RobAVA: A Large-scale Dataset and Baseline Towards Video based Robotic Arm Action Understanding (Supplementary Material)

A. RobAVA

A.1. Details of Dataset Collection

RobAVA-S. The RobAVA-S dataset aims to provide a diverse and large-scale data resource for robotic manipulation by combining virtual environment simulations and real human demonstrations. It captures both normal and anomalous trajectories, addressing the lack of diversity and anomalous data in existing datasets. During data collection, human operators remotely controlled robots in simulation environments to perform a small number of demonstration tasks across various scenarios. These included basic tasks (e.g., picking and assembling), contact-based tasks (e.g., precise insertion and cleanup), long-horizon tasks (e.g., sequential pick-and-place), and high-precision industrial tasks (e.g., gear assembly and frame construction). Subsequently, these human demonstrations were processed and expanded using the MimicGen system [1]. MimicGen decomposes task trajectories into object-centric subtask segments, transforms and reconstructs them for new scenes and initial state distributions, and generates diverse trajectory data. The system employs real-time task success evaluation to filter the generated trajectories, retaining only those that successfully complete the task to ensure the quality and usability of the data. However, during the data collection process, we observed that while excluding anomalous data ensures consistency and reliability, it also neglects the potential benefits of including diverse anomalous trajectories, such as for anomaly detection and fault recovery tasks. To address this, we intentionally retained anomalous data during the generation process. MimicGen supports multiple simulation environments, including robosuite and Factory. The former emphasizes flexibility and scene diversity, while the latter supports sub-millimeter precision tasks, covering a range of requirements from everyday tasks to complex industrial applications. Furthermore, MimicGen demonstrated significant data expansion capabilities. Using fewer than 100 human demonstrations, we generated 17,000 trajectories covering 8 tasks. By incorporating various robotic arms, the dataset's breadth and diversity were further enhanced. By including anomalous data and capturing complete action sequences, RobAVA-S provides a solid foundation for advancing research in robotic manipulation, including anomaly detection, recovery strategies, and the development of more adaptable robotic systems.

RobAVA-R. We collected and selected videos of real-world robotic arm actions from the publicly available robotic manipulatio dataset, Open X-Embodiment [2]. The dataset consists of 100 distinct manipulation tasks, each clearly annotated and containing both successful and unsuccessful task attempts. Objects used for these tasks are divided into two sets: Object Set 1, used for 21 training tasks, and Object Set 2, used for 79 training tasks. For evaluation purposes, 4 tasks were generated among objects within Object Set 2, and 24 holdout tasks utilized objects from both sets. Multiple instances of these objects were used, with slight physical variations to enhance diversity. For example, the ceramic bowl in Object Set 1 appears red in the original image, but data was also collected using ceramic bowls painted green or blue. Similarly, sponges used in tasks could be either blue or white with subtle shape differences, while erasers and peppers varied in size and material. The data collection process was distributed across multiple robots at 1 to 4 physical locations. This setup enabled the development of a policy that could handle variations across different robot hardware, backgrounds, and scene configurations. Each data collection station utilized a unique set of objects with minor physical differences. This variability introduces higher sample complexity, as the policy must learn to generalize across these variations to achieve the desired performance level with each object set. Throughout each episode, the camera view remained fixed to ensure consistency in observation.

Prompt Template of Attribute Generation

Given the category name of a robotic arm action, perform the following steps:
(1) Imagine a scene involving the specified action.
(2) Generate M concise descriptions of the continuous motion steps of the action, with each description limited to 12 words or fewer.

Figure 1. Prompt supplied for the attribute prompt generation.

A.2. More Example

Here, we provide more examples of our RobAVA-S (Figure 2) and RobAVA-R (Figure 3, 4, 5), including both normal ✅ and anomalous ❌ action executions.

B. AGPT-Net

B.1. Attribute Prompts Generation

For each category, we design a prompt template to query LLMs as shown in Figure 1, aiming to obtain video descriptions with local contextual information across different time intervals.

The attribute descriptions for some categories in RobAVA-R are given below.

"drag the pepper across the table": [
"Pepper is grabbed by the robotic arm.",
"Arm moves pepper across the table.",
"Pepper slides smoothly along the surface.",
"Pepper reaches the end of its path."
],
"knock the bottle over": [
"Arm moves toward the bottle on table.",
"Fingers lightly push the bottle's edge.",
"Bottle tilts and starts to fall.",
"Bottle falls completely, liquid spills."
],
"move the arm in a circular motion": [
"Arm begins a circular sweeping motion.",
"Motion follows a consistent circular path.",
"Arm rotates smoothly in the circle.",
"Circle is completed with perfect precision."
],
"pick up the ceramic bowl": [
"Arm moves towards the ceramic bowl.",
"Fingers grasp the bowl firmly.",
"Bowl is lifted from the surface.",
"Arm retracts, holding the bowl."
],
"place apple in metal cup": [
"Apple is picked up by robotic arm.",
"Arm moves apple toward the metal cup.",
"Fingers adjust to lower the apple.",
"Apple is placed securely in the cup."
],
"place towel in tray": [
"Towel is picked up by robotic gripper.",
"Arm moves towel toward the tray.",
"Towel is placed gently inside the tray.",
"Towel rests securely in the tray."
],
"push the ceramic bowl across the table": [
"Arm moves to the ceramic bowl's side.",
"Fingers gently push the bowl forward.",
"Bowl slides smoothly across the table.",
"Bowl reaches the other side of table."
],
"push the eraser across the table": [
"Eraser is touched gently by the arm.",
"Arm pushes eraser across the surface.",
"Eraser glides smoothly across the table.",
"Eraser comes to rest at table's edge."
],
"stack bowls into tray": [
"Bowls are picked up by robotic gripper.",
"Arm moves toward the tray carefully.",
"Bowls are stacked one by one inside.",
"Bowls rest securely stacked in the tray."
],
"stack bowls on top of each other": [
"Bowls are picked up one by one.",
"Each bowl is placed on top of another.",
"Bowls stack securely without shifting.",
"Stack of bowls is stable and balanced."
],
"stack cups into tray": [
"Cups are grabbed by robotic fingers.",
"Arm moves toward the tray carefully.",
"Cups are stacked neatly in the tray.",
"Stacked cups rest securely inside tray."
],
"stack cups on top of each other": [
"Cups are picked up one by one.",
"Each cup is stacked carefully on another.",
"Cups are stacked without tilting.",
"Cup stack is stable and balanced."
],
"stand the bottle upright": [
"Bottle is grabbed gently by the arm.",
"Arm moves bottle into upright position.",
"Bottle is placed vertically on table.",
"Bottle stands upright, secure and stable."
],
"wipe purple bowl with brush": [
"Brush is picked up from the surface.",
"Arm moves brush toward purple bowl.",
"Brush scrubs the surface of the bowl.",
"Brush lifts, bowl surface is clean."
],
"wipe purple bowl with eraser": [
"Eraser is lifted by robotic gripper.",
"Arm moves eraser toward purple bowl.",
"Eraser wipes the inside of the bowl.",
"Bowl surface is cleared of debris."
],
"wipe purple bowl with sponge": [

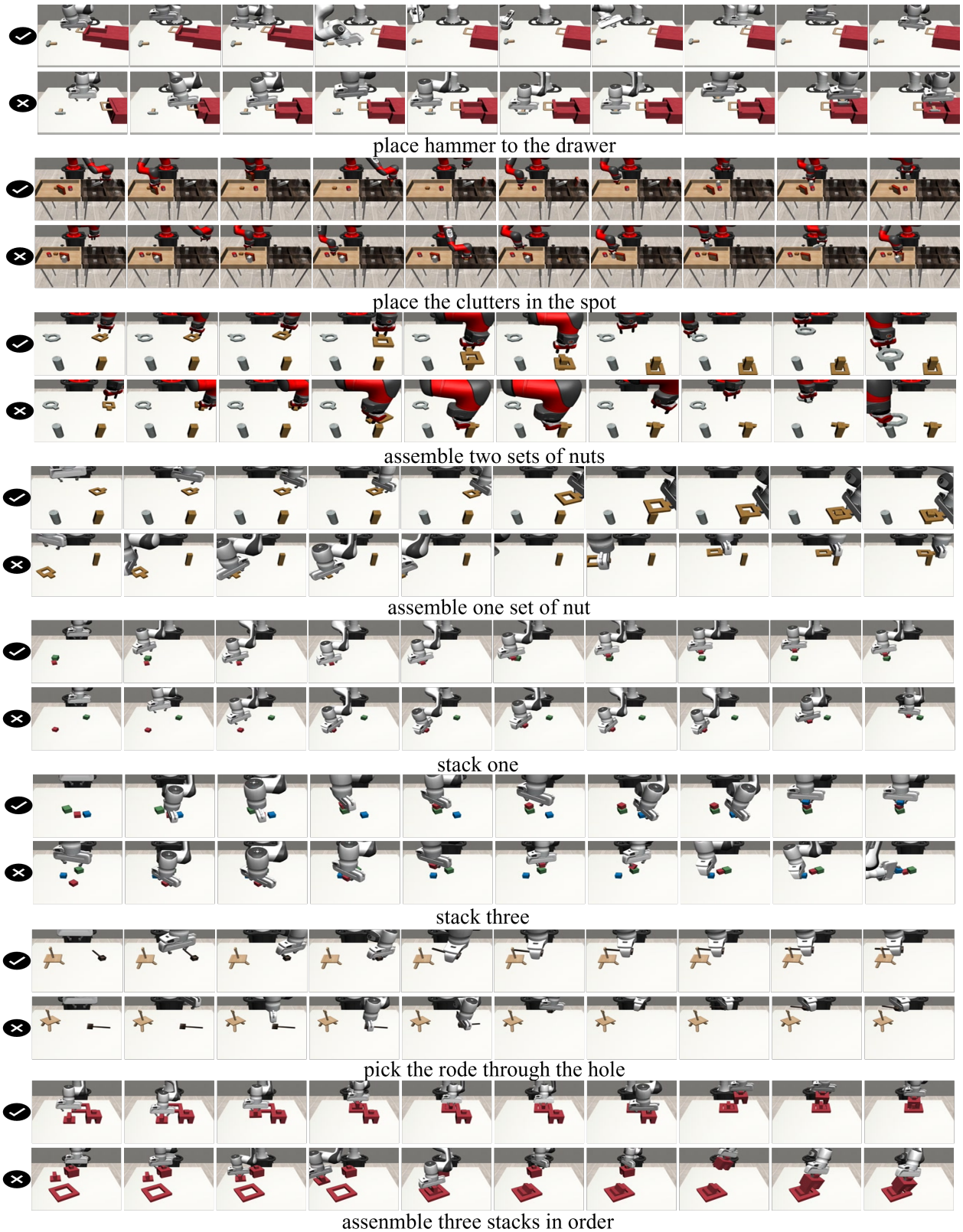

Figure 2. Examples of action categories in our RobAVA-S.

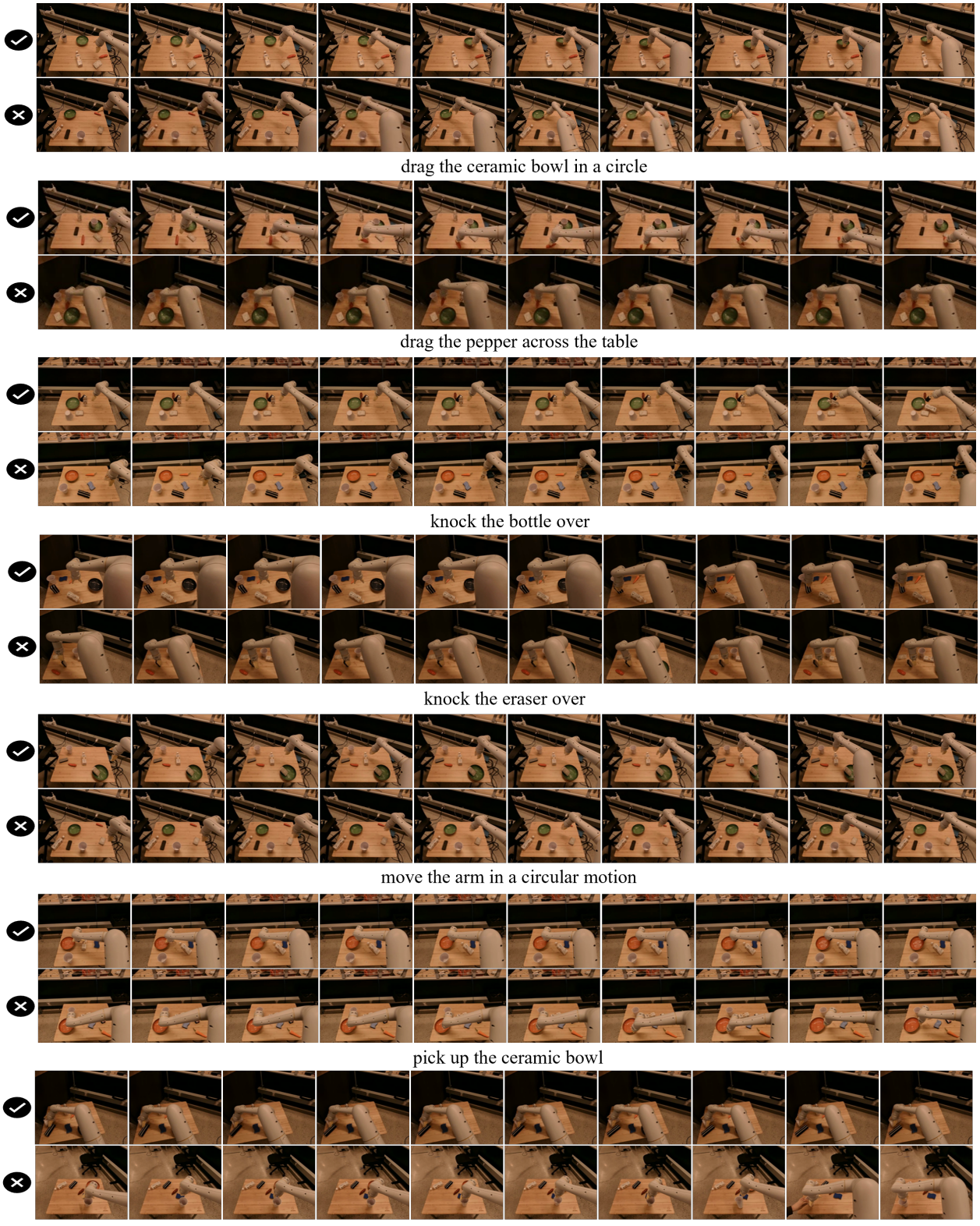

Figure 3. Examples of action categories in our RobAVA-R (part 1).

Figure 4. Examples of action categories in our RobAVA-R (part 2).

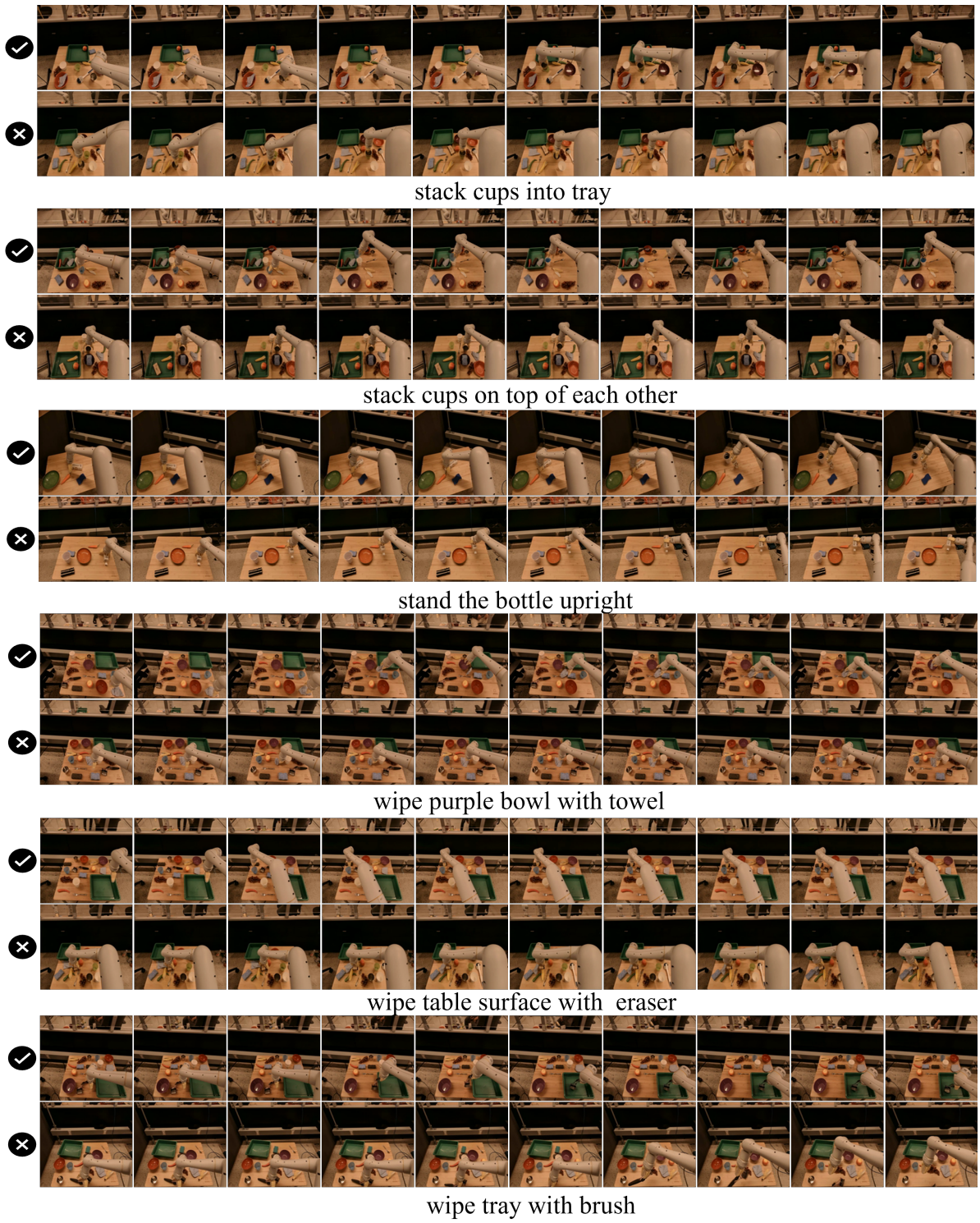

Figure 5. Examples of action categories in our RobAVA-R (part 3).

"Sponge is picked up by robotic arm.",
"Arm moves sponge toward purple bowl.",
"Sponge wipes the inside of the bowl.",
"Bowl surface is cleaned with sponge."
],

References

[1] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj S. Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *Conference on Robot Learning, CoRL, November*, pages 1820–1864, 2023. 1

[2] Abby O'Neill, Abdul Rehman, and et. al. Open x-embodiment: Robotic learning datasets and RT-X models : Open x-embodiment collaboration. In *IEEE, ICRA, May*, pages 6892–6903, 2024. 1