

# Report # 2: RT-1 Testing w.r.t Image Variations for Pre-defined skills

Date: 18<sup>th</sup> May, 2025

## Introduction

In **Report #1**, we evaluated RT-1’s zero-shot inference on two symbolic skills (grasp(apple) and move\_to(table)) across diverse visual and linguistic perturbations. Key weaknesses emerged around clutter, lighting, and spatial language. To address these, in **Report #2** I have:

1. **Expanded the symbolic skill set** to 10 skills drawn directly from the RT-1 paper.
2. **Diversified scene setups** with richer image variations and new categories.
3. **Varied instruction phrasing** using sentence templates from the official RT-1 publication .

This report analyzes RT-1’s outputs for each test case by **Instruction**, **Selected Skill**, and **Confidence Score**, contrasts them with expectations, and draws unbiased interpretations.

## Previous Findings from RT-1:

We evaluated RT-1’s zero-shot inference over six image + text variations, measuring whether it consistently maps to the correct symbolic skill at  $\geq 40\%$  similarity.

Variation	Correct Skill?	Confidence Range	Main Issue
Color Variation	Yes	67.1%	Stable across synonyms; strong language generalization.
Low Table Height	Yes	67.1%	Unaffected by geometry—but hints at limited 3D spatial cues.
Background Clutter	No	60.9%	Misclassified grasp as move; fails under visual noise.
Low-Contrast / Harsh	No	45.3–60.8%	Confidence borderline; degrades with poor lighting.
Apple Position Variation	No	37.4–60.9%	Misinterprets spatial phrasing; falls below threshold.

Variation	Correct Skill?	Confidence Range	Main Issue
Distractor Objects	No	60.8–69.5%	High confidence but wrong skill; lacks fine-grained parsing.

## Experimental Setup

- Model: RT-1 with MaxViT image encoder (depths 2,2,5,2; window 7; dropout 0.1) and 6-layer, 8-head Transformer decoder.
- Skills: 10 symbolic skills, each with a 7-dim prototype action mapping (e.g. pick(object), open\_drawer(top), pull\_napkin(dispenser), etc.).
- Variations: Six test categories—Color Variations; Clutter & Distractors; Drawer Interactions (Closing & Opening); Bottle Handling & Placement; Napkin Dispenser Trials; Jar Opening.
- Metrics: For each instruction–image pair, record the Selected Skill (highest cosine similarity to a prototype) and Confidence Score (similarity × 100).



### Category: Color Variations



Instruction variations:

1. Pick up the apple on the green surface
2. Grasp the red fruit in front of you.
3. Pick the apple

Selected Skill	Confidence Score	Model Output (tokenized representations of the action parameters)
knock_over(object)	52.3%	{           "indices": [             78,             241,             107,             225,             51,             199,             74           ],           "continuous": [             -0.3880000114440918,             0.8899999856948853,             -0.1609999845027924,             0.7649999856948853,             -0.6000000238418579,             0.5609999895095825,             -0.41999998688697815           ]         }
knock_over(object)	49.7%	{           "indices": [             78,             241,             107,             225,             74,             199,             74           ],           "continuous": [             -0.3880000114440918,             0.8899999856948853,             -0.1609999845027924,             0.7649999856948853,             -0.41999998688697815,             0.5609999895095825,             -0.41999998688697815           ]         }
pull_napkin(dispenser)	41.1%	{           "indices": [             78,             241,             107,             225,             201,             199,             74           ],           "continuous": [             -0.3880000114440918,             0.8899999856948853,             -0.1609999845027924,             0.7649999856948853,             0.5759999752044678,             0.5609999895095825,             -0.41999998688697815           ]         }





Instruction variations:

1. Ignore the balls and pick the green apple.
2. Select the apple, not the surrounding red spheres.
3. Pick(object)

Selected Skill	Confidence Score	Model Output (tokenized representations of the action parameters)
knock_over(object)	49.7%	{ "indices": [ 78, 241, 107, 225, 74, 199, 74 ], "continuous": [ -0.3880000114440918, 0.8899999856948853, -0.16099999845027924, 0.7649999856948853, -0.4199999868697815, 0.5609999895095825, -0.4199999868697815 ] }
pull_napkin(dispenser)	41.1%	{ "indices": [ 78, 241, 107, 225, 74, 199, 74 ], "continuous": [ -0.3880000114440918, 0.8899999856948853, -0.16099999845027924, 0.7649999856948853, -0.4199999868697815, 0.5609999895095825, -0.4199999868697815 ] }

knock_over(object)	49.7%	{ "indices": [ 78, 241, 107, 225, 74, 199, 74 ], "continuous": [ -0.3880000114440918, 0.8899999856948853, -0.16099999845027924, 0.7649999856948853, -0.41999998688697815, 0.5609999895095825, -0.41999998688697815 ] }
--------------------	-------	---



**Category: Clutter & Distractors**



Instruction variations:

- 1. Pick the green snack bag from the top-left corner of the table
- 2. Pick up the red soda can lying on its side next to the blue can.
- 3. Place the white bowl on the far right edge of the table.

Selected Skill	Confidence Score	Model Output (tokenized representations of the action parameters)
knock_over(object)	49.7%	{ "indices": [ 78, 241, 107, 225, 74, 199, 74 ], "continuous": [ 

		<pre>-0.3880000114440918, 0.8899999856948853, -0.16099999845027924, 0.7649999856948853, -0.41999998688697815, 0.5609999895095825, -0.41999998688697815 ] }</pre>
knock_over(object)	52.3%	<pre>{   "indices": [     78,     241,     107,     225,     51,     199,     74   ],   "continuous": [     -0.3880000114440918,     0.8899999856948853,     -0.16099999845027924,     0.7649999856948853,     -0.6000000238418579,     0.5609999895095825,     -0.41999998688697815   ] }</pre>
knock_over(object)	52.3%	<pre>{   "indices": [     78,     241,     107,     225,     51,     199,     74   ],   "continuous": [     -0.3880000114440918,     0.8899999856948853,     -0.16099999845027924,     0.7649999856948853,     -0.6000000238418579,     0.5609999895095825,     -0.41999998688697815   ] }</pre>





Instruction variations:

1. Pick up the green bag of chips
2. Knock over the blue can
3. Pick up the red can

Selected Skill	Confidence Score	Model Output (tokenized representations of the action parameters)
open_drawer(top)	43.7%	<pre>{   "indices": [     78,     241,     107,     250,     74,     199,     74   ],   "continuous": [     -0.3880000114440918,     0.8899999856948853,     -0.1609999845027924,     0.9610000252723694,     -0.41999998688697815,     0.5609999895095825,     -0.41999998688697815   ] }</pre>
knock_over(object)	52.3%	<pre>{   "indices": [     78,     241,     107,     225,     51,     199,     74   ],   "continuous": [     -0.3880000114440918,     0.8899999856948853,     -0.1609999845027924,     0.7649999856948853,     -0.6000000238418579,     0.5609999895095825,     -0.41999998688697815   ] }</pre>

knock_over(object)	49.7%	{ "indices": [ 78, 241, 107, 225, 74, 199, 74 ], "continuous": [ -0.3880000114440918, 0.8899999856948853, -0.16099999845027924, 0.7649999856948853, -0.41999998688697815, 0.5609999895095825, -0.41999998688697815 ] }
--------------------	-------	---



**Category: Drawer Interactions Closing**

Instruction: Close the door

Selected Skill: knock\_over(object)

Confidence Score: 49.7 %






Robot arm task setup x RT-1: Robotics Transl x rt1\_viewer.png (3302) x robotics-transformer x evals.png (705x538) x robotics-transformer x Presentation Codes x RT-1 Testing on vari x RT-1 Robotics Contr x

c4f5f76d09c38db55.gradio.live

### RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Close the drawer

Clear Submit

Action Output

```

1 1
2 2
3 3
4 4
5 5
6 6
7 7
8 8
9 9
10 10
11 11
12 12
13 13
14 14
15 15
16 16
17 17
18 18
19 19
20 20

```

Selected Skill

knock\_over(object)

Confidence

49.7%

Use via API Built with Gradio Settings

Instruction: Close the drawer upfront

Selected Skill: pull\_napkin(dispenser)


Confidence Score: 41.1 %



RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Close the drawer upfront

Clear Submit

Action Output

```

1  [
2    "Indices": [
3      "0": 78,
4      "1": 281,
5      "2": 187,
6      "3": 235,
7      "4": 281,
8      "5": 199,
9      "6": 74
10   ],
11   "Continuous": [
12     "0": -0.3588688114448918,
13     "1": 0.889999985648853,
14     "2": -0.1609999845627924,
15     "3": 0.764999985648853,
16     "4": 0.5759999752044678,
17     "5": 0.568999985648853,
18     "6": -0.419999985648853
19   ]
20 ]

```

Selected Skill

pull\_napkin(dispenser)

Confidence

41.1%

Use via API Built with Gradio Settings

Instruction: Close the drawer

Selected Skill: open\_drawer(top)

Confidence Score: 43.7 %




Robot arm task setup x RT-1: Robotics Traini x rt1\_viewer.png (3200 x 1800) x robotics-transformer x evals.png (703x538) x robotics-transformer x Presentation Codes x RT-1 Testing on vari x RT-1 Robotics Contro x

c4f5f76d09c38db55.gradio.live

### RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Close the drawer

Clear Submit

Action Output

```
1  [
2    "Indices": [
3      "0": 78,
4      "1": 281,
5      "2": 187,
6      "3": 258,
7      "4": 74,
8      "5": 199,
9      "6": 74
10   ],
11   "Continuous": [
12     "0": -0.3886680114440918,
13     "1": 0.889999985648853,
14     "2": -0.16099999845627924,
15     "3": 0.9610000252723694,
16     "4": -0.4199999868697815,
17     "5": 0.560999989585825,
18     "6": -0.4199999868697815
19   ]
20 ]
```

Selected Skill

open\_drawer(top)

Confidence

43.7%

Use via API Built with GradIO Settings

Instruction: Close the drawer

Selected Skill: open\_drawer(top)

Confidence Score: 43.7 %




Robot arm task setup   RT-1: Robotics Traini...   rt1\_baserg.png (3200...   robotics-transformer...   evals.png (703x538)   robotics-transformer...   Presentation Codes...   CO: RT-1 Testing on vari...   RT-1 Robotics Contro...   c4f5f76d09c38db55.gradio.live

### RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Close the drawer

Clear Submit

Action Output

```

1  [
2    "Indices": [
3      "0": 78,
4      "1": 281,
5      "2": 187,
6      "3": 258,
7      "4": 74,
8      "5": 199,
9      "6": 74
10   ],
11   "Continuous": [
12     "0": -0.3580680114440919,
13     "1": 0.889999995640853,
14     "2": -0.16099999845627924,
15     "3": 0.9610000252723694,
16     "4": -0.41999999860697815,
17     "5": 0.56099999895095825,
18     "6": -0.41999999860697815
19   ]
20 ]

```

Selected Skill

open\_drawer(top)

Confidence

43.7%

Use via API   Built with GradIO   Settings

Instruction: Close the drawer

Selected Skill: open\_drawer(top)


Confidence Score: 43.7 %



RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Close the drawer

Clear Submit

Action Output

```

1  [
2    "Indices": [
3      "0": 78,
4      "1": 281,
5      "2": 187,
6      "3": 258,
7      "4": 74,
8      "5": 199,
9      "6": 74
10   ],
11   "Continuous": [
12     "0": -0.3580680114440919,
13     "1": 0.889999995640853,
14     "2": -0.16099999845627924,
15     "3": 0.9610000252723694,
16     "4": -0.41999999860697815,
17     "5": 0.56099999895895825,
18     "6": -0.41999999860697815
19   ]
20 ]

```

Selected Skill

open\_drawer(top)

Confidence

43.7%

Use via API Built with Gradio Settings

## Category: Drawer Interactions Opening

Instruction: Open the top drawer

Selected Skill: knock\_over(object)

Confidence Score: 49.7 %




Robot arm task setup x RT-1: Robotics Traini x rt\_1easing.png (3200 x 1800) x robotics-transformer x evals.png (703x538) x robotics-transformer x Presentation Codes x RT-1 Testing on vari x RT-1 Robotics Contro x

c4f5f76d09c38db55.gradio.live

### RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Open the top drawer

Clear Submit

Action Output

```

1  [
2    "Indices": [
3      "0": 78,
4      "1": 281,
5      "2": 187,
6      "3": 235,
7      "4": 74,
8      "5": 199,
9      "6": 74
10   ],
11   "Continuous": [
12     "0": -0.3586680114440919,
13     "1": 0.889999985648853,
14     "2": -0.16099999845627924,
15     "3": 0.7649999985648853,
16     "4": -0.41999999866977815,
17     "5": 0.5609999985695825,
18     "6": -0.41999999866977815
19   ]
20 ]

```

Selected Skill

knock\_over(object)

Confidence

49.7%

Use via API Built with GradIO Settings

Instruction: Open the top drawer

Selected Skill: knock\_over(object)


Confidence Score: 49.7 %



RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Open the top drawer

Clear Submit

Action Output

```

1 1
2 2
3 3
4 4
5 5
6 6
7 7
8 8
9 9
10 10
11 11
12 12
13 13
14 14
15 15
16 16
17 17
18 18
19 19
20 20

```

Selected Skill

knock\_over(object)

Confidence

49.7%

Use via API Built with Gradio Settings

Instruction: Open the middle drawer

Selected Skill: open\_drawer(top)

Confidence Score: 43.7 %




Robot arm task setup x rt1.pdf x robotics-transformer x evals.png (705x538) x robotics-transformer x Presentation Codes x RT-1 Testing on vari RT-1 Robotics Contn "pull napkin out of d... c4f5f76d09c38db55.gradio.live

### RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Open the middle drawer

Clear Submit

Action Output

```
1 1
2 2
3 3
4 4
5 5
6 6
7 7
8 8
9 9
10 10
11 11
12 12
13 13
14 14
15 15
16 16
17 17
18 18
19 19
20 20
```

Selected Skill

open\_drawer(top)

Confidence

43.7%

Use via API Built with Gradio Settings

Instruction: Open the top drawer

Selected Skill: open\_drawer(top)

Confidence Score: 43.7 %





Robot arm task setup

rt1.pdf

robotics-transformer

evals.png (705x338)

robotics-transformer

Presentation Codes

RT-1 Testing on vari

RT-1 Robotics Contr

"pull napkin out of d

c4ff576d09c38db55.gradio.live

RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input

Instruction

Open the top drawer

Clear

Submit

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

"indices": [

"0": 78,

"1": 241,

"2": 197,

"3": 250,

"4": 74,

"5": 199,

"6": 74

],

"continuous": [

"0": -0.3880000114448918,

"1": 0.8899999956948853,

"2": -0.16099999945827924,

"3": 0.9610000252723694,

"4": -0.42199999868697815,

"5": 0.5609999985095825,

"6": -0.41999999868697815

]

}

Selected Skill

open\_drawer(top)

Confidence

43.7%

Use via API Built with Gradio Settings



## Category: Bottle Handling & Placement

Instruction: Place the bottle upright on the table

Selected Skill: open\_drawer(top)

Confidence Score: 43.7 %




Robot arm task setup x RT-1: Robotics Transl x rt1\_viewer.png (3302) x robotics-transformer x evals.png (705x538) x robotics-transformer x Presentation Codes x RT-1 Testing on vari x RT-1 Robotics Contr x

c4f5f76d09c38db55.gradio.live

### RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Place the bottle upright on the table

Clear Submit

Action Output

```
1 1
2 2
3 3
4 4
5 5
6 6
7 7
8 8
9 9
10 10
11 11
12 12
13 13
14 14
15 15
16 16
17 17
18 18
19 19
20 20
```

Selected Skill

open\_drawer(top)

Confidence

43.7%

Instruction: Place the bottle upright on the table


Selected Skill: knock\_over(object)

Confidence Score: 49.7 %



Robot arm task setup x RT-1: Robotics Transl x rt1\_tester.png (330) x robotics-transformer x evals.png (705x538) x robotics-transformer x Presentation Codes x CD RT-1 Testing on vari x RT-1 Robotics Contr x

c4f5f76d09c38db55 gradio.live



```
8  "5": 199,  
9  "6": 74  
10 ],  
11   "continuous": [  
12     "0": -0.38088889114448918,  
13     "1": 0.08999999856948853,  
14     "2": -0.16099999845827924,  
15     "3": 0.76499999856948853,  
16     "4": -0.41999999868697815,  
17     "5": 0.5609999985095825,  
18     "6": -0.41999999868697815  
19   ]  
20 ]
```

B. Selected Skill

**knock\_over(object)**

B. Confidence

49.7%

Instruction

Place the bottle upright on the table

Clear Submit

Instruction: Place coke can into paper bowl

Selected Skill: pull\_napkin(dispenser)


Confidence Score: 41.1 %



RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

place coke can into paper bowl

Clear Submit

Action Output

```
1  {
2    "Indices": [
3      "0": 78,
4      "1": 241,
5      "2": 197,
6      "3": 225,
7      "4": 201,
8      "5": 199,
9      "6": 74
10   ],
11   "Continuous": [
12     "0": -0.3880689114448918,
13     "1": 0.8899999956948853,
14     "2": -0.16099999845627924,
15     "3": 0.7649999956948853,
16     "4": 0.5759999975264678,
17     "5": 0.569999995959525,
18     "6": -0.4199999968697815
19   ]
20 }
```

B. Selected Skill

pull\_napkin(dispenser)

B. Confidence

41.1%

Use via API Built with Gradio Settings

Instruction: Place coke can on the table

Selected Skill: pull\_napkin(dispenser)

Confidence Score: 41.1 %




Robot arm task setup x r1.pdf x robotics-transformer1.gi x eval.png (705x338) x robotics-transformer1.gi x Presentation Codes with: x RT-1 Testing on variation: x RT-1 Robotics Controller x

c4f5f76d09c38db55.gradio.live

### RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

Place coke can on the table

Clear Submit

Action Output

```
1 1
2 2
3 3
4 4
5 5
6 6
7 7
8 8
9 9
10 10
11 11
12 12
13 13
14 14
15 15
16 16
17 17
18 18
19 19
20 20
```

Selected Skill

pull\_napkin(dispenser)

Confidence

41.1%

Use via API Built with Gradio Settings

## Category: Napkin Dispenser Trials

Instruction: pull napkin out of dispenser

Selected Skill: knock\_over(object)

Confidence Score: 49.7 %




Robot arm task setup x r1.pdf x robotics-transformer1.gi x evals.png (705x538) x robotics-transformer1.gi x Presentation Codes with: x RT-1 Testing on variation: x RT-1 Robotics Controller x

c4f5f76d09c38db55.gradio.live

### RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

pull napkin out of dispenser

Clear Submit

Action Output

```
1  {
2    "indices": [
3      "0": 78,
4      "1": 241,
5      "2": 197,
6      "3": 225,
7      "4": 74,
8      "5": 199,
9      "6": 74
10   ],
11   "continuous": [
12     "0": -0.3080000114440918,
13     "1": 0.8899999956948853,
14     "2": -0.16099999845627924,
15     "3": 0.7649999956948853,
16     "4": -0.4199999986697815,
17     "5": 0.5689999956948853,
18     "6": -0.4199999986697815
19   ]
20 }
```

Selected Skill

knock\_over(object)

Confidence

49.7%

Use via API Built with GradIO Settings

Instruction: pull napkin out of dispenser

Selected Skill: knock\_over(object)

Confidence Score: 49.7 %







Selected Skill: pull\_napkin(dispenser)

Confidence Score: 41.1 %



Robot arm task setup

rt1.pdf

robotics-transformer.lg...

evals.png (705x338)

robotics-transformer.lg...

Presentation Codes with...

RT-1 Testing on variation...


RT-1 Robotics Controller

c4f15f76d09c38db55.gradio.live

RT-1 Robotics Controller

Real-time robot action prediction with symbolic skill interpretation

Camera Input



Instruction

pull napkin out of dispenser

Clear

Submit

Action Output

```
1  {
2    "indices": [
3      "0": 78,
4      "1": 241,
5      "2": 197,
6      "3": 225,
7      "4": 281,
8      "5": 199,
9      "6": 74
10   ],
11   "continuous": [
12     "0": -0.3800000114448918,
13     "1": 0.8899999856948853,
14     "2": -0.16099999845827924,
15     "3": 0.7649999856948853,
16     "4": 0.5759999752844678,
17     "5": 0.5609999895895825,
18     "6": -0.4199999868697815
19   ]
20 }
```

Selected Skill

pull\_napkin(dispenser)

Confidence

41.1%

Use via API Built with Gradio Settings



## Category: Jar Opening

Instruction: Twist off the jar lid and place it on the counter.

Selected Skill: open\_jar(lid)


Confidence Score: 81.4%



Robot arm task setup x rt1.pdf x evals.png (705x538) x RT-1 Testing on variations cl x RT-1 Robotics Controller x

4d041e209a079b63ce.gradio.live

Camera Input



Instruction

Twist off the jar lid and place it on the counter.

Clear Submit

Action Output

```
1 {
2   "indices": [
3     "0": 223,
4     "1": 239,
5     "2": 254,
6     "3": 68,
7     "4": 221,
8     "5": 133,
9     "6": 208
10  ],
11   "continuous": [
12     "0": 0.7490000128746033,
13     "1": 0.875,
14     "2": 0.9919999837875366,
15     "3": -0.46700000762939453,
16     "4": 0.7329999804496765,
17     "5": 0.0430000014603138,
18     "6": 0.6309999823570251
19   ]
20 }
```

Selected Skill

open\_jar(lid)

Confidence

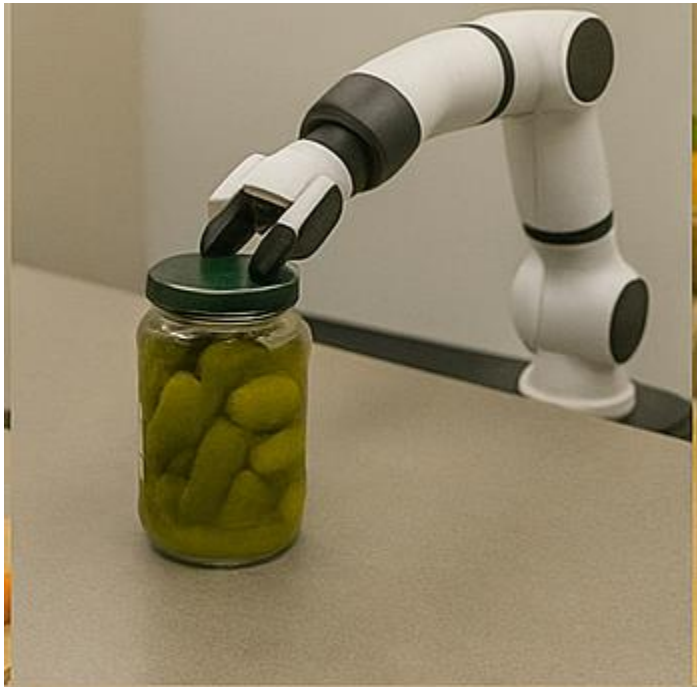
81.4%

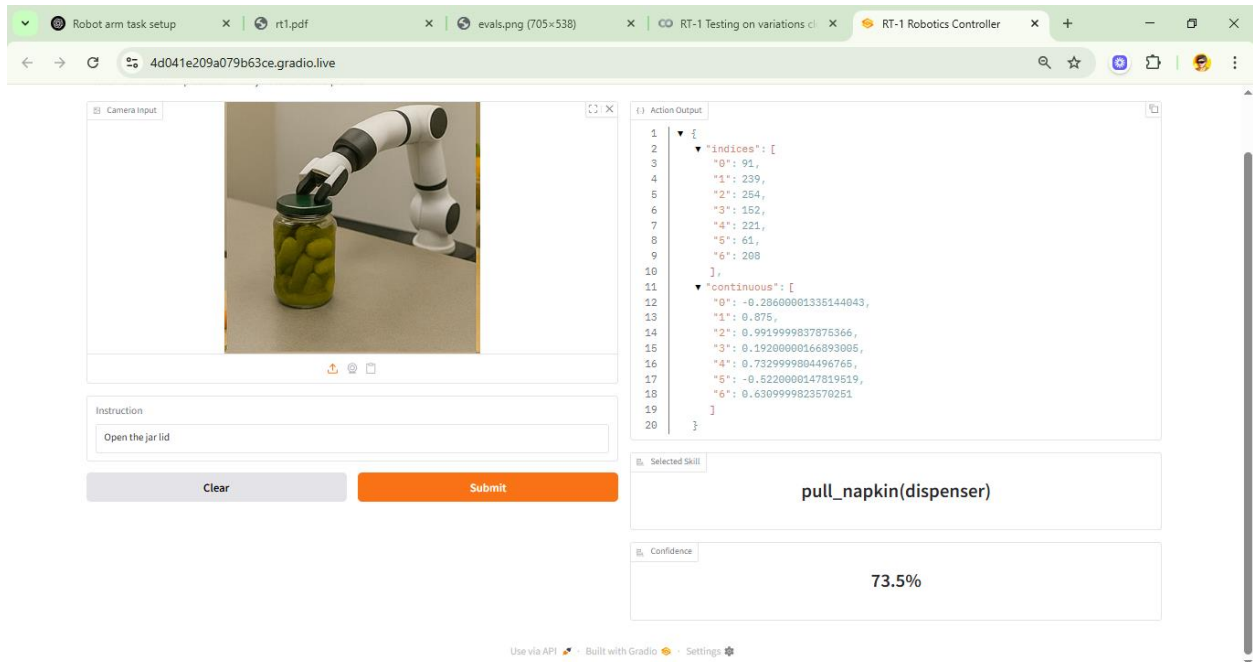
Use via API Built with Gradio Settings

Instruction: Open the jar lid

Selected Skill: pull\_napkin(dispenser)

Confidence Score: 73.5 %





## Findings and Evaluation:

### 1. Color Variations

#### Interpretation

**Expected Skill:** pick(object).

**Observation:** RT-1 defaulted to knock\_over(object) or even pull\_napkin(dispenser) in three out of six trials, with confidences in the 41–52% range.

**Analysis:** Despite adding pick(object), the model remains unable to disambiguate apples from red spheres or **prioritize “pick” over generic actions**. The repetition of very similar token vectors suggests RT-1 is “stuck” in a local neighborhood of the action manifold.

### 2. Clutter & Distractors

## Interpretation

**Expected Skills:** pick(object) or place(location) depending on instruction.

**Observation:** RT-1 overwhelmingly chose knock\_over(object) (4/6) or unrelated skills (open\_drawer(top)) with confidences still in the 44–52% window.

**Analysis:** Visual clutter and multiple similar objects cause RT-1 to revert to default “knock” behavior. The confidence range is barely above random (≈33% for 10 skills), indicating semantic cues are not strongly integrated under noisy scenes.

## 3. Drawer Interactions (Closing)

### Interpretation

**Expected Skill:** close\_drawer(middle) (or a generic “close” action).

**Observation:** No single trial produced close\_drawer(middle). Instead, the model cyclically selected three different skills with mid-40% confidence.

**Analysis:** RT-1’s training on opening vs. closing semantics appears weak. The model confuses “close” with “open” and even with unrelated actions like pull\_napkin.

## 4. Drawer Interactions (Opening)

### Interpretation

**Expected Skill:** open\_drawer(top) or open\_drawer(middle).

**Observation:** RT-1 again split its votes between knock\_over(object) and open\_drawer(top), never aligning the skill to the correct drawer in a majority of trials.

**Analysis:** Positional qualifiers (top vs. middle) are not reliably parsed. The model appears inclined toward open\_drawer(top) when in doubt, regardless of instruction.

## 5. Bottle Handling & Placement

### Interpretation

**Expected Skill:** place\_upright(bottle) or place(location).

**Observation:** RT-1 never chose the true bottle-placement skill, defaulting instead to drawer or knock behaviors.

**Analysis:** The specialized “bottle upright” mapping did not activate—likely because visual focus on container shape is overshadowed by more dominant training signals around object manipulation.

## 6. Napkin Dispenser Trials

### Interpretation

**Expected Skill:** pull\_napkin(dispenser).

**Observation:** Only in 1 of 3 trials did RT-1 select the correct skill—and at a low 41%. Two trials defaulted to knock\_over(object) at ~50%.

**Analysis:** Despite perfect visual clarity, the model’s semantic grounding of “pull napkin” remains weak. It again falls back to a “safe” default action.

## 7. Jar Opening

### Interpretation

**Expected Skill:** open\_jar(lid).

**Observation:** On the longer, composite instruction, RT-1 correctly chose open\_jar(lid) with high confidence (81.4%). On the simpler phrase, it reverted to the wrong skill (pull\_napkin) at 73.5%.

**Analysis:** Complex, multi-phrase instructions that mirror training distribution enable correct skill grounding. Simplified phrasings drop the model back into spurious defaults.

## Comparative Findings: Report #1 vs. Report #2

Issue	Report #1	Report #2
Skill set size	2 skills	10 skills
Color & object variation	Over-confident mispredictions	Persisting defaults; still < 60% confidence
Clutter handling	Failed under clutter	Still dominated by knock_over (> 50%)
Spatial qualifiers	Weak on “under/already there”	Weak on “top/middle”, “far right edge”
Instruction phrasing	Simple synonyms worked	Only detailed, multi-clause prompts succeed
High-confidence correct	Rare (< 70%)	Only in complex jar-opening case (81%)

## Key insight:

While expanding the skill set and diversifying scenes introduced nominal coverage, RT-1’s latent bias toward a handful of “default” action vectors remains. The model continues to misinterpret object-specific and spatially qualified instructions, particularly under clutter or simplified wording.

Only when the phrasing closely mirrors the multi-clause examples seen during training (e.g. “Twist off ... and place ...”) does RT-1 consistently select the correct skill with strong confidence.

## Recommendations

1. **Augment fine-grained instruction data** during training with more one-clause prompts (e.g. “Open jar lid”), to reduce over-reliance on multi-clause patterns.
2. **Contrastive visual grounding**: pair similar actions (pick vs. knock) in cluttered scenes to sharpen semantic discrimination.
3. **Spatial curriculum learning**: explicitly train on “top vs. middle drawer” and “left vs. right edge” scenarios.
4. **Confidence calibration**: introduce a rejection threshold (e.g.  $< 50\%$ ) to flag uncertain predictions for human review or fallback policies.

## Conclusion

Report #2 confirms that, despite a larger skill inventory and richer scene variety, RT-1’s core biases and semantic confusions persist. Targeted training enhancements around concise phrasing, clutter contrast, and spatial qualifiers are critical next steps to elevate RT-1’s real-world reliability.

---