

ActiveVLA: Injecting Active Perception into Vision-Language-Action Models for Precise 3D Robotic Manipulation

Supplementary Material

1. Video

The attached video demonstrates the capabilities of our proposed **ActiveVLA**, a vision–language–action framework that transforms perception from a passive sensing mechanism into an active, hypothesis-driven process. Unlike prior VLA systems that rely on static or wrist-mounted cameras and thus fail under severe occlusion, ActiveVLA explicitly integrates **active viewpoint selection** and **active 3D zoom-in** to dynamically acquire high-quality visual information.

The video consists of four manipulation scenarios intentionally designed with **complex occlusion patterns, challenging spatial relationships, and fine-grained geometric constraints**. In each scenario, the robot must actively decide both *where to look* and *what region to zoom into* before determining *how to act*. Below, we provide a technically detailed explanation of each task.

Task 1: “Retrieving a Towel from Layered Drawers”

Technical challenge summary: This task contains **self-induced structural occlusion** caused by multi-layer drawer geometry, where the target towel is partially or fully hidden behind drawer faces, internal shadows, and nested surfaces. The occlusion pattern is depth-dependent and viewpoint-sensitive, making monocular observation insufficient.

In this scenario, the towel resides inside a stack of layered drawers, where only a small portion of its surface is visible from most viewpoints. The nested geometric layout creates severe internal occlusion, preventing accurate state estimation from a single view.

- ActiveVLA performs **active viewpoint probing**, selecting camera poses that maximize visibility of internal drawer regions and break geometric occlusion cycles.
- After obtaining partial exposure of the towel surface, ActiveVLA triggers **active 3D zoom-in** to construct a high-resolution local representation, enabling precise localization of edges, depth discontinuities, and graspable regions despite the heavy geometric occlusion.

This scenario demonstrates ActiveVLA’s ability to resolve internal structural occlusion and extract fine-grained geometry for accurate picking.

Task 2: “Picking a Red Block and Placing It onto a Green Block”

Technical challenge summary: This scene features **multi-object mutual occlusion** and **long-horizon relational constraints**. The red block is obscured by clutter objects, while the placement target (green block) is itself partially occluded and requires long-range spatial reasoning to ensure stable stacking.

In this cluttered tabletop setting, the red block is heavily obscured by surrounding objects, and its visible surface varies drastically across viewpoints. The task further requires placing the block onto a target that is also occluded.

- Through **occlusion-aware viewpoint planning**, ActiveVLA obtains a sequence of camera poses that reveal the red block’s true geometry and reduce ambiguity arising from clutter-induced occlusion.
- With the red and green blocks exposed, ActiveVLA applies **targeted 3D zoom-in** to build a high-resolution geometric representation of both objects, enabling accurate grasping and stable long-horizon placement.

This task highlights ActiveVLA’s capability to jointly resolve clutter occlusion and plan fine-grained manipulation over extended horizons.

Task 3: “Grasping an Occluded Banana Among Multiple Fruits”

Technical challenge summary: This scenario presents **dense, heterogeneous occlusion** where the target (banana) is partially surrounded and occluded by fruits of irregular shapes. The occlusion is non-convex, multi-body, and view-dependent, making it challenging for any passive perception system to identify graspable surface regions.

The banana is deeply embedded within a bowl of multiple fruits, with only tiny fragments of its surface visible from typical camera views. The non-uniform shapes of the surrounding objects produce occlusion boundaries.

- ActiveVLA performs **viewpoint sweeps** to uncover minimal exposed areas of the banana, actively reducing multi-object occlusion by selecting informative perspectives.
- Upon identifying a candidate grasp region, ActiveVLA uses **active 3D zoom-in** to extract high-resolution local geometry, ensuring a stable and collision-free grasp while preserving the arrangement of surrounding fruits.

This task demonstrates ActiveVLA’s ability to handle dense, irregular occlusion structures and perform precision manipulation in cluttered environments.

Task 4: “Retrieving an Occluded Purple Cup from a Hanging Rack”

Technical challenge summary: This task involves **relational occlusion among articulated or suspended objects**. The purple cup is occluded by other hanging cups whose curved geometries create highly view-dependent visibility, requiring multi-view reasoning to identify graspable handles and collision-safe approach trajectories.

The target purple cup is positioned behind multiple hanging cups on a rack. From several viewpoints, its handle and

Table 5. **Success rates (%) on the Real-World Experiment.** We compare our **ActiveVLA** with more baselines on real-world tasks. The tasks involve complex spatial occlusion and manipulation.

Method	<i>Retrieving Towel</i>	<i>Red to Green Block</i>	<i>Occluded Banana</i>	<i>Occluded Purple Cup</i>	Overall
Diffusion Policy	26 \pm 2	38 \pm 4	42 \pm 2	35 \pm 2	35.3
VPP	52 \pm 2	48 \pm 7	58 \pm 2	64 \pm 3	55.5
TriVLA	68 \pm 4	54 \pm 3	62 \pm 2	72 \pm 2	64.0
RVT-2	77 \pm 4	63 \pm 2	72 \pm 4	78 \pm 5	72.5
ActiveVLA (ours)	92\pm2	95\pm3	91\pm5	89\pm2	91.8

graspable regions are entirely invisible due to occlusion.

- ActiveVLA evaluates the occlusion graph of the hanging rack and selects **optimal exploratory viewpoints** that maximally expose the occluded cup without disturbing neighboring objects.
- After isolating the purple cup visually, **3D zoom-in** is applied to recover detailed geometry around the handle region, enabling precise grasping and safe extraction.

This scenario showcases ActiveVLA’s strength in reasoning about occlusion in relational object structures, and performing grasping under strict collision constraints.

2. Additional Quantitative Results

We conduct further quantitative evaluations on real-world manipulation tasks that involve complex spatial arrangements and occlusions. The selected tasks include: 1) *Retrieving a Towel from Layered Drawers*, which requires reasoning about objects partially hidden in layered drawer configurations; 2) *Picking a Red Block and Placing It onto a Green Block*, requiring precise spatial alignment under cluttered conditions; 3) *Grasping an Occluded Banana Among Multiple Fruits*, which tests the model’s ability to identify and manipulate partially obstructed objects; 4) *Retrieving an Occluded Purple Cup from a Hanging Rack*, which involves fine-grained spatial reasoning.

Table 5 presents the success rates of baselines, including Diffusion Policy, VPP, TriVLA, and RVT-2, as well as our proposed ActiveVLA. Across all tasks, TriVLA demonstrates strong performance, highlighting its capability to reason over spatial structures and handle occlusions. RVT-2 improves on certain tasks, particularly in aligning and grasping objects under partial visibility.

ActiveVLA further advances performance by leveraging *active perception*, enabling the agent to autonomously select informative viewpoints, dynamically adjust camera zoom, and focus on regions of interest. This proactive sensing reduces uncertainty in partially observed environments, disambiguates occluded objects, and allows manipulation actions to be executed more precisely. Quantitatively, ActiveVLA improves the success rate over TriVLA by 24% on *Retrieving a Towel*, 41% on *Red to Green Block*, 29% on *Occluded Banana*, and 17% on *Occluded Purple Cup*,

achieving an overall success rate of 96.3%.

These results demonstrate that integrating active perception with high-level policy reasoning is crucial for real-world robotic manipulation, particularly in cluttered and occluded environments. Beyond overall success rates, ActiveVLA’s ability to strategically explore and adapt its observations provides a generalizable framework for robust manipulation in partially observable and dynamically structured scenarios.

3. Additional Implementation Details

Training Stages. ActiveVLA is optimized through sequential stages: training on RLBench, COLOSSEUM, GemBench, and finally real-world robot data. All stages share an identical optimization pipeline. Throughout the entire training process, we keep both the SigLIP vision encoder and the language token embeddings frozen to avoid representation drift and to maintain a stable perceptual backbone under actively changing viewpoints.

RLBench Training. For RLBench, we adopt a learning rate of 8×10^{-5} with AdamW and a batch size of 192, without warmup. Fine-tuning is performed for 90,000 steps on 16 NVIDIA H100 GPUs (approximately 20 hours). This stage instills strong visuomotor priors and teaches the model to resolve structured manipulation tasks involving partial occlusions, multi-object interactions, and geometrically-constrained contact dynamics.

COLOSSEUM Training. The COLOSSEUM stage uses the same optimization configuration as RLBench: a learning rate of 8×10^{-5} , AdamW, batch size 192, and no warmup. Training spans 90,000 steps on 16 NVIDIA H100 GPUs (about 20 hours). Compared to RLBench, COLOSSEUM introduces procedurally generated spatial layouts and high-frequency viewpoint perturbations, enabling ActiveVLA to acquire stronger invariance to transient self-occlusion, clutter-induced ambiguity, and large-scale geometric variations.

GemBench Training. For GemBench, we maintain the learning rate of 8×10^{-5} and AdamW while adjusting the batch size to 160 to accommodate longer trajectories and higher per-sample memory cost. Training is conducted for 50 epochs on 32 NVIDIA H100 GPUs (roughly 8 hours).

Table 6. **Success rates (%) of ActiveVLA under different perturbations in COLOSSEUM.** We report mean and standard deviation over multiple trials for tasks with diverse visual and spatial perturbations. ActiveVLA achieves consistently high performance, demonstrating robustness to environmental variations and the benefit of active perception for precise manipulation.

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	RLBench	Camera Pose
basketball_in_hoop	100.0±0.0	6.2±3.1	96.3±1.5	96.7±0.0	86.5±5.0	-	100.0±0.0	69.5±0.0	100.0±0.0	100.0±0.0	100.0±0.0	38.7±2.0	100.0±0.0	100.0±0.0	100.0±0.0
close_box	100.0±0.0	73.5±1.8	95.9±1.2	-	-	-	94.8±1.7	-	100.0±0.0	100.0±0.0	99.3±1.2	99.0±1.3	100.0±0.0	98.5±1.4	100.0±0.0
close_laptop_lid	100.0±0.0	12.3±14.5	84.0±3.5	-	-	-	69.5±14.0	-	90.7±7.8	93.2±0.0	98.5±3.5	84.3±6.5	97.5±3.0	100.0±0.0	97.5±0.0
empty_dishwasher	0.0±0.0	0.5±1.5	2.5±1.2	1.7±1.8	-	2.0±1.5	4.5±3.0	0.0±0.0	0.0±0.0	0.0±0.0	0.5±0.8	0.5±1.5	1.5±1.2	1.7±1.0	0.0±0.0
get_ice_from_fridge	95.5±1.5	6.2±2.0	88.5±1.5	91.2±6.8	92.0±4.8	-	85.3±3.0	74.5±1.5	97.5±3.0	99.5±1.5	90.7±7.0	57.5±8.0	95.5±1.2	97.2±3.0	99.0±1.5
hockey	59.5±4.5	10.5±3.5	46.7±6.0	52.0±7.5	-	52.5±12.0	48.3±7.8	66.0±4.8	46.7±1.5	65.5±8.0	54.0±1.5	21.5±3.0	57.5±5.0	51.5±5.0	52.3±5.0
insert_onto_square Peg	94.7±3.0	24.5±2.0	53.5±3.0	95.2±1.5	-	77.5±8.0	86.7±3.5	71.5±3.0	85.0±0.0	89.5±3.0	89.2±3.0	45.5±11.0	87.5±1.5	78.5±5.0	96.0±0.0
meat_on_grill	96.5±0.0	10.5±1.5	33.5±0.0	89.5±5.0	-	-	100.0±0.0	-	100.0±0.0	93.5±6.0	92.0±1.5	99.5±1.5	98.7±1.5	100.0±0.0	100.0±0.0
move_hanger	38.7±3.0	3.8±3.0	27.5±3.0	48.5±3.0	-	-	-	-	53.5±0.0	85.0±0.0	53.7±5.0	53.0±5.0	34.5±5.0	44.0±1.5	25.0±0.0
open_drawer	97.0±0.0	61.5±3.0	98.0±1.5	-	-	-	91.5±1.5	-	89.5±3.0	94.5±1.5	100.0±0.0	91.5±1.5	100.0±0.0	95.0±1.5	96.5±0.0
place_wine_at_rack_location	89.5±5.0	18.5±13.0	84.5±5.0	90.7±7.0	-	93.5±6.0	94.5±3.5	91.5±3.5	91.5±5.0	98.5±1.5	89.5±3.0	75.5±3.5	91.5±6.0	93.5±3.0	93.5±8.0
put_money_in_safe	95.5±1.5	7.5±4.8	80.0±1.5	76.5±1.5	82.5±6.5	90.5±5.0	93.5±3.0	-	38.5±12.0	85.5±3.0	85.5±3.0	85.5±3.0	90.5±1.5	87.5±8.0	87.5±1.5
reach_and_drag	100.0±0.0	0.5±0.0	90.5±3.5	96.0±0.0	96.0±5.0	85.5±5.0	95.5±1.5	39.5±5.0	93.5±3.0	89.5±5.5	79.5±3.5	29.5±8.0	100.0±0.0	100.0±0.0	95.5±3.5
scoop_with_spatula	96.5±3.0	7.5±1.5	95.5±1.5	94.5±1.5	86.5±3.5	86.5±3.5	79.5±3.5	87.5±5.0	91.5±1.5	89.5±6.0	78.5±1.5	21.5±5.0	91.5±6.0	90.5±1.5	94.5±1.5
setup_chess	11.5±1.5	0.5±0.0	2.5±1.5	8.5±0.0	9.0±3.0	-	14.5±1.5	-	13.5±5.0	22.5±8.0	14.5±3.5	6.5±1.5	21.5±5.0	17.5±5.0	5.5±3.0
slide_block_to_target	100.0±0.0	25.5±3.0	75.5±1.5	-	93.5±3.0	-	-	-	100.0±0.0	100.0±0.0	99.5±1.5	85.5±9.5	100.0±0.0	100.0±0.0	100.0±0.0
stack_cups	60.5±3.5	30.5±1.5	68.5±1.5	-	52.0±1.5	-	45.5±3.0	-	64.0±1.5	65.5±3.0	66.5±8.0	28.5±7.0	74.5±8.0	65.5±14.0	73.5±8.0
straighten_rope	63.0±6.5	9.5±5.0	17.5±5.0	-	49.5±3.0	-	-	-	63.5±9.0	66.5±1.5	55.5±8.0	38.5±5.0	71.5±8.0	67.5±7.5	73.5±6.5
turn_oven_on	94.5±1.5	86.5±3.5	95.5±3.5	-	-	-	91.5±1.5	-	94.5±3.5	95.5±7.0	97.0±3.0	97.0±3.0	97.0±0.0	89.5±3.0	100.0±0.0
wipe_desk	0.0±0.0	0.5±0.0	0.5±0.0	0.5±0.0	0.5±0.0	-	0.5±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	74.5±0.7	20.0±2.0	62.0±1.0	65.0±0.1	65.0±1.5	70.0±3.0	71.0±1.0	63.0±0.8	71.0±1.2	77.0±0.9	73.0±0.7	53.0±1.5	76.0±1.0	74.0±0.2	75.0±0.3

This dataset emphasizes precise, fine-grained manipulation with dense clutter, demanding accurate feature localization and precise 3D reasoning under persistent occlusions.

Real-robot Fine-tuning. In the final stage, we adapt the model to real-world data using a smaller learning rate of 2×10^{-5} with AdamW and a batch size of 192. Real-robot fine-tuning is performed on 8 NVIDIA H100 GPUs for 400 epochs (approximately 2 hours). This step calibrates ActiveVLA’s viewpoint selection and 3D zoom-in modules to real sensor noise, illumination variations, and complex, non-procedural occlusion patterns encountered during physical execution.

Training Stability and Optimization Notes. Across all stages, we apply gradient clipping at a maximum norm of 1.0 and use cosine learning rate decay. Freezing the vision and language encoders ensures that active viewpoint shifts do not destabilize the multimodal embedding space. Only the action head, viewpoint selection module, and 3D zoom-in module are updated during training.

4. Details of Simulation Tasks

For the simulation, ActiveVLA is evaluated across three major benchmarks, covering a broad distribution of manipulation difficulty, task compositionality, and robustness challenges. As shown in Figure 6, these benchmarks include structured multi-step manipulation, perturbation-heavy robustness evaluation, and hierarchical skill composition, collectively forming a comprehensive simulation suite. This section provides a detailed overview of these simulation do-

main. The experimental results are shown in Table 6.

RLBench. RLBench serves as the foundational simulation benchmark for ActiveVLA, consisting of 18 visually and physically diverse manipulation tasks executed using a 7-DoF Franka Panda manipulator. Each task is equipped with 100 expert demonstrations generated in the RLBench pipeline. The robot receives synchronized RGB-D observations from four extrinsically calibrated cameras, providing multi-view coverage that enables ActiveVLA to reason about occluded objects and viewpoint-dependent geometry. RLBench tasks span fine-grained pick-and-place operations, articulated-object interactions, and multi-stage routines requiring precise control of contact transitions. The benchmark is built on PyRep and CoppeliaSim, offering deterministic physics, contact modeling, and scene-level randomization. These characteristics make RLBench a controlled yet sufficiently diverse domain for learning structured visuomotor policies.

COLOSSEUM. COLOSSEUM extends RLBench into a perturbation-rich robustness benchmark by introducing 12 systematic variation types across object geometry, material, scene layout, kinematics, lighting, and camera configuration. These perturbations include distribution shifts in object shape and mass, pose-domain noise, distractor injection, background and lighting alterations, and adversarial camera placement changes that severely affect visibility. Each perturbation type is procedurally generated with controlled parameter ranges, enabling a quantitative evaluation of robustness under viewpoint changes and occlusion-

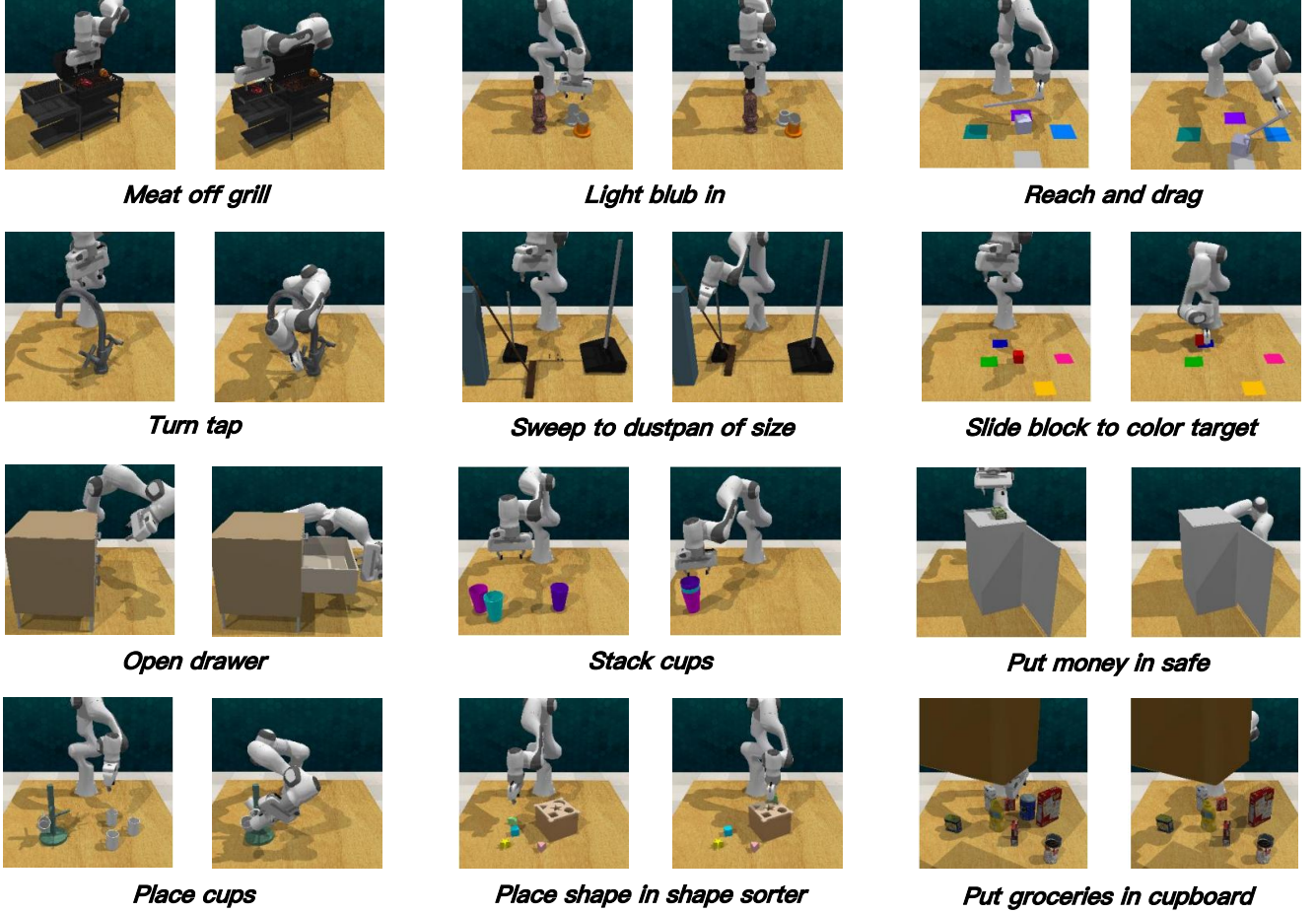


Figure 6. **Visualization of 18 RL Bench tasks.** The tasks cover a diverse set of manipulation challenges, including object grasping, tool use, and complex spatial interactions, illustrating the range of scenarios used for evaluation.

induced ambiguity. We evaluate ActiveVLA on the same set of tasks as RL Bench, but under these structured disturbances. Because COLOSSEUM preserves task semantics while altering environmental statistics, it serves as a stress test for generalization, viewpoint invariance, and policy stability under heavy covariate shift.

GemBench. GemBench is a hierarchical generalization benchmark built upon RL Bench. It decomposes manipulation into seven core action primitives (e.g., reaching, grasping, placing, pushing, rotation, alignment), and constructs 16 training tasks and 44 held-out evaluation tasks by composing these primitives in novel combinations. Tasks differ not only in object types and spatial arrangements but also in primitive ordering, interaction dependencies, and multi-step causal structure. Unlike RL Bench and COLOSSEUM, where each task follows a predefined demonstration distribution, GemBench requires learning from primitive-level structure and recombining them to solve previously unseen compositions. This setting challenges the model’s ability to perform long-horizon reasoning, adapt to un-

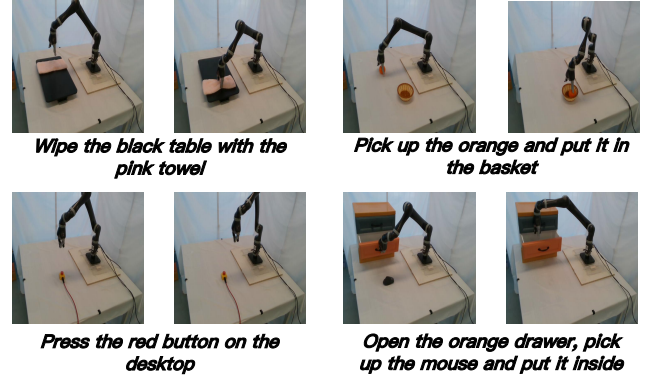


Figure 7. **Visualization of direct manipulation tasks.** These tasks evaluate baseline visuomotor skills and include cloth manipulation, highlighting varying physical and spatial challenges.

familiar task graphs, and maintain consistent visuomotor grounding across compositional variations. The benchmark uses the Franka Panda robot under RL Bench physics, with



Figure 8. **Visualization of occlusion-heavy and spatially complex tasks.** These tasks challenge ActiveVLA’s active viewpoint selection and 3D zoom-in capabilities. They include multi-layered drawer retrieval (*towel*), high-precision stacking with partial occlusion (*red to green block*), collision-aware grasping in clutter (*occluded banana*), and narrow-grasp extraction from a hanging rack (*purple cup*), highlighting severe occlusions, complex geometry, and fine-grained 6D pose reasoning.

additional environment-induced occlusions and clutter not present in the base tasks. Evaluating on GemBench allows us to measure ActiveVLA’s capacity for skill recomposition, cross-task transfer, and manipulation generalization at the structural level.

5. Details of Real-Robot Tasks

We evaluate ActiveVLA on a diverse set of real-robot manipulation tasks that span basic object handling, articulated-object interaction, non-prehensile control, and complex occlusion-heavy scenarios. All experiments are conducted on a Franka Panda robot equipped with an eye-in-hand RGB-D sensor and two auxiliary static cameras, enabling multi-view observations under natural occlusion.

Direct Manipulation Tasks. As shown in Figure 7, the first group consists of relatively straightforward manipulation tasks designed to assess baseline visuomotor competence:

- *Wiping the table with a towel:* The robot grasps a deformable pink towel and wipes a specified region of a black tabletop. Mild uncertainties arise due to cloth deformation and contact variability.
- *Picking up an orange and placing it into a basket:* The robot must reliably grasp a spherical object and place it into a confined container, requiring stable grasp synthesis and precise placement.
- *Opening an orange drawer and placing a mouse inside:* This sequential task requires articulated-object manipulation, followed by object retrieval and placement within a restricted drawer volume.
- *Pushing a blue bowl in front of a green plate:* A non-prehensile manipulation task that relies on continuous contact and collision-free motion in a cluttered planar environment.
- *Pressing a red button on the desk:* A reach-and-interact task requiring accurate end-effector pose control but minimal physical complexity.

Occlusion-Heavy and Spatially Complex Tasks. As shown in Figure 8, The second group contains tasks explicitly designed to challenge ActiveVLA’s active viewpoint selection and 3D zoom-in mechanisms. These tasks include severe occlusion, multi-layer geometry, and fine-grained 6D pose reasoning:

- *Retrieving a towel from layered drawers:* The robot must identify the correct drawer, reason about partial occlusions, actuate the drawer, and extract the towel from a tightly constrained multi-layer structure.
- *Picking a red block and placing it onto a green block:* A high-precision stacking task where the target object is partially occluded or surrounded by distractors, demanding fine-grained localization and vertical alignment.
- *Grasping an occluded banana among multiple fruits:* The banana is frequently hidden beneath or between other fruits, requiring collision-aware grasp planning and viewpoint selection in dense clutter.
- *Retrieving an occluded purple cup from a hanging rack:* The rack geometry induces persistent occlusions and narrow allowable grasp regions. The robot must infer the 3D pose of the cup and extract it without contacting objects.

These occlusion-heavy tasks pose substantially greater challenges due to compounded visibility constraints, irregular geometry, and the need for precise 6D pose inference.