# IN-CONTEXT ROBOT LEARNING MADE SIMPLE: A DEMOCRATIZED RECIPE FOR MANIPULATION TASKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study robotic in-context learning (ICL), an emerging paradigm that enables robots to infer and execute tasks from visual demonstrations. Despite its growing promise, the problem itself remains under-defined: a visual demonstration simultaneously conveys action trajectories, object semantics, manipulation affordances, spatial relations, and task goals, making it unclear what information the robot is actually expected to follow. In this work, we first provide a clear problem definition of robot ICL that explicitly defines its learning target and resolves this fundamental prompt ambiguity. Building on this definition, we develop a minimalist and reproducible ICL framework (**SimpleICL**) with a visual prompt encoder and a low-cost data collection protocol. Without massive pre-training or specialized data infrastructure, our framework achieves strong performance in both simulation and real-world environments. Extensive experiments further reveal several key properties of robot ICL, including action, semantic, composition, and affordance discrimination. We will fully open-source our data and training pipeline to facilitate systematic and reproducible research on robot ICL.

## 1 INTRODUCTION

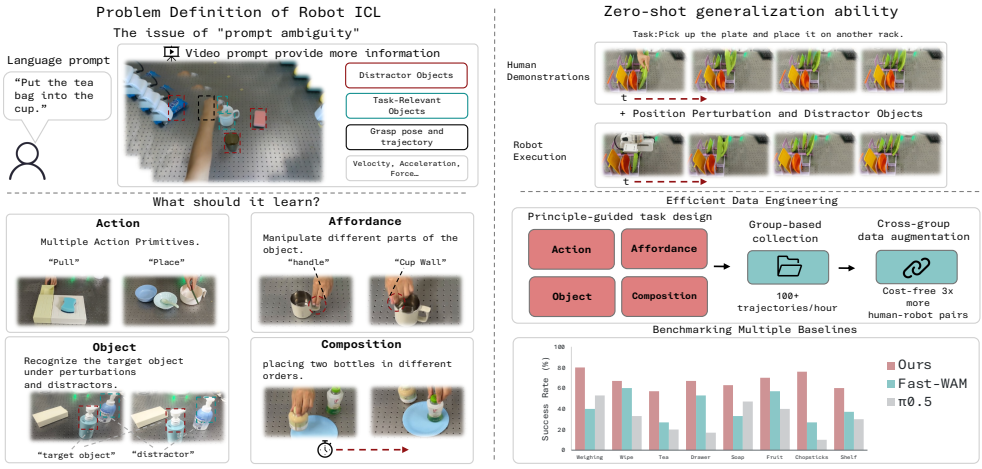



Figure 1: **Overview of SimpleICL.** We identify the prompt ambiguity problem in robot ICL and clarify the intrinsic semantics that the model should learn. We further improve task understanding and generalization through efficient data engineering. Our method achieves the best performance across eight evaluation tasks.

In recent years, the learning paradigm of embodied intelligence has undergone multiple shifts: from reinforcement learning (RL) in simulation, to imitation learning reliant on large-scale teleoperation data, to real-world RL post-training, and further to the utilization of egocentric data. This surge of interest naturally begs a fundamental question: *what is the most intrinsic way of robot learning?*

Although it is hard to answer such a big question, we can first look back at how humans themselves acquire manipulation skills. We typically learn a skill by observing others perform it and then au-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tonomously imitating them. Unlike existing imitation learning methods, humans are rarely forced to learn by having someone physically control our arms. Instead, we perceive and understand the correct operational procedures and logic through visual observation. Instead, we perceive and understand the correct operational procedures and logic through visual observation. In essence, this is a form of in-context learning (ICL), where the visual observations we receive act as visual prompts. Such visual prompts carry vastly richer information than language prompts, including spatial locations, object appearance and shape, and fine-grained action trajectories. Consequently, in-context learning incorporating visual prompts is widely considered more likely to exhibit zero-shot generalization capabilities.

In very recent months, several concurrent works have independently begun exploring in-context robot learning, such as HOST (Chen et al., 2026), GEN 1.5 (Team, 2026), and Zero-WAM (Zhou et al., 2026). These works have demonstrated preliminary ICL capabilities, achieving zero-shot generalization across novel tasks. However, as an under-explored learning paradigm, in-context robot learning still harbors numerous open problems waiting to be investigated. First and foremost among these is the ambiguity of ICL prompts—that is, because video conveys an overabundance of information, it often fails to precisely convey the intent of the human demonstration. For instance, should the robot strictly follow the human's motion trajectory, or should it adapt based on the high-level task/object semantics? Should it strictly replicate the specific way an object is manipulated, or focus solely on achieving the final goal? Figure 1 illustrates this inherent ambiguity in ICL problems. Beyond this, several other open questions remain, including but not limited to: (1) What are the minimal necessary factors for ICL to work in robotics? (2) How should data for ICL be designed and collected? (3) Can ICL learn fine-grained manipulation affordances?

To address these questions, this paper presents a simple yet effective framework for robot ICL, termed **SimpleICL**, adhering to the principles of being simple and easily reproducible. Within this framework, to tackle the critical challenge of prompt ambiguity, we first establish a problem definition for robot ICL – what exactly we expect ICL to learn from human demonstrations. This definition logically eliminates intent ambiguity during context learning. Based on the definition, we further introduce a universal visual prompt encoding module alongside a standardized protocol for data design and collection. This prompt encoding module can be seamlessly integrated into mainstream model architectures. Our data collection standards likewise follow the ambiguity-avoiding principle, encompassing a comprehensive definition for actions, objects, and spatial layouts, coupled with a cost-free data augmentation strategy. This collection methodology enables us to achieve promising ICL performance at a low data cost, without relying on massive pre-training or specialized data infrastructure. This holds significant value for democratizing and decentralizing ICL research across the broader community. In summary, our contributions are threefold:

- **Identifying and addressing prompt ambiguity:** We are the first to expose and thoroughly examine the issue of prompt ambiguity in visual-prompt-based ICL, offering a clear problem definition of ICL to resolve it.
- **A simple, minimalist architecture:** We propose a minimalist robot ICL model architecture without bells and whistles and conduct extensive experiments in both simulation and real-world environments. Our results demonstrate the effectiveness of ICL, while uncovering a range of underlying properties and open challenges for the first time.
- **Low-cost and reproducible data recipe:** We contribute a low-cost, easily reproducible data infrastructure for ICL and will fully open-source the entire data and training pipeline, facilitating the further adoption and advancement of this emerging research paradigm within the robotics community.

## 2 Related Work

**Vision-Language-Action and World Action Model Policies** In their seminal works, embodied studies generally utilize Vision-Language-Action (VLAs) models and World Action Models (WAMs) as their foundation models. The VLAs (Kim et al., 2024; Black et al., 2024; Intelligence et al., 2025; Bjorck et al., 2025; Zitkovich et al., 2023; Liu et al., 2024; Bu et al., 2025; Shukor et al., 2025; Team et al., 2025; Team, 2025; Wen et al., 2025) align the text instruction and visual observations into the latent space of a pretrained large vision-language model, which is finetuned to produce the robotic actions. In parallel, the WAMs (Du et al., 2023; Wu et al., 2023; Zhou et al.,

2024; Feng et al., 2025; Bharadhwaj et al., 2024; Won et al., 2025; Cheang et al., 2024; Jang et al., 2025; Zhao et al., 2025; Cen et al., 2025a;b; Zhou et al., 2025; Zheng et al., 2025; Zhang et al., 2025; Zhu et al., 2025; Liang et al., 2025; Kim et al., 2026; Liao et al., 2025; Pai et al., 2025; Li et al., 2026; Bi et al., 2025; Ye et al., 2026; Yuan et al., 2026) are based on video generation models, which consider future video prediction as a prior to infer the actions. However, most existing policies rely on task-specific finetuning to generalize to novel tasks (Kim et al., 2024; Black et al., 2024; Team et al., 2025). To reduce the high costs of robotic data collection and model retraining, recent works investigate methodologies to learn from human demonstration videos.

**Learning from Human Demonstrations** Unlike robot demonstrations, human manipulation data typically consist only of video, rather than multimodal signals such as states and actions. WAM-TTT (Feng et al., 2026) proposes a test-time adaptation that optimizes a lightweight memory to acquire new skills from human video. ReCAP (Park et al., 2026) appends the new task demonstration to the retrieval pool and retrieves it when executing. Very recent works propose in-context learning on robotics. Similar to ICL in large language models (LLMs) (Brown et al., 2020), ICL on robotics regards the human demonstration videos as part of the prompt along with text instructions. The video prompts bring richer information such as spatial constraints, intermediate states, and temporal structure. In addition to directly using the entire video as context (Zhou et al., 2026; Li et al., 2026), another approach, HOST (Chen et al., 2026), is to temporally align the video in real time and use the corresponding current frame as the conditioning input. Nevertheless, in-context learning (ICL) on robotics remains an under-explored learning paradigm. As a novel area, our work intends to address the issues of definition, phenomenon, data design, and model architecture in ICL.

# 3 METHOD

## 3.1 PROBLEM DEFINITION OF ROBOT ICL

As articulated in Section 1, resolving prompt ambiguity necessitates establishing a clear problem definition of In-Context Robot Learning (ICL). Fundamentally, these definitions can be approached from two complementary perspectives: (1) *What do we expect ICL to learn?* and (2) conversely, *What do we expect ICL NOT to learn?*

Regarding the first perspective, we categorize the target knowledge into four distinct semantic dimensions. In other words, the model needs to clearly distinguish different semantic instances at each dimension. (1) **Action Semantics:** The ICL model must emulate human behavior at the categorical level of actions, such as grasping, placing, rotating, pressing, and translating. (2) **Compositional Semantics:** The ICL model should discern different structural compositions and sequential orderings of actions within human demonstrations. For instance, executing a rotation prior to picking up an object conveys a semantics fundamentally distinct from picking up the object before rotating it. (3) **Object Semantics:** The model must comprehend that it is manipulating a specific object rather than blindly mimicking spatial motion trajectories. Specifically, if the object's location during robot execution differs from that in the human demonstration, the model should align its execution with the object's actual pose rather than blindly replicating the demonstrator's spatial trajectory. (4) **Affordance Semantics:** When interacting with an object, the ICL model should learn *how* to manipulate it. For example, it ought to infer from the demonstration whether to grasp the body of a mug or its handle. While real-world deployments may occasionally demand ICL to condition on additional domain-specific factors, we posit that these four semantic tiers suffice to support the vast majority of common manipulation tasks without rendering ICL learning overly rigid.

Conversely, we further consider what factors ICL should not learn—that is, the aspects to which the model must exhibit sufficient robustness. First, as highlighted above, ICL should be robust against spatial mismatches of target objects between human demonstrations and actual robot execution environments. However, extreme positional disparities are treated as distinct semantics. Second, ICL must remain invariant to low-level nuances of human actions, such as execution speed, arm morphology, hand appearance, specific grasp angles, and personal reaching habits. Although certain scenarios might intentionally require fine-grained conditioning on these factors (e.g., matching the execution speed to the human demonstration), within the scope of this paper, we temporarily treat these factors as irrelevant noise to establish precise boundaries and clearly dissect the fundamental properties of ICL. Finally, ICL should also demonstrate strong robustness to visual scene variations, including lighting conditions, background colors, and non-target distractor objects.

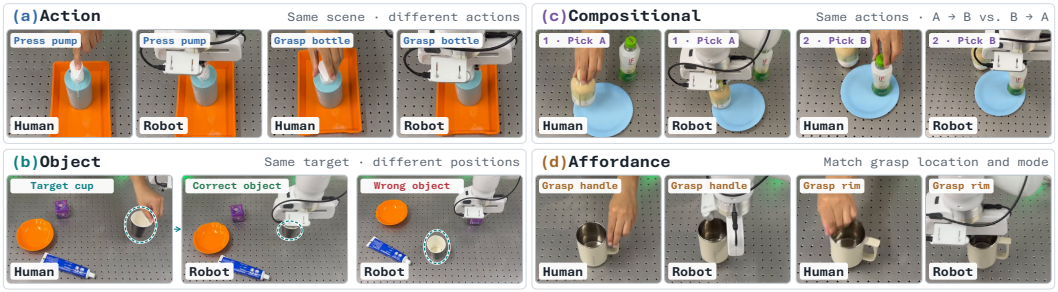



Figure 2: Four data types constructed for ICL based on our problem definition.

In summary, these definitions establish clear boundaries regarding the core capabilities and robustness expected from ICL in this work. This conceptualization enables us to provide human demonstrations conveniently while unambiguously communicating intent during execution.

### 3.2 Data Design and Collection Protocol

Based on the problem definitions of ICL introduced above, we propose a corresponding suite of data construction methods tailored to enforce these core semantic properties and invariances. Specifically, we first introduce semantic-discriminative data construction (Figure 2) as follows.

- **Action-discriminative data:** To compel the model to distinguish between different action semantics, we collect multiple demonstrations within the same scene using distinct actions. For example, given a toothpaste tube and a long box, one demonstration pair depicts grasping the tube and rotating it into the box, while another depicts grasping the tube and simply placing it on top of the box. The robot must differentiate these fine-grained actions and underlying intents.
- **Object-discriminative data.** To ground object semantics, we intentionally perturb object positions such that the relative spatial arrangement of the target object during human demonstration differs from that during robot execution. Furthermore, we introduce semantically distinct distractor objects around the target object. This design forces the model to learn the semantic identity of the target object rather than mechanically memorizing the spatial trajectory traversed by the human hand. Additionally, we enforce a strict uniqueness assumption—where only one instance of a specific target object exists per scene—to guarantee semantic unambiguousness.
- **Composition-discriminative data.** To capture compositional semantics, we design tasks sharing identical action sets but arranged in different temporal orders (e.g., grasping the banana first then the apple versus grasping the apple first then the banana), thereby compelling the model to attend to sequential dependencies.
- **Affordance-discriminative data.** To ground affordance semantics, we establish a human-robot consistency principle within each data pair: the human hand and the robot gripper must maintain identical grasping locations and modes. Across different pairs, however, the manipulation mode varies (e.g., one pair demonstrates grasping the body of a mug, whereas another pair demonstrates grasping its handle).

**Data augmentation for desired invariances** To suppress unwanted dependencies, we leverage targeted data augmentation strategies that force the model to ignore non-essential visual variations. (1) *Demonstrator invariance.* We collect videos across multiple human demonstrators featuring variations in hand appearance, including bare hands, different colored gloves, and diverse skin tones. (2) *Environmental invariance.* We systematically vary visual conditions, such as lighting, color temperature, and camera viewpoints in the conditioning videos while pairing them with the same robot execution trajectory, training the model to remain invariant to superficial scene variations.

**Group-based collection and augmentation strategy** While the individual data pair design is outlined above, our practical pipeline employs a *collection-group-based strategy*. Specifically, we define a collection group as a distinct scene coupled with a set of executable tasks within that scene, where each task corresponds to one collection instance. Upon completing a collection group, we

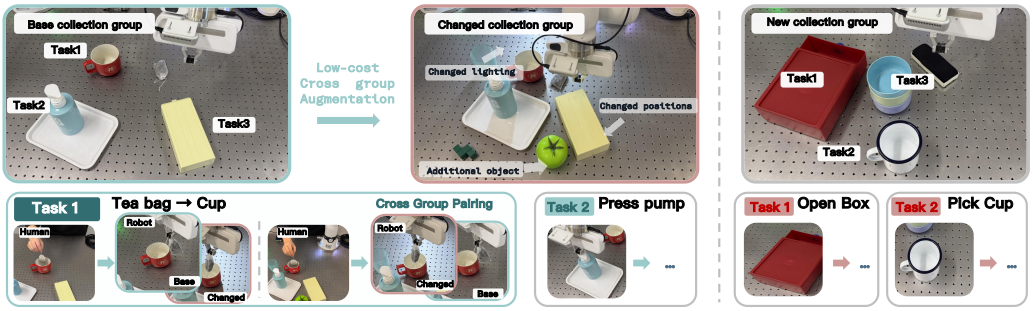



Figure 3: Schematic illustration of the real-world ICL instance construction.

transition to a new group via two types of scene switches: **Major transitions:** A complete change to an entirely new scene. **Minor transitions:** Perturbation-based variations within the existing scene, such as altering object placement, adjusting illumination, or swapping background distractor objects. These minor transitions allow us to conveniently construct new human-robot data pairs with perturbation-based augmentation across groups, significantly enriching the dataset at virtually zero additional collection cost. With this structured design, a single robotic platform can produce up to 3,000 valid human-robot data pairs per working day. Figure 3 provides a schematic illustration of the instance construction.

### 3.3 Unified Visual Prompt Encoding Module

To extract task-relevant information from the raw human demonstration video while filtering out irrelevant distractors, we design a unified visual prompt encoding module.

The module operates as follows. First, the raw conditioning video is processed by a pretrained visual encoder (e.g., a ViT or a frozen visual backbone) to produce a sequence of frame-level embeddings. These embeddings are then flattened and projected into a shared latent space. Next, a small set of learnable *context tokens*—initialized randomly—are introduced. Through a cross-attention mechanism, these tokens iteratively query the video embedding sequence, attending to the most task-relevant spatiotemporal cues while suppressing background noise and irrelevant hand movements. The final output of this module is a compact set of conditioned token embeddings. We concatenate these embeddings into the context of both video and action DiT, which originally contain text and state embeddings. They are then fed into the network via cross-attention, resulting in the final conditional policy. This design offers three key advantages: (1) *Arbitrary-length support*: the number of context tokens remains fixed regardless of video duration, enabling natural extension to long-horizon tasks. (2) *Task-relevant focus*: the attention-based querying mechanism explicitly suppresses distractors (hand appearance, background clutter) and attends to actions, temporal order, and object identities, aligning with our problem definition. We further analyze this in Section 4.4. (3) *End-to-end optimization*: the entire module is jointly differentiable with the downstream policy, avoiding brittle heuristic key-frame selection and ensuring coherent co-adaptation.

## 4 Experiments

### 4.1 Implementation Details.

The pretrained Wan2.2-5B (Wan et al., 2025) is utilized as the backbone. Following Yuan et al. (2026), we interpolate to obtain the 1B action DiT and set the action horizon to 32. Images from multiple cameras are concatenated into a single image before being fed into the VAE. 128 query tokens conduct cross-attention via the query-based module on the latent of the conditioning video, which is then fed into the context of both video and action DiTs. The query-based module stacks 4 blocks with a hidden dimension of 3072, resulting in 0.6B parameters. The video DiT is trained utilizing LoRA, while the other components are trained with full parameters.

### 4.2 Simulation Experiments

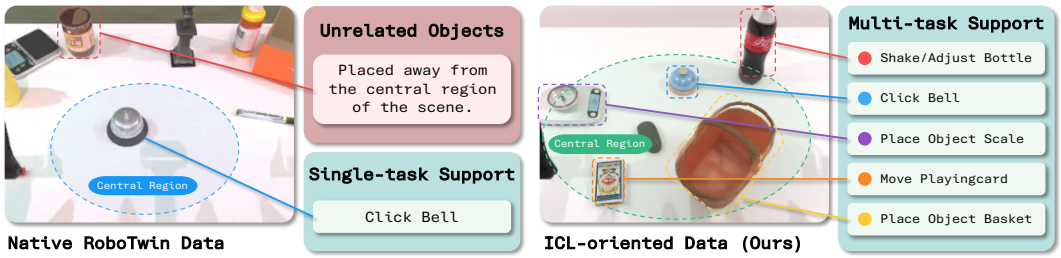



Figure 4: Left: Native RoboTwin data. Right: Proposed ICL-oriented data.

**Training data.** We adopt a widely used simulation framework RoboTwin 2.0 (Chen et al., 2025). We follow the real-world data setup in Section 3.2 to create ICL-oriented data in simulation. The native RoboTwin follows a task-centric data production paradigm. However, this paradigm cannot effectively evaluate the capability of ICL, as unrelated objects outside the action area do not constitute genuine task ambiguity, as shown on the left of Figure 4. Instead, our proposed data production pipeline first constructs a scene in which multiple task-relevant objects are simultaneously present within the reachable areas, and then collects multiple tasks under diverse object placements, as shown on the right of Figure 4. In this vein, we finally acquire ICL-oriented data, introducing genuine visual prompt ambiguity while retaining diverse scene variations, leading to 18,000 demonstrations for training. We also train our proposed model on 27,500 native RoboTwin demonstrations following Yuan et al. (2026); Bi et al. (2025); Li et al. (2026).

Table 1: Simulation evaluation results on the visual ambiguity test of RoboTwin tasks.

| Model | Trained on | Average |
|---|---|---|
| Fast-WAM | Native | 37.6 |
| SimpleICL (Ours) | Native | 40.3 |
| | ICL-oriented | **72.1** |
| w/o. video cond. | ICL-oriented | 61.6 |
| w/o. future pred. | ICL-oriented | 20.2 |

Table 2: Simulation evaluation results on novel (OOD) tasks. VC. refers to Video Condition. FP. refers to Future Prediction. Tasks are abbreviated.

| Task | Fast-WAM | Ours | w/o. VC. | w/o. FP. |
|---|---|---|---|---|
| Cup_Basket | 18.2 | 72.7 | 45.5 | 18.2 |
| Touch_Toycar | 8.3 | 58.3 | 8.3 | 0 |
| Mouse_Bowl | 25.7 | 63.3 | 20.0 | 16.7 |
| Block_Pad | 3.3 | 50.0 | 13.3 | 3.3 |
| Close_Laptop | 33.3 | 63.3 | 13.3 | 23.3 |
| Toycar_Box | 40.0 | 90.0 | 36.7 | 16.7 |
| Cards_Plate | 43.3 | 63.3 | 46.7 | 36.7 |
| **Average** | 24.6 | **65.9** | 26.3 | 17.9 |

**Test metrics and models.** For evaluation, we utilize our ICL-oriented data pipeline combining partial RoboTwin tasks to form the visual ambiguity test. We further design 7 novel (OOD) tasks (Figure 7) to evaluate our models that are neither natively supported by RoboTwin nor present in the training set. These tasks for evaluation are produced under ICL-oriented data pipeline as well. For the sake of the experiment, we utilize AGILEX for both the executor entity and the demonstrator (conditioning) platform. Finally, we have the following findings.

**ICL-oriented data helps resolve visual ambiguity.** Table 1 shows the model trained on ICL-oriented data achieves higher success rates under visual ambiguity even without video conditioning.

**Video conditioning enables in-context learning of novel tasks.** While the model without video conditioning performs reasonably well on visual ambiguity, its performance drops sharply on novel OOD tasks (as shown in Table 2). This indicates that video conditioning enables the model to leverage demonstrations for in-context learning of unseen tasks.

**Future prediction is crucial for effective video-conditioned ICL.** Removing future prediction causes substantial drops on both visual ambiguity and OOD tasks, from 72.1% to 20.2% and from 65.9% to 17.9%, respectively. This suggests that future prediction is crucial for effectively leveraging task-relevant temporal information from video demonstrations.

### 4.3 REAL-WORLD EXPERIMENTS

We instantiate the real-world data collection and evaluation pipeline following the framework established in Section 3.2. We conduct experiments using FRANKA robotic platform, equipped with three Intel RealSense D435 cameras. The teleoperation is conducted with 3D connexion space mouse.

Table 3: Real-world evaluation on eight unseen tasks. Success rates (%) are reported for each method. Task abbreviations: **Wipe** = wiping a plate; **Tea** = serving tea; **Drawer** = pulling a drawer; **Soap** = pressing a soap dispenser; **Fruit** = placing fruit; **Chopsticks** = organizing chopsticks and bowls. **Shelf** = organizing items on a shelf. Please refer to appendices for detailed task design.

| | Method | Weighing | Wipe | Tea | Drawer | Soap | Fruit | Chopsticks | Shelf | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| **Easy** | FAST-WAM | 0.90 | 0.87 | 0.83 | 0.80 | 0.63 | 0.90 | 0.73 | 0.77 | 0.80 |
| | $\pi_{0.5}$ | 0.87 | 0.97 | 0.80 | 0.60 | 0.73 | 0.60 | 0.53 | 0.67 | 0.72 |
| | **SimpleICL (ours)** | 0.93 | 0.76 | 0.73 | 0.83 | 0.73 | 0.86 | 0.80 | 0.83 | **0.81** |
| **Medium** | FAST-WAM | 0.83 | 0.50 | 0.63 | 0.43 | 0.60 | 0.83 | 0.63 | 0.73 | 0.65 |
| | $\pi_{0.5}$ | 0.63 | 0.46 | 0.37 | 0.43 | 0.57 | 0.47 | 0.23 | 0.57 | 0.47 |
| | **SimpleICL (ours)** | 0.83 | 0.66 | 0.60 | 0.80 | 0.67 | 0.63 | 0.73 | 0.63 | **0.70** |
| **Hard** | FAST-WAM | 0.40 | 0.60 | 0.27 | 0.53 | 0.33 | 0.57 | 0.27 | 0.37 | 0.42 |
| | $\pi_{0.5}$ | 0.53 | 0.33 | 0.20 | 0.17 | 0.47 | 0.40 | 0.10 | 0.30 | 0.31 |
| | **SimpleICL (ours)** | 0.80 | 0.67 | 0.57 | 0.67 | 0.63 | 0.70 | 0.76 | 0.60 | **0.68** |

#### 4.3.1 MAIN RESULTS

**Real-world Task Setup.** We evaluate our method on eight unseen novel tasks. These tasks are: *weighing, wiping a plate, serving tea, pulling a drawer, pressing a soap dispenser, placing fruit, organizing chopsticks, and organizing items on a shelf.* Critically, the sponge and plate used in the wiping task, the drawer in the pulling task, the fruit in the placing task, the items to be organized in the shelf task, the soap dispenser being pressed, and the target teacup in the tea-serving task are all objects that never appeared in the training set. This ensures that the evaluation genuinely tests the model's ability to generalize to novel objects and tasks through in-context learning from human demonstrations.

**Evaluation Protocol.** We measure model capability across three levels of difficulty. **Easy**: the scene contains only one executable task with a unique way to perform it. This is the most commonly used evaluation setting in current literature. **Medium**: the scene contains multiple distinct executable tasks, requiring the model to possess discriminative capability to select the correct task. **Hard**: the task itself admits multiple valid execution modes. For example, in the plate-wiping task, two plates are placed adjacent to each other, and the model must correctly determine which one to wipe. This level demands the strongest discrimination ability and is the most challenging.

**Results and Findings.** The quantitative results are reported in Table 3, organized into these three levels. Each entry represents the success rate (%) of the corresponding method. We compare our method against two baselines: FAST-WAM and $\pi_{0.5}$. Figure 5 presents a qualitative comparison between our method and existing baselines. Based on these results, we have the following findings.

- **ICL model has better generalization to novel objects and scenes.** At the Easy Level, where the scene contains only a fixed task, the baseline methods also perform reasonably well, but our model achieves a higher success rate. This indicates that visual in-context learning endows the model with better generalization.
- **ICL model can accurately identify and follow human actions**. Although the tasks involve diverse action types, such as placing, pulling, and pressing, our model consistently follows the action type demonstrated by humans with high probability and successfully completes the tasks, with little confusion across different action types.
- **ICL model can accurately identify the semantic identity of the target object**, rather than simply imitating actions or memorizing spatial locations. In the hard-level tasks, we introduce distractor objects that are highly similar to the target in both spatial location and semantic identity. Even under such challenging conditions, our ICL model maintains a high task success rate.
- **ICL model better disentangles scene observations from conditioned task intents**. In the medium- and hard-level tasks, the same observation may correspond to multiple task intents. Language-conditioned models, such as Fast-WAM, often exhibit shortcut learning, binding a particular action intent to a fixed scene pattern rather than properly interpreting the conditioning signal. Figure 5 illustrates qualitative examples.

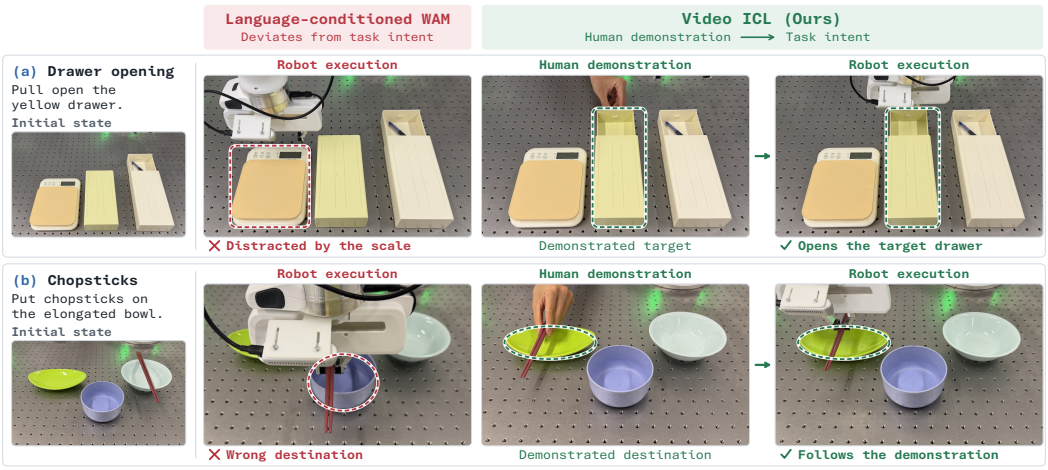



Figure 5: Qualitative comparison between our method and existing baselines. At the **Medium** level (top), FAST-WAM is distracted by a nearby scale, while our method correctly locates the drawer. At the **Hard** level (bottom), FAST-WAM mistakenly places chopsticks onto the round bowl, whereas ours reaches the correct destination.

### 4.3.2 Analysis of Composition and Affordance

In this section, we conduct a further dedicated analysis of semantics learning of composition and affordance.

Table 4: Analysis of composition and affordance under discriminative sampling (ours) versus without discriminative sampling. The metric is intent-following rate (%). †: DC means discriminative data collection.

| Setting | Composition | | | Affordance | | |
|---|---|---|---|---|---|---|
| | Snack | Cup | Bag | Cup | Measure | Tape |
| w/. DC† | **0.97** | **1.00** | 0.90 | **0.87** | **0.77** | **0.83** |
| w/o. DC | 0.93 | 0.97 | **0.93** | 0.60 | 0.40 | 0.47 |

**Experiment Control of Composition.** For the analysis of composition, we collect training data where tasks are executed in the same scene in different temporal orders. For example, we place two visually distinct apples into a fruit basket in different orders, and the two demonstrations with different manipulation orders are all recorded as training data. Conversely, for comparison, we disable composition discriminative data by only incorporating a specific order of demonstration in a scene into the training set. For evaluation, we evaluate on the task of sequentially organizing items on a table, using novel items never seen during training: snacks, cups, and toys.

**Experiment Control of Affordance.** For the analysis of affordance, we collect data where the same object is manipulated multiple times with different affordances. To disable this strategy, for each object, we only collect demonstrations involving the same affordance for training. For example, for black mugs, we always grasp the mug body, whereas for white mugs, we always grasp the handle. For evaluation, we use three classic objects with multiple affordances: a tape measure (grasping the ring or the body), a cup with a handle (grasping the handle or the body), and tape (grasping the left side or the right side).

Table 5: Robutness analysis. Success rates (%) are reported under multiple disturbance categories. †: CP means cross-group pairing strategy for data augmentation.

| Method | Orig. | Env | | Hand | | Obj. Disturb | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | | Light | Bg. | App. | Sty. | Loc. | Clut. | |
| **Ours** | **0.81** | 0.78 | 0.79 | 0.77 | 0.80 | 0.79 | 0.75 | **0.78** |
| w/o. CP† | 0.80 | 0.73 | 0.71 | 0.78 | 0.79 | 0.51 | 0.66 | 0.70 |

Table 4 demonstrates the experimental results. Here we have the following findings.

- Our model demonstrates strong **composition discrimination** and **affordance discrimination** capabilities, allowing it to effectively distinguish and follow different composition and affordance patterns exhibited in human demonstrations.

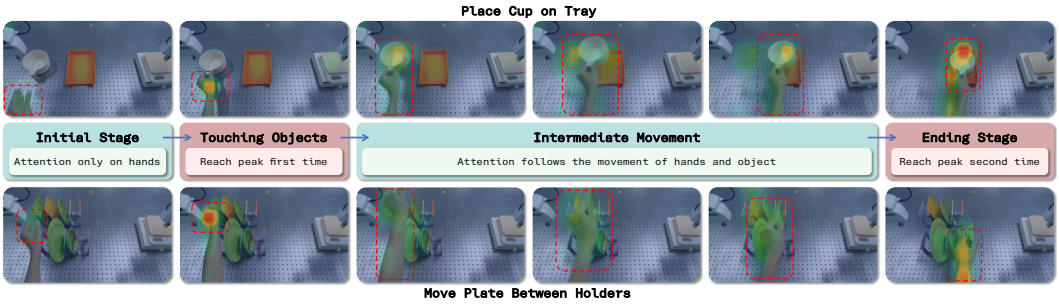



Figure 6: The timeline of attention visualization heat map for the human conditioning videos on two real tasks of *Place Cup on Tray* and *Move Plate Between Holders*.

- For **composition discrimination**, it is unnecessary to explicitly construct data with different compositions for the same scene. The ICL model can spontaneously learn to disentangle scene content from composition patterns directly from the training data.
- In contrast, for **affordance discrimination**, training data with different affordances must be explicitly constructed for the same object to force the model to acquire this capability. Under our current setting, the ICL model cannot spontaneously disentangle object identity from affordance. This indicates that within the scope of ICL, fine-grained affordance may be harder to learn than global temporal and compositional relationships.

#### 4.3.3 Robustness Analysis

As stated in the method section, we argue that ICL should ignore task-irrelevant disturbances present in the human demonstration video context. In this section, we analyze the impact of such disturbances on our model's execution performance.

We create different disturbances. Specifically, **Light** denotes varying illumination, and **Bg.** denotes adding background color distractions. **App.** denotes changing the appearance of the hand, e.g., wearing a glove, and **Sty.** denotes different human subjects and motion styles, e.g., testing with a different person from those in training. **Loc.** denotes that during robot execution the target object position is shifted relative to the human demonstration, and **Clut.** denotes that irrelevant clutter appears in the scene during execution relative to the demonstration. As shown in Table 5, our model effectively handles different disturbances. In addition, when the grouped pairing strategy is not used, each robot trajectory is paired with a human video whose environment and object positions are exactly identical. This causes the model to bind these task-irrelevant factors to the task itself, leading to a drop in success rate.

### 4.4 Attention Visualization

Figure 6 demonstrates a timeline of attention heat map for the human conditioning videos during task execution. We discover that: **a) At the initial stage,** the attention only appears on the human hands; **b) The peaks of attention intensity occur twice:** when the hand first touches on the target object and when the task is completed, respectively; **c) During the intermediate stage,** the attention follows the movement of both hands and the object with moderate intensity. These observations suggest that the model exhibits temporally structured and interaction-centric attention. Rather than uniformly attending to the scene, it focuses on the action-relevant cues at different execution stages, with pronounced responses at key interaction events such as object contact and task completion.

## 5 Conclusion

In this paper, we formulate a series of open questions surrounding the emerging area of robot in-context learning. To address these questions, we propose a specially designed data paradigm, an efficient data collection pipeline, and a network module for video conditioning. We validate these contributions in both simulation and real-world environments. More importantly, our experiments reveal several insightful findings regarding the nature of robot ICL. We believe that this field remains a promising and important direction that warrants further investigation.

## AI USE STATEMENT

We used OpenAI ChatGPT 5.6 solely to assist with method implementation, language translation, and the polishing of figure formatting. ChatGPT 5.6 was not used for research design, data collection, data analysis, result interpretation, scientific reasoning, or drawing conclusions. All scientific content, experimental results, and conclusions were developed and verified by the authors.

## REFERENCES


Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model, 2025. URL https://arxiv.org/abs/2512.13030.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi_0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Jun Cen, Siteng Huang, Yuqian Yuan, Kehan Li, Hangjie Yuan, Chaohui Yu, Yuming Jiang, Jiayan Guo, Xin Li, Hao Luo, Fan Wang, Fan Wang, and Deli Zhao. Rynnvla-002: A unified vision-language-action and world model. *arXiv preprint arXiv:2511.17502*, 2025a.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:*, 2025b.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Guangyan Chen, Meiling Wang, Te Cui, Zichen Zhou, Qi Shao, Shalfun Li, Hang Su, Roy Gan, Hao Wang, Mengyin Fu, et al. Robots acquire manipulation skills in seconds from a single human video. *arXiv preprint arXiv:2607.20033*, 2026.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023. URL https://arxiv.org/abs/2302.00111.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation, 2025. URL https://arxiv.org/abs/2507.12898.

Yusen Feng, Bingchen Han, Jiangran Lyu, Kai Liu, Yixin Zheng, Yuxuan Wan, Weiheng Liu, Sun Han, Ruiqin Li, Yulong Zhang, et al. Wam-ttt: Steering world-action models by watching human play at test time. *arXiv preprint arXiv:2607.06988*, 2026.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. Dreamgen: Unlocking generalization in robot learning through video world models, 2025. URL https://arxiv.org/abs/2505.12705.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint 2512.15692*, 2025.

Jeongeun Park, Juhan Park, Taekyung Kim, Sungjoon Choi, Dongyoon Han, and Sangdoo Yun. Retrieve, don't retrain: Extending vision language action models to new tasks at test time. *arXiv preprint arXiv:2606.15631*, 2026.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Galaxea Team. Galaxea g0: Open-world dataset and dual-system vla model. *arXiv preprint arXiv:2509.00576v1*, 2025.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Generalist Team. Gen-1.5: Embodied foundation models are one-shot learners. *Generalist AI Blog*, 2026. https://generalistai.com/blog/gen-1.5.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang,

Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025.

John Won, Kyungmin Lee, Huiwon Jang, Dongyoung Kim, and Jinwoo Shin. Dual-stream diffusion for world-model augmented vision-language-action model, 2025. URL https://arxiv.org/abs/2510.27607.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation, 2023.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026. URL https://arxiv.org/abs/2603.16666.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. *CoRR*, abs/2507.04447, 2025. doi: 10.48550/ARXIV.2507.04447. URL https://doi.org/10.48550/arXiv.2507.04447.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models, 2025. URL https://arxiv.org/abs/2503.22020.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. Flare: Robot learning with implicit world modeling, 2025. URL https://arxiv.org/abs/2505.15659.

Jiaming Zhou, Qihang Zhang, Gangwei Xu, Cunxin Fan, Yujie Zhao, Ruilin Wang, Yiming Luo, Shuai Yang, Xing Zhu, Yujun Shen, et al. Zero-wam: In-context world-action modeling from human videos for open-ended task generalization. *arXiv preprint arXiv:2608.26103*, 2026.

Pengfei Zhou, Liliang Chen, Shengcong Chen, Di Chen, Wenzhi Zhao, Rongjun Jin, Guanghui Ren, and Jianlan Luo. Act2goal: From world model to general goal-conditioned policy, 2025. URL https://arxiv.org/abs/2512.23541.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets, 2025. URL https://arxiv.org/abs/2504.02792.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

## A Simulation Novel (OOD) Tasks

We design 7 novel (OOD) tasks for simulation evaluation; from left to right, they are *Put Toycar in Plasticbox*, *Put Cup in Basket*, *Touch Toycar*, *Place Playingcards Plate*, *Close Laptop*, *Move Block on Pad*, and *Place Mouse in Bowl*.

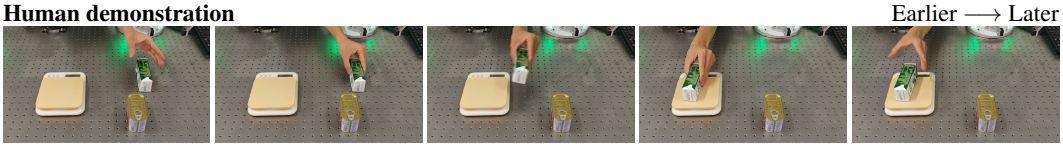



Figure 7: Novel (OOD) tasks for simulation evaluation.

## B Real-World Demonstrations

The supplementary material will include **a demo video** of all eight real-world evaluation tasks. Below, each task is illustrated by five human demonstration keyframes (top row) and five corresponding robot execution keyframes (bottom row), ordered chronologically from left to right.

**Human demonstration** Earlier ⟶ Later



**Robot execution**

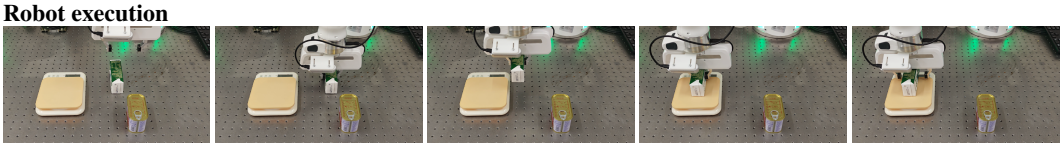
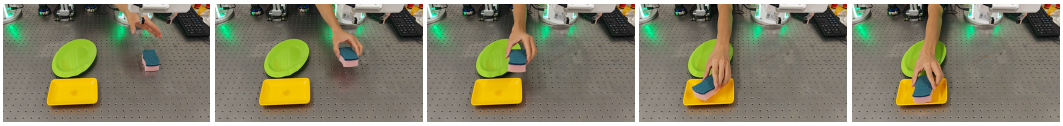


Figure 8: **Weighing.**

**Human demonstration** Earlier ⟶ Later



**Robot execution**

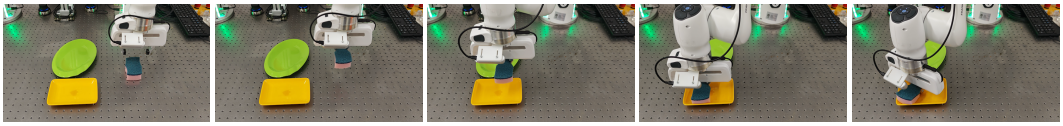


Figure 9: **Wiping a plate.**

**Human demonstration** Earlier ⟶ Later



**Robot execution**



Figure 10: **Serving tea.**

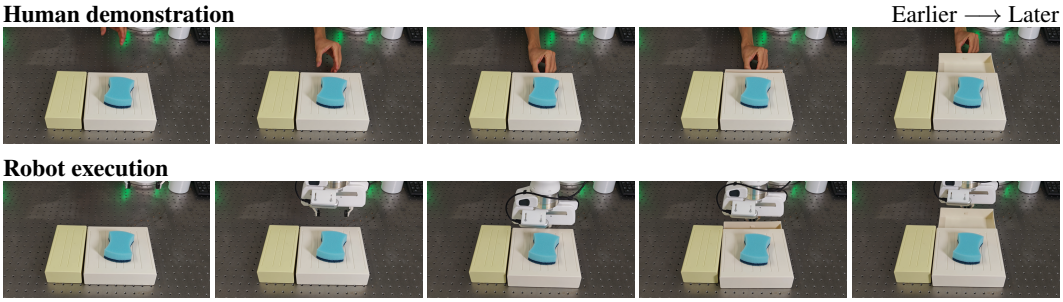



Figure 11: **Pulling a drawer.**

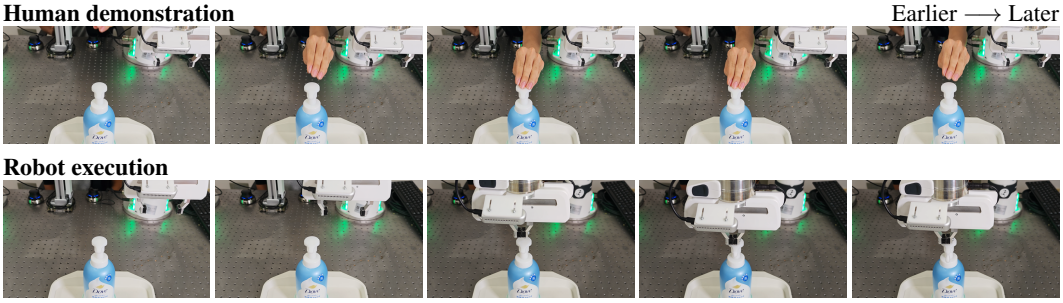



Figure 12: **Pressing a soap dispenser.**

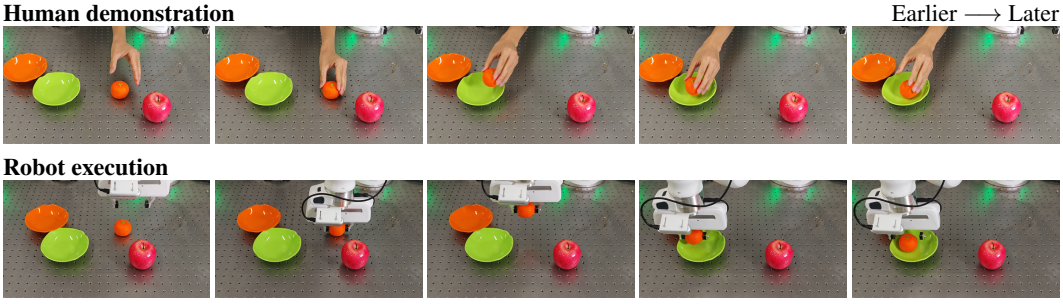



Figure 13: **Placing fruit.**

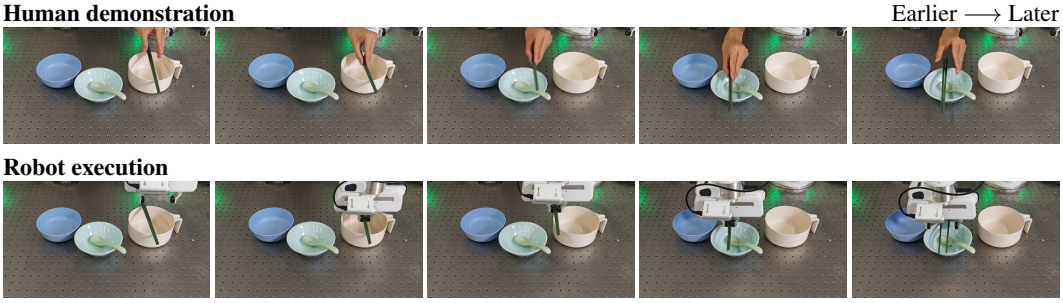



Figure 14: **Organizing chopsticks and bowls.**

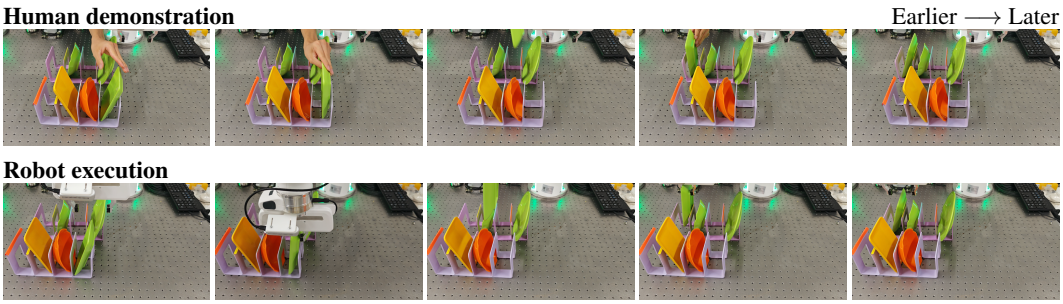



Figure 15: **Organizing items on a shelf.**