

Appendix

Table 3: Training hyperparameters for BridgeVLA

	Pretrain	RLBench Finetune	Colosseum Finetune	Real-robot Finetune
learning rate	5e-5	8e-5	8e-5	2e-5
optimizer	AdamW	AdamW	AdamW	AdamW
batch size	384	192	192	192
warmup steps	400	-	-	-

A Training & Inference Details

Detailed training configurations are summarized in Tab. 3. Throughout both pre-training and fine-tuning, we keep the SigLIP vision encoder and language token embeddings frozen.

Computational Resources:

1. Pre-training: 8 NVIDIA A100 GPUs for 3,800 steps (≈ 2 hours)
2. RLBench fine-tuning: 48 NVIDIA H100 GPUs for 83,000 steps (≈ 20 hours)
3. COLOSSEUM fine-tuning: 48 NVIDIA H100 GPUs for 83,000 steps (≈ 20 hours)
4. GemBench fine-tuning: 40 NVIDIA A100 GPUs for 50 epochs (≈ 2.1 hours)
5. Real-world fine-tuning: 8 NVIDIA A100 GPUs for 300 epochs (≈ 1.5 hours)

For inference, we run BridgeVLA on a machine equipped with an NVIDIA RTX 4090 GPU. To evaluate its inference speed, we conducted 100 trials. From point cloud input to action output, the average end-to-end inference time is 0.21 seconds.

B Simulation Experiments

B.1 Experiments on COLOSSEUM

Setup. The COLOSSEUM benchmark is an extension to the RLBench benchmark. The model is trained on the data from the original RLBench benchmark but evaluated in environments spanning 12 axes of perturbations. These perturbations, which are unseen during training, encompass changes in object texture, color, and size, backgrounds, lighting, distractors and camera poses. In total, the COLOSSEUM creates 20,371 unique task perturbations instances to comprehensively evaluate the generalization capabilities of the model. Specifically, our evaluation includes three steps: 1) train the model with the original RLBench data without perturbations (100 trajectories per task) on 20 tasks, 2) evaluate each task over 25 trials per perturbation, 3) compute the average success rate of all evaluated tasks for every perturbation. Besides the 12 types of perturbations, we also evaluate on basic variations from the original RLBench (denoted as **RLBench** in Tab. 4), and a more challenging setting which combines all the 12 types of perturbations (denoted as **All Perturbations** in Tab. 4).

Baselines. We compare BridgeVLA with five baseline methods. **R3M-MLP** and **MVP-MLP** are two 2D methods that utilize pre-trained visual encoders to process observation images and an MLP for action prediction. Specifically, R3M-MLP uses R3M [33] that is pre-trained on large-scale egocentric human videos; MVP-MLP uses MVP [46] that is pre-trained on millions of in-the-wild data. Both visual encoders show strong adaptability on various robotics tasks in both simulation and the real world. We also compare with three 3D methods introduced in Sec. 4.1.1, *i.e.*, **PerAct** [39], **RVT** [14], and **RVT-2** [15].

Overall			Success Rate (%)					
Models	Avg. SR (%) \uparrow	Avg. Rank \downarrow	All Perturbations	MO-COLOR	RO-COLOR	MO-TEXTURE	RO-TEXTURE	MO-SIZE
R3M-MLP [33]	0.8	5.71	0.6	0.4	0.0	0.0	0.0	1.8
MVP-MLP [46]	1.6	5.0	0.8	1.2	0.0	0.4	0.0	4.44
PerAct [18]	27.9	3.71	7.2	24.0	29.2	28.8	17.71	35.6
RVT [14]	35.4	3.28	6.4	26.0	31.3	44.8	41.1	35.3
RVT-2 [15]	56.7	1.92	15.6 \pm 0.8	53.0 \pm 0.9	54.6 \pm 0.6	59.7 \pm 0.7	56.7 \pm 1.4	60.9 \pm 0.9
BridgeVLA (Ours)	64.0	1.07	18.7 \pm 2.2	60.5 \pm 1.1	63.8 \pm 0.1	63.5 \pm 1.5	68.4 \pm 3.3	69.3 \pm 1.0
Models	RO-SIZE	Light Color	Table Color	Table Texture	Distractor	Background Texture	RLBench	Camera Pose
R3M-MLP [33]	0.0	1.0	1.4	0.2	1.6	1.2	2.0	0.8
MVP-MLP [46]	0.0	1.6	1.6	1.0	3.8	2.2	2.0	2.6
PerAct [18]	29.3	29.1	30.4	23.2	27.1	33.5	39.4	36.3
RVT [14]	40.5	34.0	30.0	45.2	18.8	46.4	53.4	42.2
RVT-2 [15]	53.4 \pm 1.5	58.0 \pm 1.1	62.6 \pm 0.9	56.6 \pm 0.9	60.8 \pm 0.5	68.7 \pm 1.1	68.8 \pm 1.3	64.4 \pm 0.5
BridgeVLA (Ours)	61.7 \pm 0.8	69.7 \pm 1.2	75.7 \pm 0.9	71.3 \pm 0.7	51.8 \pm 1.5	74.8 \pm 1.0	73.1 \pm 0.2	73.8 \pm 0.3

Table 4: **Results on the COLOSSEUM Benchmark.** The table shows the success rates across 14 generalization settings. The ‘‘Avg. Rank’’ column reports the average rank of each method across all perturbations, where lower values indicate better overall performance. Compared to the state-of-the-art baseline, BridgeVLA improves the average success rate by 7.3%.

Results. Results are shown in Tab. 4. We use the results of R3M-MLP [33], MVP-MLP [46], RVT [14], and PerAct [39] from the original COLOSSEUM paper [35]. For RVT-2 [15] and BridgeVLA, we perform our own training and evaluation process. We performed three test repetitions and report the average success rate and variance of BridgeVLA and RVT-2 for each task under different perturbations in Tab. 6 and Tab. 7, respectively. BridgeVLA outperforms all the comparing baseline methods in terms of average success rate, significantly outperforming the best baseline method by 7.3%. Among all the 14 evaluated perturbations, our method ranks the best among all methods in 13 of them. These results address Q3, showcasing that BridgeVLA possesses strong robustness against visual perturbation.

B.2 Experiments on GemBench

Setup. GemBench [12] is a hierarchical generalization benchmark built on the RLBench simulator [19]. Its training set contains 16 tasks (31 variations) covering seven core action primitives—press, pick, push, screw, close, open, and stack/put. The test set consists of 44 tasks (92 variations), categorized into four increasingly challenging settings:

L1 (Novel Placements): L1 consists of the original 16 tasks (31 variations). The object placements are randomized within the workspace. In addition, chromatic distractors are introduced to test the ability to handle additional visual complexity.

L2 (Novel Rigid Objects): L2 involves 15 unseen tasks (28 variations) that require interaction with 8 novel rigid objects using learned primitives. The generalization capabilities are evaluated across two categories: novel object-color compositions and novel object shapes.

L3 (Novel Articulated Objects): L3 consists of 18 unseen tasks (21 variations) that involve interacting with articulated objects. It evaluates the generalization capabilities across three categories: novel action-part compositions, novel object instances, and novel object categories.

L4 (Novel Long-Horizon Tasks): L4 includes 6 complex long-horizon tasks (12 variations) that require combining multiple actions to finish a whole task.

Baselines. In total, we compare with six baseline methods. **3D-LOTUS** [12] processes point cloud inputs through a language-conditioned point cloud transformer architecture [45]. It showcases notable multi-tasking capabilities and high training efficiency. Its enhanced variant, **3D-LOTUS++** [12], integrates the generalization capabilities of large-scale models into 3D-LOTUS with a modular architecture consisting of three components: (1) LLM-based task planning [1], (2) VLM-based object grounding [32, 27], and (3) motion control inherited from 3D-LOTUS. We also compare with four methods introduced in Sec. 4.1.1, *i.e.*, **Hiveformer** [16], **PolarNet** [9], **3D Diffuser Actor** [25], **RVT-2** [15]

Results. Overall results are shown in Tab. 5 and per-task success rates on the four settings of GemBench are shown in Tab. 8, 9, 10, 11. The results of baseline methods are sourced from [12]. In total, we evaluate on 5 random seeds to reduce statistical variance. And for every seed, we

Method	Average	L1	L2	L3	L4
Hiveformer [16]	30.4	60.3 \pm 1.5	26.1 \pm 1.4	35.1 \pm 1.7	0.0 \pm 0.0
PolarNet [9]	38.4	77.7 \pm 0.9	37.1 \pm 1.4	38.5 \pm 1.7	0.1 \pm 0.2
3D Diffuser Actor [25]	44.0	91.9 \pm 0.8	43.4 \pm 2.8	37.0 \pm 2.2	0.0 \pm 0.0
RVT-2 [15]	44.0	89.1 \pm 0.8	51.0 \pm 2.3	36.0 \pm 2.2	0.0 \pm 0.0
3D-LOTUS [12]	45.7	94.3 \pm 1.4	49.9 \pm 2.2	38.1 \pm 1.1	0.3 \pm 0.3
3D-LOTUS++ [12]	48.0	68.7 \pm 0.6	64.5 \pm 0.9	41.5 \pm 1.8	17.4 \pm 0.4
BridgeVLA (Ours)	50.0	91.1 \pm 1.1	65.0 \pm 1.3	43.8 \pm 1.2	0.0 \pm 0.0

Table 5: **Results on GemBench.** We show the average success rates on the four evaluation settings of GemBench. BridgeVLA establishes a new state of the art on this benchmark, achieving an average success rate of 50.0%.

run 20 trials per task variation. BridgeVLA consistently outperforms all the comparing baseline methods in terms of average success rate across the four evaluation settings. Notably, BridgeVLA achieves state-of-the-art results in both the L2 and L3 settings, demonstrating strong generalization capabilities, addressing Q4. However, similar to most baseline approaches, BridgeVLA exhibits limited performance in the L4 setting, where each task comprises multiple sub-tasks. In the future, we plan to explore leveraging large language models (LLMs) for long-horizon task decomposition and further improve the performance in such setting.

B.3 Key frame Selection

For all the simulation and real-robot experiments, we adopt the same key frame selection strategy as PerAct [39]. A time step is labeled as a key frame if (i) the robot is stationary, (ii) the gripper state changes, or (iii) the step is the final state of the episode. The robot is considered stationary when the absolute velocities of all joints fall below 0.1 rad/s.

B.4 Data

Following [39, 14, 15], we select 18 tasks from RLBench [19] to evaluate the performance of our method on complex manipulation tasks. These tasks are visualized in Fig. 5.

To assess the generalization capability of BridgeVLA, we also evaluate on the COLOSSEUM benchmark [35] and GemBench [12]. The COLOSSEUM benchmark includes 20 basic tasks and 12 types of perturbations. These perturbations, which are unseen during training, encompass changes in object texture, color, and size, backgrounds, lighting, distractors and camera poses. The benchmark evaluates on all the 12 types of perturbations, a setting with basic variations from the original RLBench, and a more challenging setting which combines all the 12 types of perturbations. We visualize all perturbations except the one from the original RLBench in Fig. 6.

For GemBench, the training set includes 16 tasks (31 variations) spanning seven fundamental action primitives (press, pick, push, screw, close, open, stack/put). The test set includes 44 tasks (92 variations) organized into four increasingly challenging settings. Unlike RLBench and COLOSSEUM, where demo augmentation is used, we train BridgeVLA using only keyframes from each trajectory without performing any demo augmentation in GemBench.

C Real-Robot Experiments

C.1 Experiment Setup

Fig. 3 illustrates our real-robot setting. The platform comprises a 7-DoF Franka Research 3 manipulator with a parallel-jaw gripper and a ZED 2i stereo camera mounted on a tripod for capturing point clouds of the workspace. We collect expert trajectories with a kinesthetic teaching approach. We first move the manipulator to keypoints of an expert trajectory and then play back the keypoints to record the observation and action at each keypoint.

C.2 Basic Setting

This setting provides a scene similar to the training dataset, where only the object layouts are modified. To highlight BridgeVLA’s advantages over existing manipulation policies, we compare it with four representative methods in this setting. The behaviors of these baselines are as follows:

SpatialVLA [37]: In the experimental setup, we initially trained SpatialVLA using only 10 trajectories per task. However, it failed on nearly all tasks, often struggling to move toward the correct target object. To improve performance, we augmented the dataset with an additional 40 trajectories per task. While this improved performance, it still lagged significantly behind BridgeVLA—particularly on more challenging tasks, such as "Put the giraffe in the lower drawer." These findings suggest that BridgeVLA provides a more effective and data-efficient solution for 3D VLA.

π_0 [5]: Similarly, π_0 fails with only 10 trajectories per task, likely due to overfitting—it performs well on the training set but often fails during online testing. Common failure modes include missing or failing to grasp the target and prematurely opening the gripper before reaching the goal. Notably, both BridgeVLA and π_0 share the same PaliGemma backbone and are trained end-to-end. This highlights a key contribution of our work: while VLAs like π_0 perform well with large-scale data, they struggle in low-data regimes—even on simpler tasks, such as "Press sanitizer." In contrast, BridgeVLA achieves near-perfect success and generalizes robustly across diverse settings.

ACT [51]: ACT also underperforms compared to BridgeVLA. It demonstrates limited spatial generalization, performing well only in areas densely covered during training, but often failing when the target is near the workspace boundaries. This behavior is consistent with its design: ACT models actions using a Gaussian prior, which assigns low probability to peripheral regions, limiting its spatial generalization capabilities.

RVT-2 [15]: RVT-2 performs the best among all the baselines. It can successfully solve most tasks, but it is not as robust as BridgeVLA. For instance, it sometimes fails to pick up the block precisely or place the object accurately, leading to task failure. Meanwhile, by utilizing the capabilities of VLM, BridgeVLA’s advantages are further amplified in generalization settings, as detailed in Sec. 4.2.

C.3 Generalization Settings

We evaluate on a total of six generalization settings: Distractor, Lighting, Background, Height, Combination, and Category. For Distractor, Lighting, Background, and Height, we visualize these settings in Fig. 10. We visualize the settings of Combination and Category in Fig. 11 and Fig. 12 respectively.

In Distractor, we add distractor objects that are visually similar to at least one target object to the scene. In Lighting, we evaluate the model in a novel lighting condition in which the lights are off. In Background, we use three different tablecloths to change the background. For Height, we evaluate all objects for manipulation with a drawer that is about 10cm high. Distractor, Lighting, Background, and Height aim to evaluate the robustness against visual disturbances.

In Combination, we combine objects and skills that are not paired together in the training datasets. That is, while the object for manipulation and the manipulation skill are seen during training, the instruction that pairs them together is novel. The setting of Combination helps us evaluate whether the model is able to generalize across novel object-skill combinations. In Category, we want to evaluate whether BridgeVLA is able to manipulate objects from categories that are *unseen* in the robot training data. In total, we test 7 novel objects.

C.4 Preservation of Object Grounding Capability after Fine-tuning

We observe that even after fine-tuning on robot action data, BridgeVLA retains the object grounding capability learned during pre-training. We visualize its predictions on the pre-training dataset after fine-tuning in Fig. 14. It is important to note that the samples in Fig. 14 are not cherry-picked. BridgeVLA does not forget its pre-training knowledge after 3D action fine-tuning.

C.5 Per-task Success Rate

We showcase per-task success rates of BridgeVLA in the basic setting in Tab. [12](#). Notably, BridgeVLA achieves exceptionally high success rates even with only 3 trajectories per task, highlighting its superb sample efficiency.

Task Name	Original	All Perturbations	MO-COLOR	RO-COLOR	MO-TEXTURE	RO-TEXTURE	MO-SIZE	RO-SIZE	Light Color	Table Color	Table Texture	Distractor	Background Texture	RL Bench	Camera Pose
basketball_in_hoop	100.0±0.0	4.0±3.3	94.7±1.9	96.0±0.0	84.0±5.7	-	100.0±0.0	68.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	37.3±1.9	100.0±0.0	100.0±0.0	100.0±0.0
close_box	100.0±0.0	72.0±0.0	94.7±1.9	-	-	-	93.3±1.9	-	100.0±0.0	100.0±0.0	98.7±1.9	98.7±1.9	100.0±0.0	97.3±1.9	100.0±0.0
close_laptop_lid	100.0±0.0	11.1±15.7	82.7±3.8	-	-	-	67.9±14.6	-	89.3±8.2	92.0±0.0	97.3±3.8	82.7±6.8	96.0±3.3	100.0±0.0	96.0±0.0
empty_dishwasher	0.0±0.0	0.0±0.0	1.3±1.9	1.3±1.9	-	1.3±1.9	4.0±3.3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	1.3±1.9	1.3±1.9	0.0±0.0
get_ice_from_fridge	94.7±1.9	5.3±1.9	86.7±1.9	90.7±7.5	90.7±5.0	-	84.0±3.3	73.3±1.9	96.0±3.3	98.7±1.9	89.3±7.5	56.0±8.6	94.7±1.9	96.0±3.3	98.7±1.9
hockey	57.3±5.0	9.3±3.8	44.0±6.5	50.7±8.2	-	50.7±13.2	46.7±8.2	65.3±5.0	45.3±1.9	64.0±8.6	53.3±1.9	20.0±3.3	56.0±5.7	49.3±5.0	50.7±5.0
insert_onto_square_peg	93.3±3.8	23.3±2.4	52.0±3.3	94.7±1.9	-	76.0±8.6	85.3±3.8	70.7±3.8	84.0±0.0	88.0±3.3	88.0±3.3	44.0±11.8	86.7±1.9	77.3±5.0	96.0±0.0
meat_on_grill	96.0±0.0	9.3±1.9	32.0±0.0	88.0±5.7	-	-	100.0±0.0	-	100.0±0.0	92.0±6.5	90.7±1.9	98.7±1.9	97.3±1.9	100.0±0.0	100.0±0.0
move_hanger	37.3±3.8	2.7±3.8	26.7±3.8	46.7±3.8	-	-	-	-	52.0±0.0	84.0±0.0	52.0±5.7	52.0±5.7	33.3±5.0	42.7±1.9	24.0±0.0
open_drawer	96.0±0.0	60.0±3.3	97.3±1.9	-	-	-	90.7±1.9	-	88.0±3.3	93.3±1.9	100.0±0.0	90.7±1.9	100.0±0.0	94.7±1.9	96.0±0.0
place_wine_at_rack_location	88.0±5.7	17.3±13.6	82.7±5.0	89.3±7.5	-	92.0±6.5	93.3±3.8	90.7±3.8	90.7±5.0	97.3±1.9	88.0±3.3	74.7±3.8	90.7±6.8	92.0±3.3	92.0±8.6
put_money_in_safe	94.7±1.9	6.7±5.0	78.7±1.9	74.7±1.9	81.3±6.8	89.3±5.0	92.0±3.3	-	37.3±12.4	84.0±3.3	84.0±3.3	84.0±3.3	89.3±1.9	86.7±8.2	86.7±1.9
reach_and_drag	100.0±0.0	0.0±0.0	89.3±3.8	96.0±0.0	94.7±5.0	84.0±5.7	94.7±1.9	38.7±5.0	92.0±3.3	88.0±5.7	78.7±3.8	28.0±8.6	100.0±0.0	100.0±0.0	94.7±3.8
scoop_with_spatula	96.0±3.3	6.7±1.9	94.7±1.9	93.3±1.9	85.3±3.8	85.3±3.8	78.7±3.8	86.7±5.0	90.7±1.9	88.0±6.5	77.3±1.9	20.0±5.7	90.7±6.8	89.3±1.9	93.3±1.9
setup_chess	10.7±1.9	0.0±0.0	1.3±1.9	8.0±0.0	8.0±3.3	-	13.3±1.9	-	12.0±5.7	21.3±8.2	13.3±3.8	5.3±1.9	20.0±5.7	16.0±5.7	4.0±3.3
slide_block_to_target	100.0±0.0	24.0±3.3	74.7±1.9	-	92.0±3.3	-	-	-	100.0±0.0	100.0±0.0	98.7±1.9	84.0±9.8	100.0±0.0	100.0±0.0	100.0±0.0
stack_cups	58.7±3.8	29.3±1.9	66.7±1.9	-	50.7±1.9	-	44.0±3.3	-	62.7±1.9	64.0±3.3	65.3±8.2	26.7±7.5	73.3±8.2	64.0±14.2	72.0±8.6
straighten_ropes	61.3±6.8	8.0±5.7	16.0±5.7	-	48.0±3.3	-	-	-	61.3±9.4	65.3±1.9	54.7±8.2	37.3±5.0	70.7±8.2	66.7±7.5	72.0±6.5
turn_oven_on	93.3±1.9	85.3±3.8	94.7±3.8	-	-	-	90.7±1.9	-	93.3±3.8	94.7±7.5	96.0±3.3	96.0±3.3	96.0±0.0	88.0±3.3	100.0±0.0
wipe_desk	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±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Task Mean	73.9±0.7	18.7±2.2	60.5±1.1	63.8±0.1	63.5±1.5	68.4±3.3	69.3±1.0	61.7±0.8	69.7±1.2	75.7±0.9	71.3±0.7	51.8±1.5	74.8±1.0	73.1±0.2	73.8±0.3

Table 6: Success Rates of BridgeVLA under Different Perturbations of COLOSSEUM.

Task Name	No variations	All Perturbations	MO-COLOR	RO-COLOR	MO-TEXTURE	RO-TEXTURE	MO-SIZE	RO-SIZE	Light Color	Table Color	Table Texture	Distractor	Background Texture	RL Bench	Camera Pose
basketball_in_hoop	100±0.0	40±4.9	99±1.7	33±1.7	96±0.0	-	99±1.7	100±0.0	87±4.4	54±2.0	91±11.4	55±9.1	100±0.0	97±1.7	100±0.0
close_box	96±4.9	32±16.7	43±3.3	-	-	-	91±5.2	-	84±2.8	78±3.5	91±9.1	96±2.8	98±3.5	96±2.8	95±3.3
close_laptop_lid	30±4.5	48±7.5	50±4.5	-	-	-	56±6.9	-	28±2.8	23±4.4	42±12.8	49±5.2	44±0.0	33±1.7	48±2.8
empty_dishwasher	0±0.0	0±0.0	0±0.0	0±0.0	-	1±1.7	1±1.7	1±1.7	0±0.0	0±0.0	0±0.0	0±0.0	0±0.0	1±1.7	1±1.7
get_ice_from_fridge	66±4.5	2±2.0	67±3.3	11±1.7	67±5.2	-	71±3.3	44±2.8	24±0.0	35±1.7	77±4.4	65±3.3	70±3.5	69±1.7	71±4.4
hockey	12±2.8	0±0.0	18±6.0	0±0.0	-	14±3.5	2±3.5	9±3.3	7±5.9	10±2.0	16±15.0	5±1.7	13±1.7	5±1.7	9±4.4
meat_on_grill	45±1.7	56±8.5	62±4.5	33±1.7	-	-	64±2.8	-	61±1.7	65±1.7	49±5.2	67±11.1	63±3.3	62±4.5	51±3.3
move_hanger	0±0.0	0±0.0	0±0.0	0±0.0	-	-	-	-	0±0.0	0±0.0	0±0.0	21±9.1	0±0.0	0±0.0	0±0.0
wipe_desk	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	0±0.0	0±0.0	0±0.0
open_drawer	97±1.7	9±5.9	100±0.0	-	-	-	100±0.0	-	90±4.5	56±0.0	100±0.0	89±5.2	100±0.0	97±1.7	96±0.0
slide_block_to_target	100±0.0	37±10.7	100±0.0	-	100±0.0	-	-	-	100±0.0	91±1.7	88±20.8	90±12.8	100±0.0	100±0.0	100±0.0
reach_and_drag	86±2.0	1±1.7	34±2.0	64±2.8	75±4.4	74±4.5	95±1.7	79±1.7	20±0.0	24±0.0	72±15.7	43±21.4	81±1.7	75±1.7	80±2.8
put_money_in_safe	63±1.7	1±1.7	62±2.0	5±1.7	58±2.0	75±1.7	64±2.8	-	60±0.0	47±3.3	82±4.5	60±4.9	60±0.0	60±2.8	48±0.0
place_wine_at_rack_location	96±4.9	59±11.4	94±4.5	94±2.0	-	96±2.8	91±1.7	88±2.8	87±5.9	93±4.4	94±2.0	80±16.0	88±2.8	95±3.3	96±4.9
insert_onto_square_peg	5±1.7	0±0.0	0±0.0	13±1.7	-	9±3.3	16±0.0	6±3.5	0±0.0	0±0.0	0±0.0	2±3.5	4±0.0	4±0.0	5±1.7
stack_cups	44±0.0	2±3.5	42±2.0	-	50±4.5	-	8±0.0	-	20±0.0	13±1.7	10±6.0	15±3.3	40±0.0	36±0.0	24±0.0
turn_oven_on	97±1.7	5±1.7	34±4.5	-	-	-	68±2.8	-	96±2.8	97±1.7	98±2.0	97±1.7	96±2.8	92±0.0	93±5.2
straighten_ropes	54±2.0	0±0.0	32±0.0	-	57±4.4	-	-	-	77±1.7	51±4.4	14±11.8	27±20.3	61±1.7	66±2.0	59±1.7
setup_chess	5±3.3	0±0.0	1±1.7	4±2.8	7±3.3	-	4±2.8	-	8±4.9	4±2.8	12±2.8	10±10.4	18±8.2	15±5.2	7±4.4
scoop_with_spatula	96±0.0	0±0.0	11±1.7	73±3.3	82±24.2	85±1.7	81±1.7	81±5.2	57±1.7	87±3.3	83±4.4	57±11.8	93±1.7	85±3.3	92±2.8
Average	55±0.5	15±1.9	42±0.6	25±0.2	59±2.6	51±1.8	54±1.0	51±1.3	45±0.3	41±0.0	51±2.7	46±2.2	56±0.5	54±0.5	54±0.4

Table 7: Success Rates of RVT-2 under Different Perturbations of COLOSSEUM.

Method	Avg.	Close Fridge+0	Close Jar+15	Close Jar+16	CloseLaptop Lid+0	Close Microwave+0	LightBulb In+17	LightBulb In+19	Open Box+0	Open Door+0	Open Drawer+0
Hiveformer [16]	60.3 \pm 1.5	96 \pm 4.2	64 \pm 13.9	92 \pm 2.7	90 \pm 3.5	88 \pm 7.6	12 \pm 4.5	13 \pm 6.7	4 \pm 4.2	53 \pm 15.2	15 \pm 12.2
PolarNet [9]	77.6 \pm 0.9	99 \pm 2.2	99 \pm 2.2	99 \pm 2.2	95 \pm 3.5	98 \pm 2.7	72 \pm 12.5	71 \pm 6.5	32 \pm 11.5	69 \pm 8.9	61 \pm 12.4
3D diffuser actor [25]	91.9 \pm 0.8	100 \pm 0.0	100 \pm 0.0	100 \pm 0.0	99 \pm 2.2	100 \pm 0.0	85 \pm 5.0	88 \pm 2.7	11 \pm 2.2	96 \pm 4.2	82 \pm 9.1
RVT-2 [15]	89.0 \pm 0.8	77 \pm 11.0	97 \pm 4.5	98 \pm 2.7	77 \pm 13.0	100 \pm 0.0	93 \pm 5.7	91 \pm 8.2	7 \pm 4.5	98 \pm 4.5	93 \pm 5.7
3D-LOTUS [12]	94.3 \pm 3.5	96 \pm 3.7	100 \pm 0.0	100 \pm 0.0	98 \pm 2.5	98 \pm 4.0	84 \pm 7.4	85 \pm 9.5	99 \pm 2.0	77 \pm 2.5	83 \pm 8.7
3D-LOTUS++ [12]	68.7 \pm 0.6	95 \pm 0.0	100 \pm 0.0	99 \pm 2.0	28 \pm 2.5	87 \pm 5.1	55 \pm 10.5	45 \pm 8.9	55 \pm 8.9	79 \pm 9.7	68 \pm 12.5
BridgeVLA (Ours)	91.1 \pm 1.1	99 \pm 2.0	98 \pm 4.0	100 \pm 0.0	97 \pm 2.5	85 \pm 5.5	90 \pm 5.5	87 \pm 7.5	76 \pm 10.2	70 \pm 12.3	86 \pm 5.8
Method	Open Drawer+2	Pick& Lift+0	Pick& Lift+2	Pick& Lift+7	PickUp Cup+8	PickUp Cup+9	PickUp Cup+11	Push Button+0	Push Button+3	Push Button+4	PutIn Cupboard+0
Hiveformer [16]	59 \pm 7.4	86 \pm 4.2	92 \pm 6.7	93 \pm 2.7	83 \pm 7.6	69 \pm 12.9	61 \pm 19.8	84 \pm 11.9	68 \pm 6.7	87 \pm 7.6	34 \pm 8.2
PolarNet [9]	90 \pm 7.1	92 \pm 9.1	84 \pm 7.4	88 \pm 5.7	82 \pm 7.6	79 \pm 4.2	72 \pm 10.4	100 \pm 0.0	100 \pm 0.0	99 \pm 2.2	52 \pm 7.6
3D diffuser actor [25]	97 \pm 4.5	99 \pm 2.2	99 \pm 2.2	99 \pm 2.2	96 \pm 2.2	97 \pm 4.5	98 \pm 2.7	98 \pm 2.7	96 \pm 4.2	98 \pm 2.7	85 \pm 5.0
RVT-2 [15]	94 \pm 4.2	99 \pm 2.2	98 \pm 2.7	100 \pm 0.0	99 \pm 2.2	99 \pm 2.2	99 \pm 2.2	100 \pm 0.0	100 \pm 0.0	100 \pm 0.0	88 \pm 8.4
3D-LOTUS [12]	93 \pm 6.0	99 \pm 2.0	100 \pm 0.0	99 \pm 2.0	97 \pm 4.0	96 \pm 3.7	94 \pm 4.9	99 \pm 2.0	99 \pm 2.0	100 \pm 0.0	89 \pm 5.8
3D-LOTUS++ [12]	75 \pm 4.5	97 \pm 6.0	94 \pm 3.7	93 \pm 5.1	86 \pm 8.0	88 \pm 6.8	91 \pm 4.9	100 \pm 0.0	100 \pm 0.0	100 \pm 0.0	1 \pm 2.0
BridgeVLA (Ours)	99 \pm 2.0	99 \pm 2.0	100 \pm 0.0	98 \pm 2.5	96 \pm 2.0	94 \pm 3.7	99 \pm 2.0	100 \pm 0.0	98 \pm 4.0	98 \pm 4.0	74 \pm 6.6
Method	PutIn Cupboard+3	PutMoney InSafe+0	PutMoney InSafe+1	Reach& Drag+14	Reach& Drag+18	Slide Block+0	Slide Block+1	Stack Blocks+30	Stack Blocks+36	Stack Blocks+39	
Hiveformer [16]	74 \pm 6.5	85 \pm 3.5	88 \pm 2.7	37 \pm 5.7	32 \pm 7.6	99 \pm 2.2	91 \pm 12.4	6 \pm 5.5	7 \pm 4.5	6 \pm 4.2	
PolarNet [9]	88 \pm 4.5	93 \pm 4.5	95 \pm 5.0	99 \pm 2.2	99 \pm 2.2	100 \pm 0.0	0 \pm 0.0	34 \pm 10.8	30 \pm 9.4	36 \pm 12.9	
3D diffuser actor [25]	82 \pm 11.5	95 \pm 5.0	98 \pm 2.7	100 \pm 0.0	99 \pm 2.2	100 \pm 0.0	89 \pm 4.2	88 \pm 7.6	85 \pm 6.1	89 \pm 5.5	
RVT-2 [15]	80 \pm 6.1	93 \pm 8.4	96 \pm 8.5	85 \pm 10.0	94 \pm 2.2	100 \pm 0.0	37 \pm 6.7	88 \pm 5.7	93 \pm 2.7	88 \pm 11.5	
3D-LOTUS [12]	72 \pm 11.2	94 \pm 3.7	99 \pm 2.0	99 \pm 2.0	100 \pm 0.0	100 \pm 0.0	100 \pm 0.0	94 \pm 5.8	91 \pm 6.6	90 \pm 4.5	
3D-LOTUS++ [12]	2 \pm 2.5	22 \pm 6.8	16 \pm 4.9	94 \pm 3.7	62 \pm 8.7	100 \pm 0.0	65 \pm 5.5	86 \pm 5.8	20 \pm 4.5	28 \pm 13.6	
BridgeVLA (Ours)	84 \pm 6.6	79 \pm 9.7	86 \pm 3.7	96 \pm 5.8	97 \pm 4.0	100 \pm 0.0	90 \pm 5.5	77 \pm 8.1	87 \pm 4.0	85 \pm 7.8	

Table 8: Per-task Success Rate on GemBench Level 1.

Method	Avg.	Push Button+13	Push Button+15	Push Button+17	Pick& Lift+14	Pick& Lift+16	Pick& Lift+18	PickUp Cup+10	PickUp Cup+12	PickUp Cup+13
Hiveformer	26.1 \pm 1.4	97 \pm 2.7	85 \pm 10.0	88 \pm 2.7	21 \pm 6.5	9 \pm 4.2	8 \pm 6.7	30 \pm 7.1	22 \pm 13.5	26 \pm 10.6
PolarNet	37.1 \pm 1.4	100 \pm 0.0	100 \pm 0.0	85 \pm 7.9	3 \pm 4.5	1 \pm 2.2	0 \pm 0.0	48 \pm 11.0	46 \pm 8.9	16 \pm 6.5
3D diffuser actor	43.4 \pm 2.8	87 \pm 13.0	81 \pm 6.5	60 \pm 9.4	9 \pm 4.2	18 \pm 9.1	0 \pm 0.0	84 \pm 5.5	60 \pm 11.7	62 \pm 13.0
RVT-2	51.0 \pm 2.3	100 \pm 0.0	100 \pm 0.0	100 \pm 0.0	47 \pm 7.6	29 \pm 9.6	8 \pm 4.5	81 \pm 8.2	59 \pm 9.6	72 \pm 9.7
3D-LOTUS	49.9 \pm 2.2	99 \pm 2.0	100 \pm 0.0	100 \pm 0.0	3 \pm 2.5	18 \pm 8.7	33 \pm 9.3	89 \pm 3.7	78 \pm 8.7	57 \pm 7.5
3D-LOTUS++	64.5 \pm 0.9	99 \pm 2.0	100 \pm 0.0	99 \pm 2.0	94 \pm 3.7	96 \pm 3.7	95 \pm 3.2	79 \pm 4.9	89 \pm 9.7	84 \pm 10.2
BridgeVLA (Ours)	65.0 \pm 1.3	100 \pm 0.0	100 \pm 0.0	100 \pm 0.0	74 \pm 9.7	89 \pm 4.9	0 \pm 0.0	91 \pm 3.7	90 \pm 3.2	90 \pm 6.3
Method	Stack Blocks+24	Stack Blocks+27	Stack Blocks+33	Slide Block+2	Slide Block+3	Close Jar+3	Close Jar+4	LightBulb In+1	LightBulb In+2	Lamp On+0
Hiveformer	0 \pm 0.0	4 \pm 4.2	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	4 \pm 4.2	0 \pm 0.0	7 \pm 4.5
PolarNet	1 \pm 2.2	2 \pm 2.7	6 \pm 8.2	0 \pm 0.0	0 \pm 0.0	20 \pm 10.6	82 \pm 5.7	22 \pm 11.5	17 \pm 8.4	14 \pm 10.8
3D diffuser actor	66 \pm 13.9	82 \pm 2.7	50 \pm 14.6	0 \pm 0.0	0 \pm 0.0	23 \pm 16.8	82 \pm 5.7	51 \pm 17.8	60 \pm 10.0	7 \pm 7.6
RVT-2	18 \pm 4.5	56 \pm 16.7	45 \pm 13.7	0 \pm 0.0	1 \pm 2.2	7 \pm 7.6	77 \pm 5.7	68 \pm 14.4	6 \pm 6.5	0 \pm 0.0
3D-LOTUS	13 \pm 8.1	40 \pm 9.5	69 \pm 5.8	0 \pm 0.0	0 \pm 0.0	71 \pm 5.8	90 \pm 4.5	24 \pm 4.9	41 \pm 8.6	0 \pm 0.0
3D-LOTUS++	22 \pm 9.3	83 \pm 7.5	59 \pm 3.7	27 \pm 9.8	5 \pm 3.2	98 \pm 2.5	96 \pm 3.7	56 \pm 9.7	43 \pm 7.5	2 \pm 2.0
BridgeVLA (Ours)	61 \pm 10.7	51 \pm 13.2	79 \pm 8.6	12 \pm 9.3	3 \pm 4.0	66 \pm 6.6	88 \pm 4.0	66 \pm 8.6	74 \pm 5.8	7 \pm 4.0
Method	Reach& Drag+5	Reach& Drag+7	PutCube InSafe+0	Pick&Lift Cylinder+0	Pick&Lift Star+0	Pick&Lift Moon+0	Pick&Lift Toy+0	PutIn Cupboard+7	PutIn Cupboard+8	
Hiveformer	1 \pm 2.2	0 \pm 0.0	4 \pm 2.2	78 \pm 5.7	73 \pm 7.6	88 \pm 2.7	87 \pm 4.5	0 \pm 0.0	0 \pm 0.0	
PolarNet	61 \pm 8.2	10 \pm 6.1	40 \pm 14.1	93 \pm 6.7	88 \pm 8.4	93 \pm 6.7	90 \pm 3.5	0 \pm 0.0	0 \pm 0.0	
3D diffuser actor	0 \pm 0.0	64 \pm 6.5	3 \pm 2.7	99 \pm 2.2	43 \pm 17.9	91 \pm 9.6	30 \pm 9.4	0 \pm 0.0	3 \pm 4.5	
RVT-2	91 \pm 2.2	89 \pm 6.5	6 \pm 5.5	98 \pm 2.7	98 \pm 4.5	94 \pm 4.2	78 \pm 8.4	0 \pm 0.0	0 \pm 0.0	
3D-LOTUS	95 \pm 4.5	18 \pm 10.8	25 \pm 5.5	88 \pm 8.7	69 \pm 6.6	80 \pm 8.4	96 \pm 3.7	0 \pm 0.0	0 \pm 0.0	
3D-LOTUS++	94 \pm 2.0	64 \pm 12.4	37 \pm 5.1	91 \pm 2.0	94 \pm 3.7	29 \pm 6.6	71 \pm 2.0	1 \pm 2.0	0 \pm 0.0	
BridgeVLA (Ours)	94 \pm 3.7	96 \pm 3.7	3 \pm 2.5	98 \pm 2.5	99 \pm 2.0	95 \pm 3.2	93 \pm 5.1	0 \pm 0.0	0 \pm 0.0	

Table 9: Per-task Success Rate on GemBench Level 2.

Method	Avg.	Close Door+0	Close Box+0	Close Fridge2+0	CloseLaptop Lid2+0	Close Microwave2+0	Open Door2+0	Open Box2+0
Hiveformer	35.1 \pm 1.7	0 \pm 0.0	1 \pm 2.2	34 \pm 9.6	52 \pm 9.1	15 \pm 7.1	32 \pm 11.5	5 \pm 3.5
PolarNet	38.5 \pm 1.7	0 \pm 0.0	0 \pm 0.0	78 \pm 5.7	26 \pm 8.2	74 \pm 6.5	33 \pm 6.7	23 \pm 8.4
3D diffuser actor	37.0 \pm 2.2	0 \pm 0.0	0 \pm 0.0	97 \pm 2.7	23 \pm 6.7	88 \pm 7.6	86 \pm 7.4	67 \pm 9.8
RVT-2	36.0 \pm 2.2	1 \pm 2.2	2 \pm 2.7	72 \pm 6.7	42 \pm 14.0	71 \pm 8.9	79 \pm 6.5	5 \pm 6.1
3D-LOTUS	38.1 \pm 1.1	0 \pm 0.0	58 \pm 8.1	36 \pm 9.7	54 \pm 10.7	85 \pm 7.1	42 \pm 6.8	11 \pm 6.6
3D-LOTUS++	41.5 \pm 1.8	1 \pm 2.0	29 \pm 8.6	93 \pm 2.5	50 \pm 9.5	99 \pm 2.0	52 \pm 10.3	16 \pm 8.0
BridgeVLA (Ours)	43.8 \pm 1.2	0 \pm 0.0	1 \pm 2.0	95 \pm 5.5	77 \pm 4.0	54 \pm 10.2	68 \pm 10.8	74 \pm 4.9
Method	Open Drawer2+0	Open Drawer3+0	OpenDrawer Long+0	OpenDrawer Long+1	OpenDrawer Long+2	OpenDrawer Long+3	Toilet SeatUp+0	Open Fridge+0
Hiveformer	59 \pm 11.9	39 \pm 11.9	78 \pm 8.4	82 \pm 4.5	49 \pm 4.2	57 \pm 11.5	6 \pm 4.2	0 \pm 0.0
PolarNet	91 \pm 4.2	29 \pm 8.2	84 \pm 11.9	88 \pm 5.7	63 \pm 8.4	37 \pm 7.6	2 \pm 2.7	4 \pm 2.2
3D diffuser actor	19 \pm 8.2	1 \pm 2.2	15 \pm 5.0	35 \pm 13.7	26 \pm 9.6	79 \pm 12.9	0 \pm 0.0	7 \pm 5.7
RVT-2	81 \pm 11.9	0 \pm 0.0	84 \pm 8.2	39 \pm 10.8	11 \pm 8.9	75 \pm 6.1	7 \pm 5.7	0 \pm 0.0
3D-LOTUS	90 \pm 3.2	22 \pm 8.1	56 \pm 13.9	33 \pm 11.2	17 \pm 8.1	75 \pm 6.3	0 \pm 0.0	4 \pm 5.8
3D-LOTUS++	70 \pm 5.5	41 \pm 4.9	72 \pm 4.0	52 \pm 10.8	23 \pm 8.1	78 \pm 5.1	8 \pm 5.1	0 \pm 0.0
BridgeVLA (Ours)	65 \pm 6.3	87 \pm 6.0	59 \pm 8.6	34 \pm 8.0	18 \pm 10.3	85 \pm 8.4	6 \pm 5.8	7 \pm 2.5
Method	OpenLaptop Lid+0	Open Microwave+0	PutMoney InSafe+2	Open Drawer+1	Close Drawer+0	Close Grill+0		
Hiveformer	100 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	83 \pm 5.7	44 \pm 10.8		
PolarNet	100 \pm 0.0	0 \pm 0.0	1 \pm 2.2	4 \pm 4.2	29 \pm 11.9	42 \pm 11.5		
3D diffuser actor	100 \pm 0.0	0 \pm 0.0	2 \pm 4.5	0 \pm 0.0	66 \pm 7.4	65 \pm 13.7		
RVT-2	93 \pm 5.7	0 \pm 0.0	0 \pm 0.0	6 \pm 2.2	78 \pm 8.4	9 \pm 4.2		
3D-LOTUS	100 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	87 \pm 8.1	29 \pm 6.6		
3D-LOTUS++	86 \pm 6.6	0 \pm 0.0	13 \pm 8.1	0 \pm 0.0	69 \pm 5.8	19 \pm 13.9		
BridgeVLA (Ours)	95 \pm 0.0	0 \pm 0.0	2 \pm 2.5	0 \pm 0.0	58 \pm 12.9	35 \pm 12.3		

Table 10: Per-task Success Rate on GemBench Level 3.

Method	Avg.	Push Buttons4+1	Push Buttons4+2	Push Buttons4+3	TakeShoes OutOfBox+0	PutItems InDrawer+0	PutItems InDrawer+2
Hiveformer	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0
PolarNet	0.1 \pm 0.2	1 \pm 2.2	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0
3D diffuser actor	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0
RVT-2	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0
3D-LOTUS	0.3 \pm 0.3	3 \pm 4.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0
3D-LOTUS++	17.4 \pm 0.4	76 \pm 7.4	49 \pm 8.6	37 \pm 8.1	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0
BridgeVLA (Ours)	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0
Method	PutItems InDrawer+4	Tower4+1	Tower4+3	Stack Cups+0	Stack Cups+3	PutAllGroceries InCupboard+0	
Hiveformer	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	
PolarNet	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	
3D diffuser actor	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	
RVT-2	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	
3D-LOTUS	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	
3D-LOTUS++	0 \pm 0.0	17 \pm 10.8	30 \pm 13.4	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	
BridgeVLA (Ours)	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	0 \pm 0.0	

Table 11: Per-task Success Rate on GemBench Level 4.



Figure 5: Visualization of 18 RL Bench [19] Tasks.

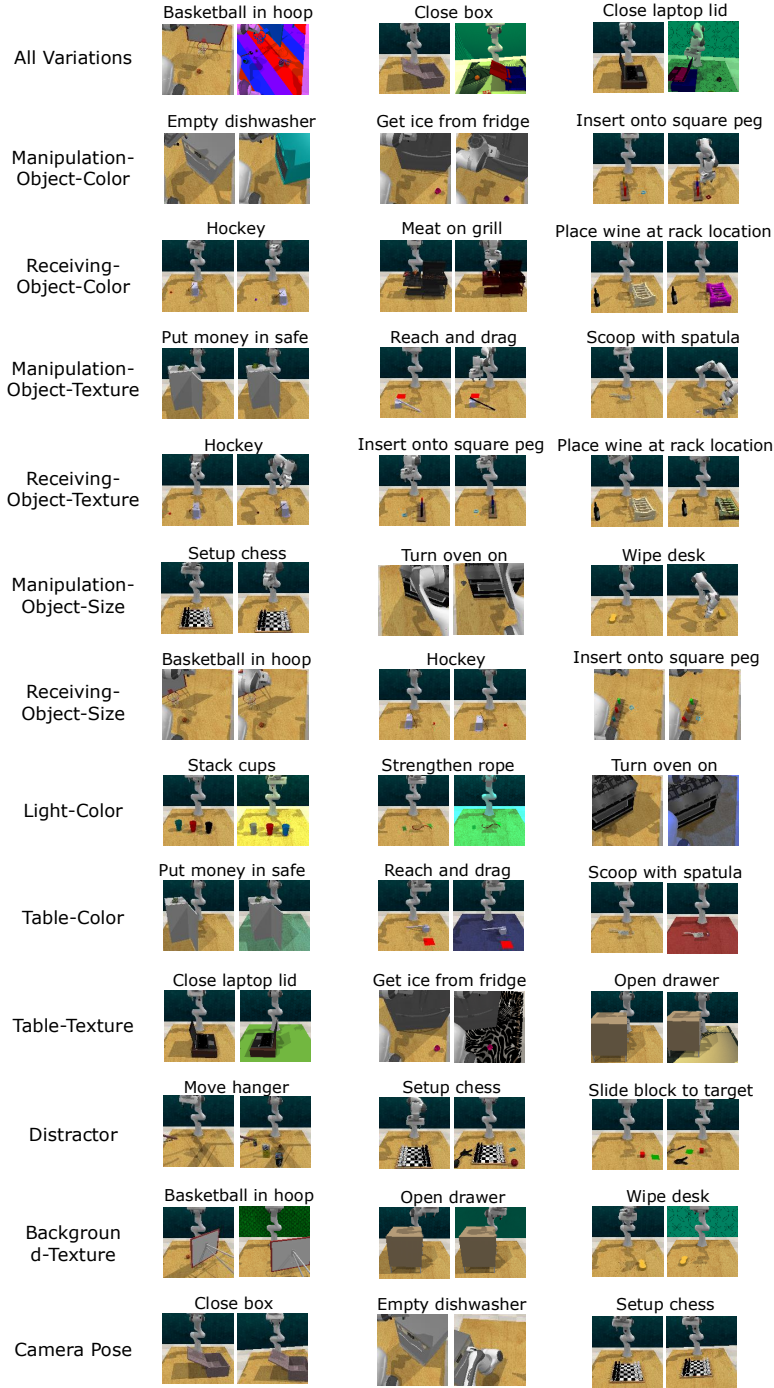


Figure 6: Visualization of Perturbations in COLOSSEUM [35].

Task	3 trajectories	10 trajectories
Put the RedBull can in the top shelf	9/10	10/10
Put the soda can in the bottom shelf	9/10	9/10
Put the RedBull can in the bottom shelf	10/10	10/10
Put the coke can in the top shelf	10/10	10/10
Place the red block in the blue plate	10/10	10/10
Place the orange block in the green plate	10/10	10/10
Put the wolf in the upper drawer	7/10	9/10
Place the red block in the purple plate	10/10	10/10
Place the yellow block in the green plate	10/10	10/10
Press sanitizer	10/10	10/10
Put the zebra in the upper drawer	9/10	9/10
Put the giraffe in the lower drawer	10/10	9/10
Put the zebra in the lower drawer	10/10	10/10

Table 12: **Per-task Success Rates of BridgeVLA in the Basic Setting.**



Figure 7: **Real-Robot Rollouts (I).**

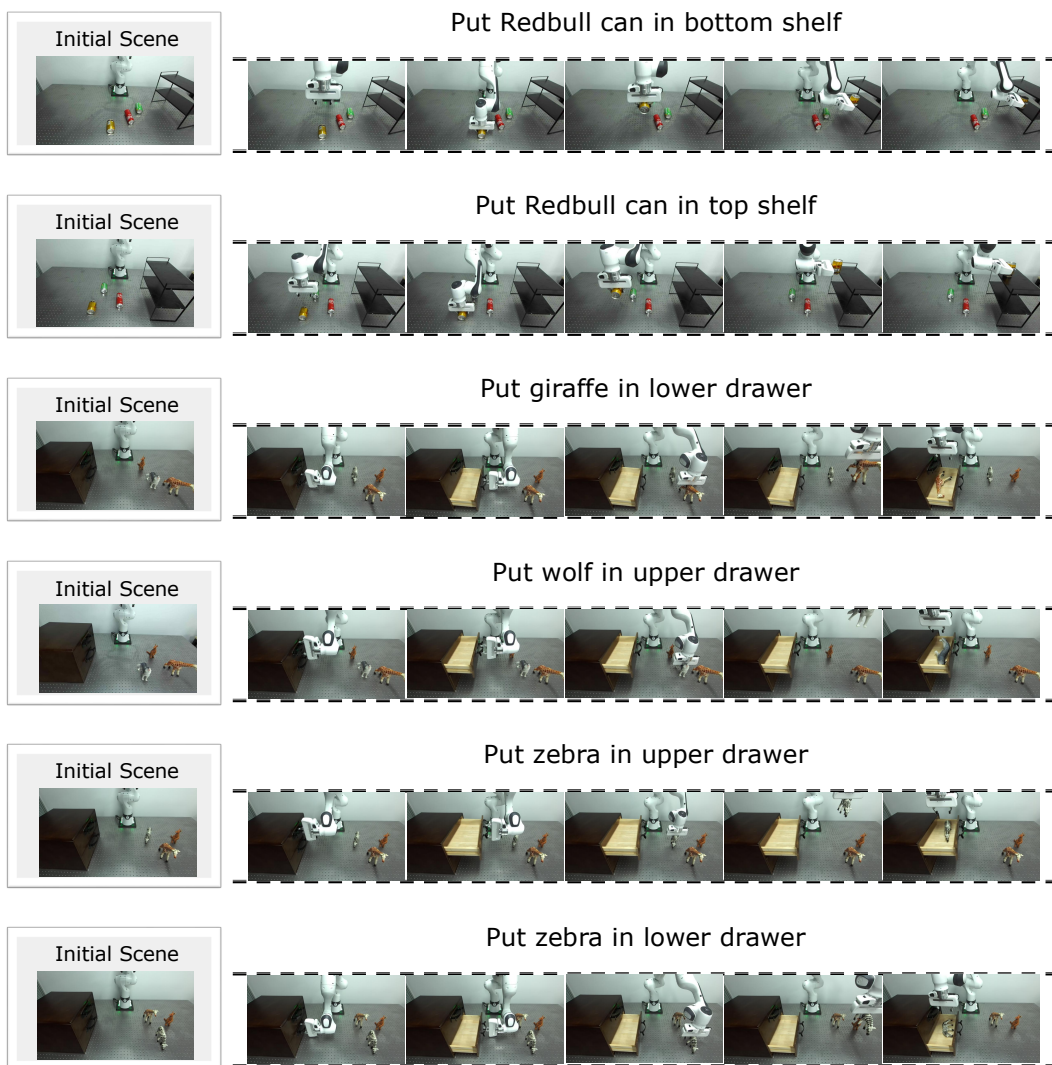


Figure 8: **Real-Robot Rollouts (II).**

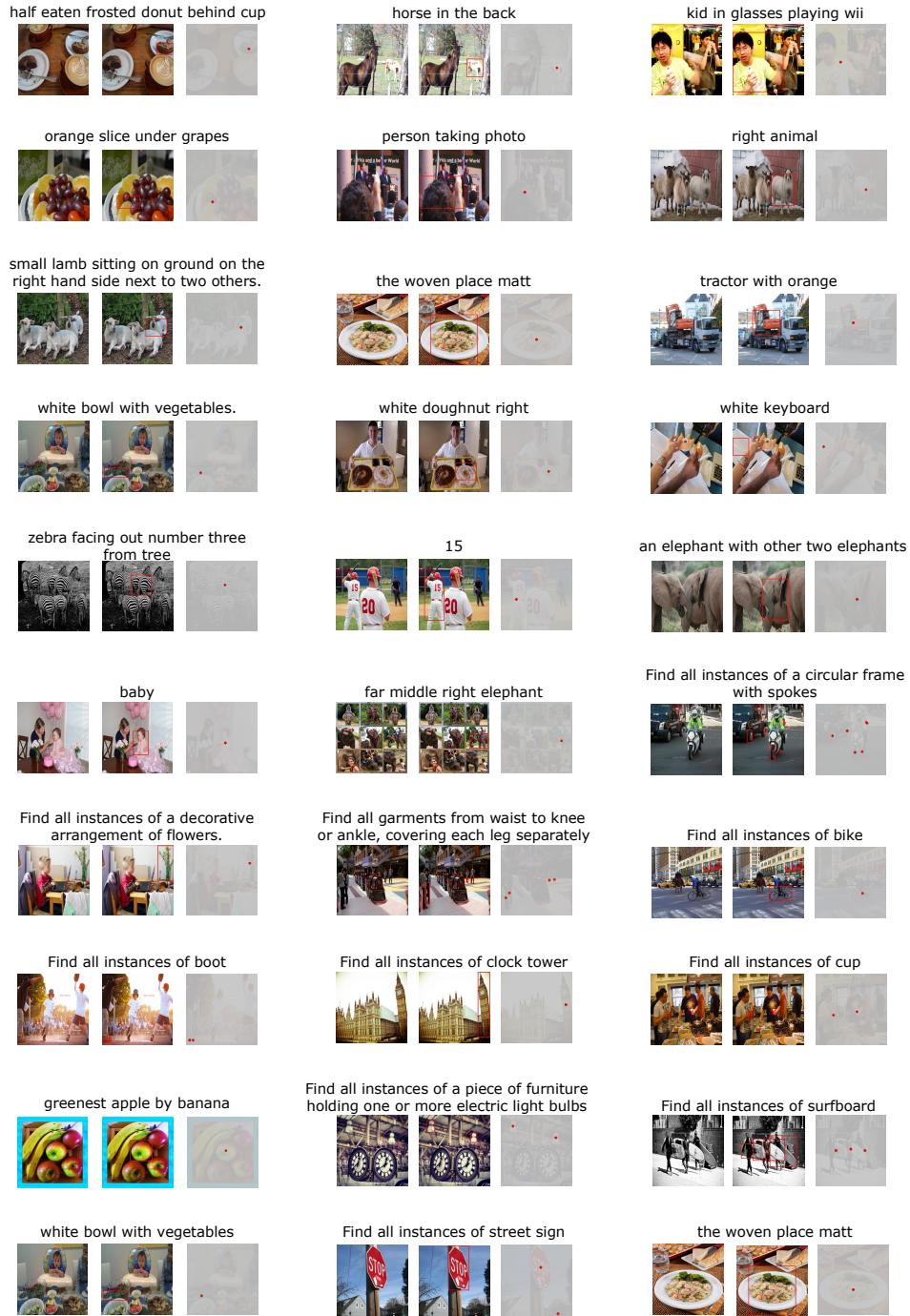


Figure 9: **Visualization of Pre-training Data.** We list some samples of pre-training data. For every sample, the left shows the original image; the middle shows the bounding boxes of the objects of interest; the right shows the ground-truth heatmap used for training.

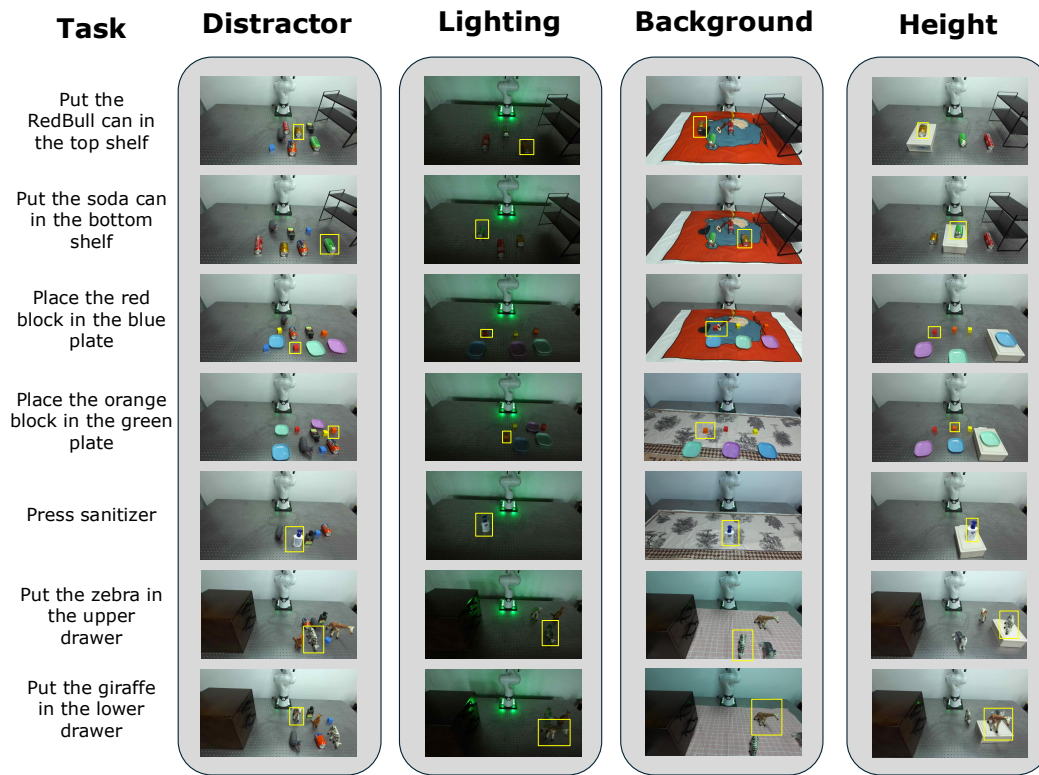


Figure 10: Visualization of the Distractor, Lighting, Background, and Height settings.

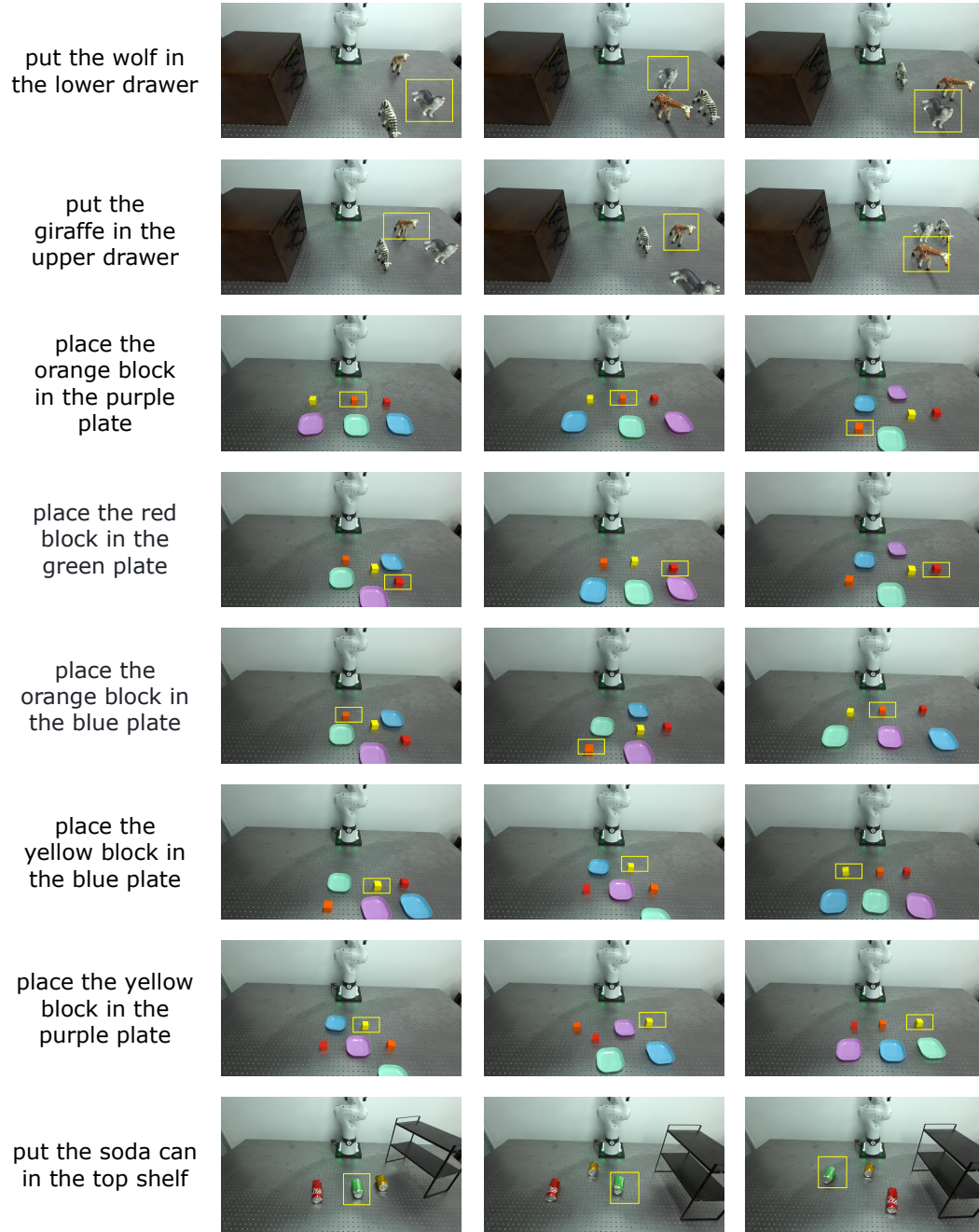
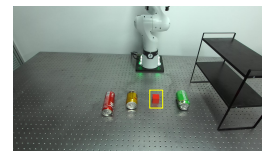
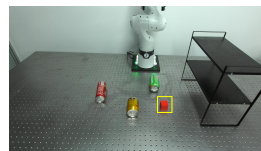
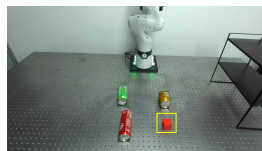
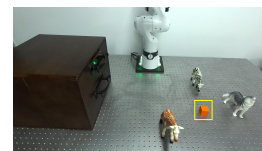
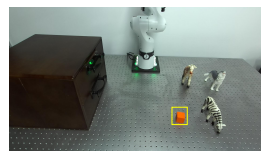
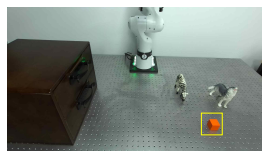


Figure 11: **Visualization of the Combination Setting (I).** During training, the manipulated objects and skills are seen, but their combinations are unseen.

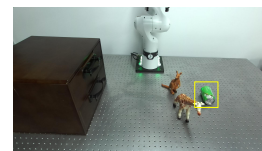
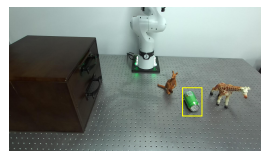
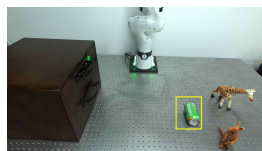
Put the red block
in the bottom
shelf



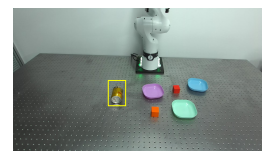
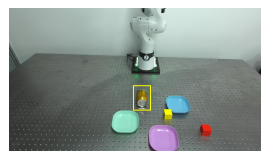
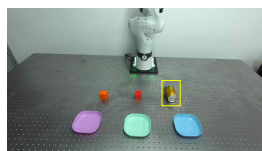
Put the orange
block in the lower
drawer



Put the soda can
in the upper
drawer



Put the Redbull
can in the green
plate



Place the zebra in
the blue plate

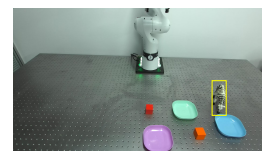
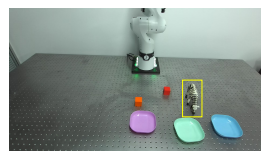
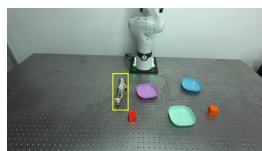


Figure 12: **Visualization of the Combination Setting (II).** During training, the manipulated objects and skills are seen, but their combinations are unseen.

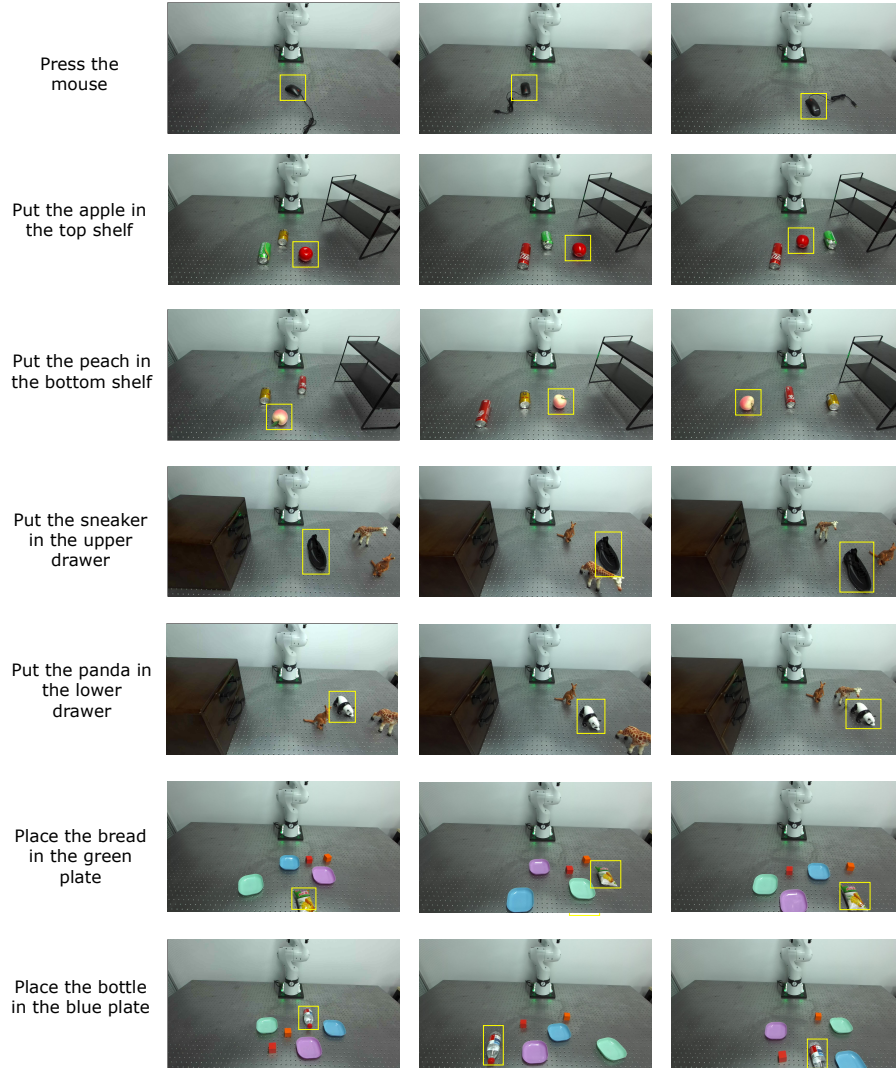


Figure 13: **Visualization of the Category Setting.** In total, we evaluate on 7 objects from novel categories that are unseen during training.

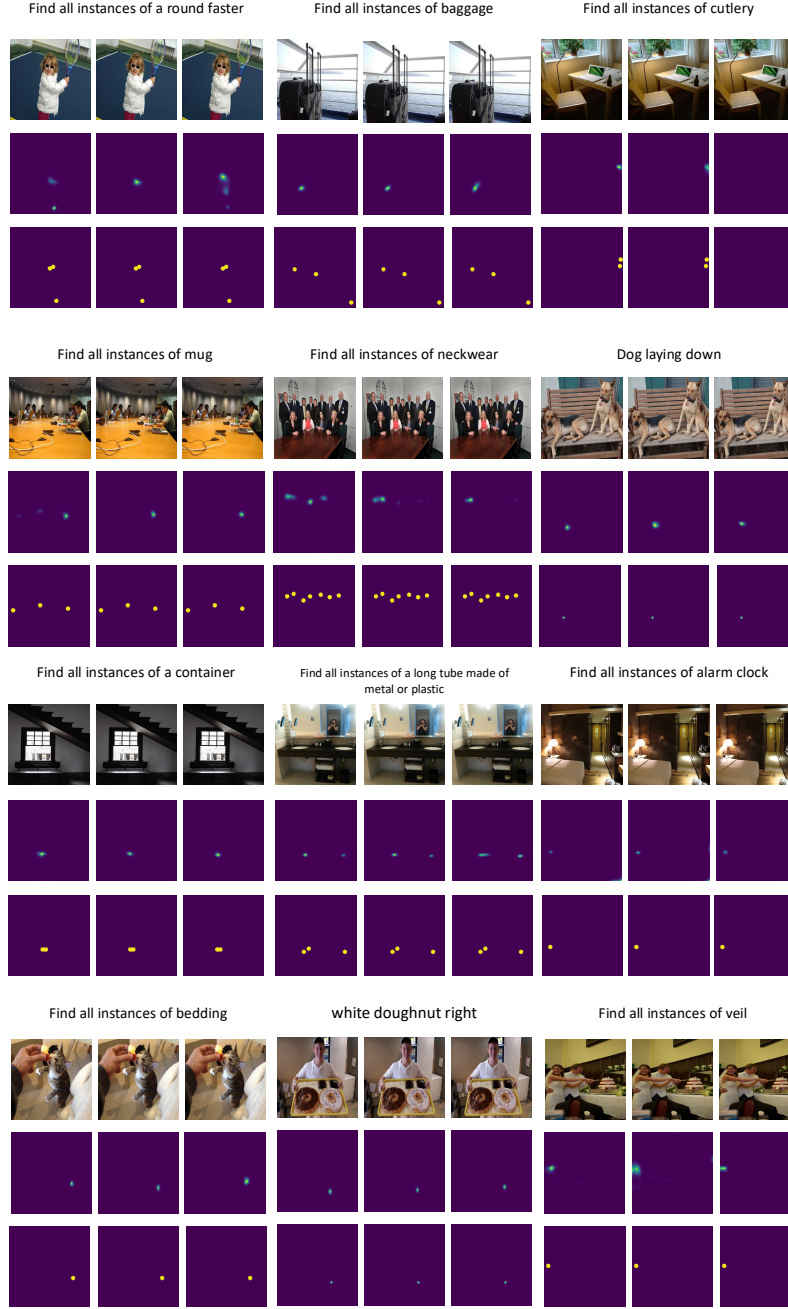


Figure 14: **Visualization of BridgeVLA’s Prediction on Pre-training Dataset after Fine-tuning.** To simulate the multi-view inputs during fine-tuning, we repeat the input image three times and feed them into the fine-tuned model to generate heatmaps. For each sample, the first row shows the input image; the second row shows the heatmap prediction; the third row shows the ground truth.