12 ± 0	53 ± 5.2

Table 4: Ablations results for RVT on RLbench with metrics for each task.

498