LogT: Logically Think with Images for Visual Search

Y. Fu, Q. You, J. Guo, Y. Lu, S. Akula, H. Wu, M. Liu, L. Yuan, S. Dutta

Atlanta Poster
Fri 11 Dec 1:30 PM – 4:30 PM PST
Hall C1 #A718

arxiv.org/abs/2611.19200

Preliminary: Think with Images

I_0



Question:

What is the color of the tissue box?

(A) Gray

(B) White

(C) Black

(D) Blue ✓

Ground-Truth Region
(**<0.07%** of Input Image Area)



Preliminary: Think with Images

I_0



Question:

What is the color of the tissue box?

(A) Gray

(B) White

(C) Black

(D) Blue ✓

Ground-Truth Region
(**<0.07%** of Input Image Area)



… The question appears to be a trick question or a distractor, as **there is no tissue box present in the image** … The final answer is **none of the above**.

Qwen2.5-VL-**7B**

… There appears to be a tissue box on the dashboard … **The tissue box appears to be white**, matching the interior color of the truck's cab … The final answer is **B**.

Qwen2.5-VL-**72B**

!!

Pure text-based CoT is not enough.

Preliminary: Think with Images

I_0



Question:

What is the color of the tissue box?

(A) Gray

(B) White

(C) Black

(D) Blue ✓

Ground-Truth Region
(**<0.07%** of Input Image Area)



… The question appears to be a trick question or a distractor, as **there is no tissue box present in the image** … The final answer is **none of the above**.

Qwen2.5-VL-**7B**

… There appears to be a tissue box on the dashboard … **The tissue box appears to be white**, matching the interior color of the truck's cab … The final answer is **B**.

Qwen2.5-VL-**72B**

!!

Pure text-based CoT is not enough.

Turn 1: … I'll select a region that includes the ***front cabin area of the truck*** …

I_1



Turn 2: … I can now see the ***tissue box***. I need to zoom in on it to determine its color …

I_2



Turn 3: … The tissue box is **blue** in color … **Answer: Blue**

✓

Our 7B

Challenges

- **Existing work** [1, 2]**:**
  - SFT + RL with **complex tool-specific designs**.
- **Fundamental challenges remain despite RL:**
  - **Inaccurate & Inefficient** tool calls.
  - **Hallucinations**.

[1] *Pixel Reasoner: Incentivizing Pixel-Space Reasoning with Curiosity-Driven Reinforcement Learning.* NeurIPS'25.
[2] *Mini-o3: Scaling Up Reasoning Patterns and Interaction Turns for Visual Search.* ICLR'26.

Challenges: Case Study

I_0 

Q: What is the color of the tissue box? **A**: **Blue**

Ground-truth Region



QWEN2.5-VL-7B-RL

Turn #1

I'll use the “zoom_in” function to focus on the area where the tissue box …
Zoom-in on I_0: bbox=[1300, 720, 1550, 960]

I_1**: Tissue Box**



Turn #2

… **It is blue with some white text** and appears to be on top of the vehicle …
Answer: Blue

PIXEL REASONER-RL

Turn #1

… Now I will zoom in to look clearer at the truck’s front grille …
Zoom-in on I_0: bbox=[896, 604, 1344, 908]

I_1**: Side of the Truck**



Turn #2

… **There is no tissue box or any object resembling a tissue box** …
Answer: Blue

CHAIN-OF-FOCUS-RL

Turn #1

… I infer that the tissue box … possibly on the front grille area …
[Zoom-in on I_0] **bbox=[1308, 874, 1584, 960]**

I_1**: Logo**



Turn 2

… This reveals a tissue box with a **blue** color …
Answer: Blue

Challenges: Case Study

I_0 

Q: What is the color of the tissue box? **A**: **Blue**

Ground-truth Region



QWEN2.5-VL-7B-RL

Turn #1

I'll use the "zoom_in" function to focus on the area where the tissue box …
Zoom-in on I_0: bbox=[1300, 720, 1550, 960]

I_1: Tissue Box



Turn #2

… **It is blue with some white text** and appears to be on top of the vehicle …
Answer: **Blue**

✓

PIXEL REASONER-RL

Turn #1

… Now I will zoom in to look clearer at the truck's front grille …
Zoom-in on I_0: bbox=[896, 604, 1344, 908]

I_1: Side of the Truck



Turn #2

… **There is no tissue box or any object resembling a tissue box** …
Answer: **Blue**

✓

CHAIN-OF-FOCUS-RL

Turn #1

… I infer that the tissue box … possibly on the front grille area …
[Zoom-in on I_0] **bbox=[1308, 874, 1584, 960]**

I_1: Logo



Turn 2

… This reveals a tissue box with a **blue** color …
Answer: **Blue**

✓

DEEPEYES-RL

Turn #1

… **The tissue box appears to be white in color** … zoom in on the tissue box …
Zoom-in on I_0: bbox=[1300, 761, 1470, 820]

I_1: Papers



Turn #2

The image shows a close-up of a tissue box, **which appears to be white** …
Answer: **White**

✗

MINI-O3-RL

Turn #1

… I will zoom in on the truck to see if there is a tissue box …
Zoom-in on I_0: bbox=[672, 604, 1568, 1210]

I_1: Truck



Turn #2

… a small object that might be a tissue box. I will zoom in on that object …
Zoom-in on I_1: bbox=[448, 123, 896, 308]

I_2: Tissue Box



Turn #3

… **The tissue box is white**. The tissue itself is also white …
Answer: **White**

✗

Challenges: Case Study

I_0 

Q: What is the color of the tissue box? **A**: **Blue**

Ground-truth Region



QWEN2.5-VL-7B-RL

Turn #1

I'll use the "zoom_in" function to focus on the area where the tissue box …
Zoom-in on I_0: bbox=[1300, 720, 1550, 960]

I_1: Tissue Box



Turn #2

… **It is blue with some white text** and appears to be on top of the vehicle …
Answer: **Blue**

✓

PIXEL REASONER-RL

Turn #1

… Now I will zoom in to look clearer at the truck's front grille …
Zoom-in on I_0: bbox=[896, 604, 1344, 908]

I_1: Side of the Truck



Turn #2

… **There is no tissue box or any object resembling a tissue box** …
Answer: **Blue**

✓

CHAIN-OF-FOCUS-RL

Turn #1

… I infer that the tissue box … possibly on the front grille area …
[Zoom-in on I_0] **bbox=[1308, 874, 1584, 960]**

I_1: Logo



Turn 2

… This reveals a tissue box with a **blue** color …
Answer: **Blue**

✓

DEEPEYES-RL

Turn #1

… **The tissue box appears to be white in color** … zoom in on the tissue box …
Zoom-in on I_0: bbox=[1300, 761, 1470, 820]

I_1: Papers



Turn #2

The image shows a close-up of a tissue box, **which appears to be white** …
Answer: **White**

✗

MINI-O3-RL

Turn #1

… I will zoom in on the truck to see if there is a tissue box …
Zoom-in on I_0: bbox=[672, 604, 1568, 1210]

I_1: Truck



Turn #2

… a small object that might be a tissue box. I will zoom in on that object …
Zoom-in on I_1: bbox=[448, 123, 896, 308]

I_2: Tissue Box



Turn #3

… **The tissue box is white**. The tissue itself is also white …
Answer: **White**

✗

LOGT-RL (OURS)

Turn #1

… I'll select a region that includes the **front cabin area of the truck** …
Zoom-in on I_0: bbox=[1155, 533, 1689, 849]

I_1: Front Cabin



Turn #2

… I can now see the **tissue box** …I need to zoom in on it to determine its color …
Zoom-in on I_0: bbox=[1399, 735, 1499, 797]

I_2: Tissue Box



Turn #3

… After zooming in, I can clearly see that the tissue box is blue in color …
Answer: **Blue**

✓

Main Idea 1: Object-Oriented QA Pair Synthesis

Previous: **Direct** Use of Input Image (**Image-oriented**).



Ground-truth Region (**~5.2%** of Input Image Area)

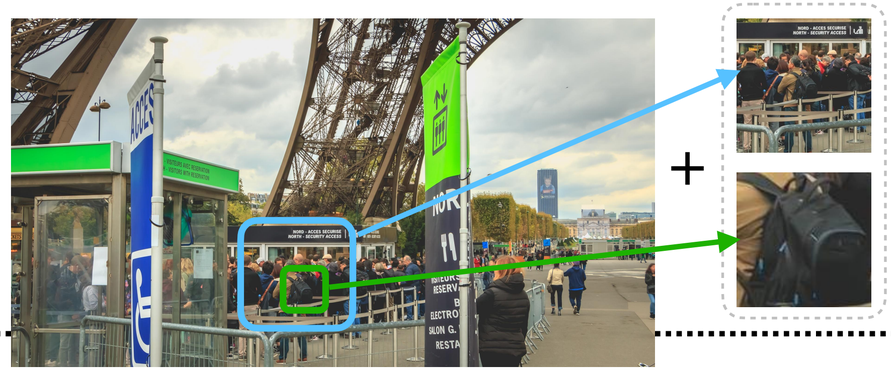


Question:
What is the color of the wheelchair accessibility symbol?
Answer:
Blue

Overly Easy Question

Main Idea 1: Object-Oriented QA Pair Synthesis

Contexts



Previous: **Direct** Use of Input Image (**Image-oriented**).

Ours: Add More **Visual Context** About **Small** Target Objects (**Object-oriented**).

Ground-truth Region (**~5.2%** of Input Image Area)



Question:
What is the color of the wheelchair accessibility symbol?
Answer:
Blue

Overly Easy Question

Question:
What color is the backpack worn by the person wearing a light brown jacket near the security access gate?
Answer:
Black

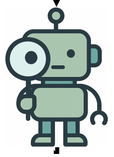
More Challenging Question

Ground-truth Region (**~0.4%** of Input Image Area)



Main Idea 2: Logic Chain-Driven Reasoning Trace Synthesis

Previous: Direct synthesis **without fine-grained control** over tool-use logic.



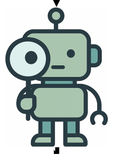
Low-quality Tool-use Reasoning Trace

Question: What color is the backpack worn by the person wearing a light brown jacket near the security access gate?

Main Idea 2: Logic Chain-Driven Reasoning Trace Synthesis

- Three **human strategies** for visual QA:
  1. **Direct answering (no tools)** for **easy QAs**.
  2. **Multi-hop reasoning** for **challenging QAs**.
  3. **Exhaustive search** for the **hardest QAs**.

Previous: Direct synthesis **without fine-grained control** over tool-use logic.



Low-quality Tool-use Reasoning Trace

Question: What color is the backpack worn by the person wearing a light brown jacket near the security access gate?

Ours: First synthesize the **logic chain**, then use it to **guide tool-use trace generation**.

Avg@8≠1

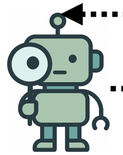
Multi-hop Reasoning / Exhaustive Search

Logic Chain

Step #1: Gate (I_0) → **Step #2:** Person (I_0) → **Step #3:** Backpack (I_2)

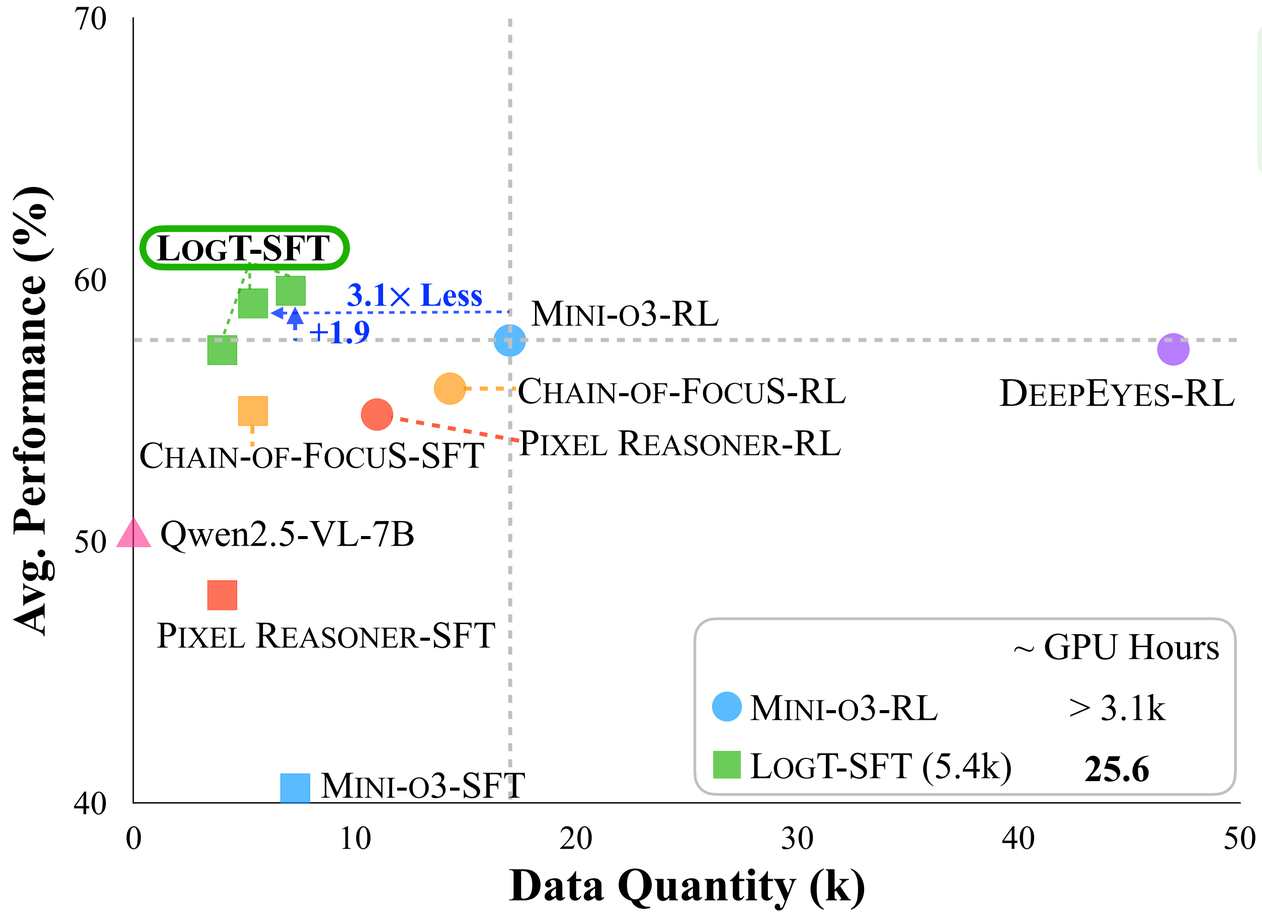
average@8=1

Direct Answering (No Tool)



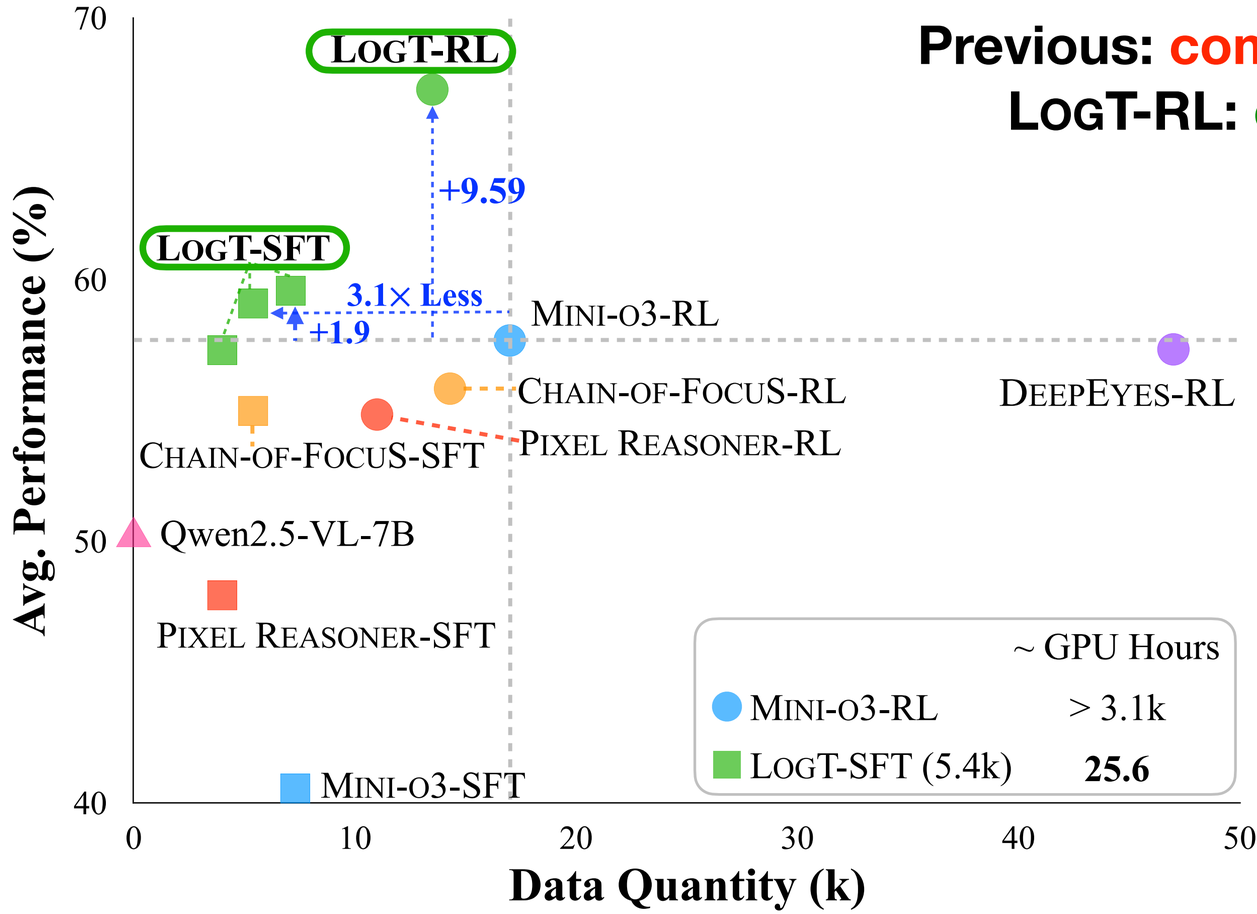
High-quality & Structured Tool-use Reasoning Trace

Experiment: LogT-SFT +1.9 vs. Open-Source SOTA with SFT + RL



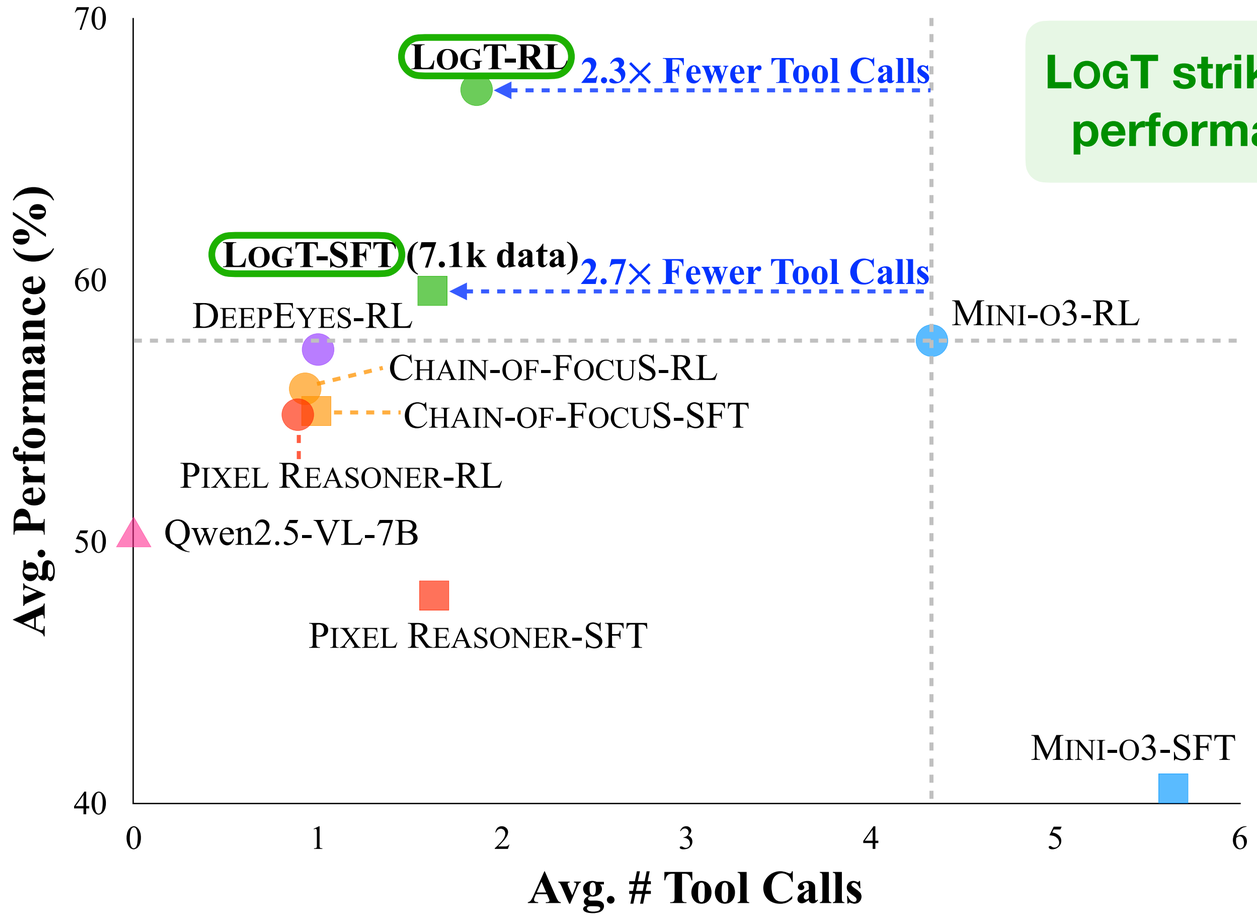
- **3.1× less data!**
- **>120× fewer GPU hours!**

Experiment: Simple RL +9.59 vs. Open-Source SOTA with SFT + RL



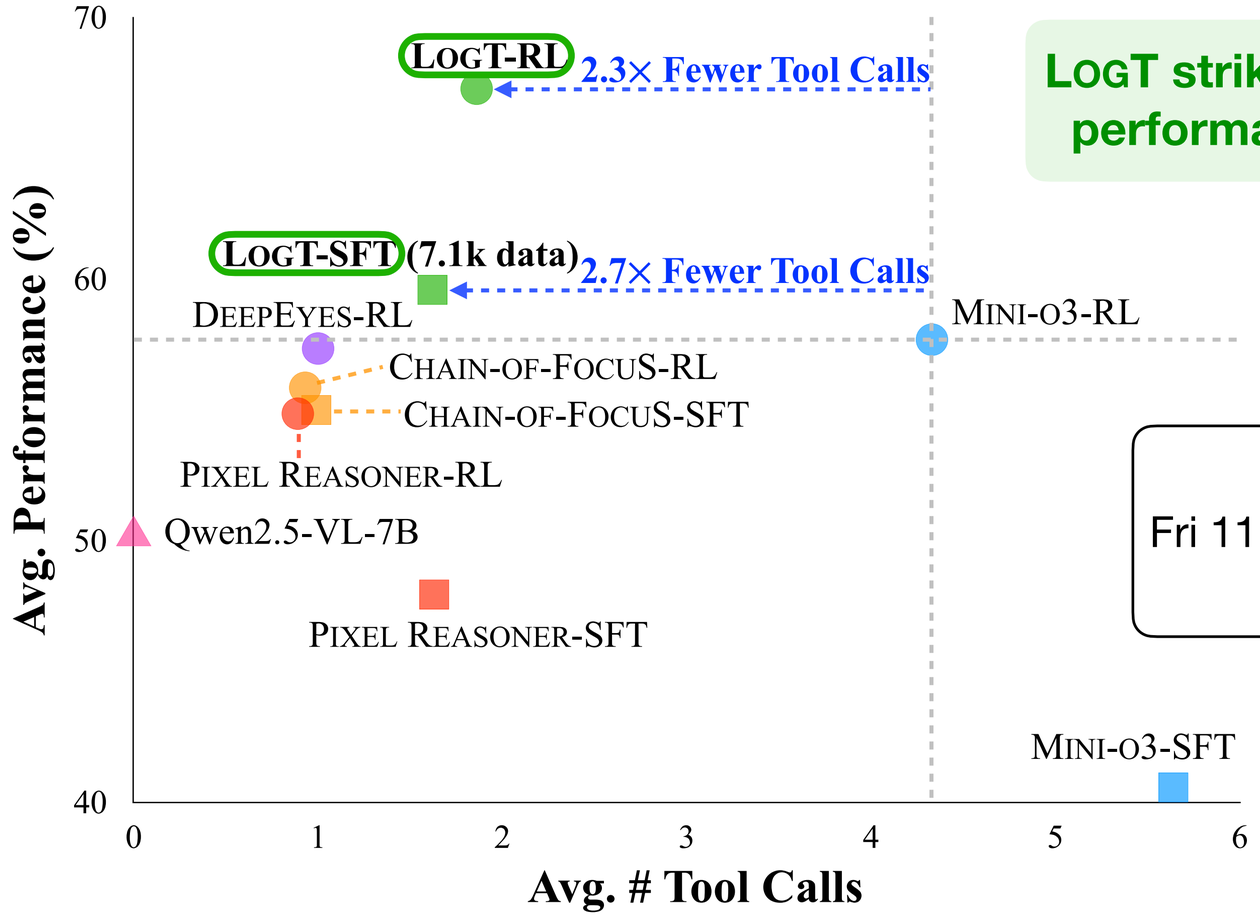
Previous: complex, tool-specific RL;
LogT-RL: off-the shelf GRPO.

Analysis: High Tool Call Efficiency



LogT strikes a strong balance between performance and tool-call efficiency.

Analysis: High Tool Call Efficiency



LogT strikes a strong balance between performance and tool-call efficiency.

Atlanta Poster
Fri 11 Dec 1:30 PM – 4:30 PM PST
Hall C1 #A718