
Supplementary Material of “Latent Chain-of-Thought for Visual Reasoning”

Anonymous Author(s)

Affiliation

Address

email

1 Proofs

Proposition 1. Let $R(z_{1:t} \top) = \log P(X z_{1:t} Y)$ be a joint-likelihood reward function.

(a) If $R(z_{1:-})$ and $R(z_{1:-+\lambda})$ are true reward and the intermediate rewards within region of length λ are constantly increment, then we approximate the intermediate reward at step $t+i$ (where $0 \leq i \leq \lambda$) as:

$$\tilde{R}(z_{1:t+i} \top) = R(z_{1:t} \top) + \frac{i}{\lambda} (R(z_{1:t+\lambda} \top) - R(z_{1:t} \top)). \quad (1)$$

(b) If λ is short enough, the interpolation reward error stays close to 0 and the flow between $F(z_{1:-})$ and $F(z_{1:-+\lambda})$ satisfies

$$F(z_i) \prod_{k=i+1}^j P_F(z_k | z_{k-1}) = F(z_j) \prod_{k=i+1}^j P_B(z_{k-1} | z_k), \quad (2)$$

Proof. (a) Assume that within the segment $\{t, t+1, \dots, t+\lambda\}$ the true reward grows linearly, i.e.

$$R(z_{1:t+i} \top) = R(z_{1:t} \top) + i \Delta, \quad \Delta := \frac{R(z_{1:t+\lambda} \top) - R(z_{1:t} \top)}{\lambda}, \quad 0 \leq i \leq \lambda.$$

Substituting this expression into Eq. (1) shows $\tilde{R}(z_{1:t+i} \top) = R(z_{1:t+i} \top)$ for every i , so the interpolation incurs zero error.

(b) Suppose R is twice-differentiable along the trajectory and its discrete second derivative is bounded:

$$|R(z_{1:s+1} \top) - 2R(z_{1:s} \top) + R(z_{1:s-1} \top)| \leq M, \quad \forall s.$$

The classical linear-interpolation error bound then yields

$$|\tilde{R}(z_{1:t+i} \top) - R(z_{1:t+i} \top)| \leq \frac{M}{8} i(\lambda - i) \leq \frac{M\lambda^2}{8}, \quad 0 \leq i \leq \lambda. \quad (3)$$

Thus the approximation error decays as $\mathcal{O}(\lambda^2)$; choosing λ sufficiently small keeps it arbitrarily close to 0.

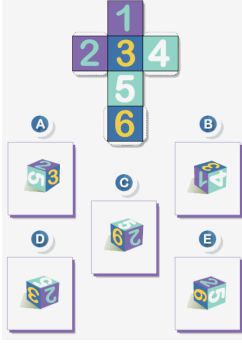
Let

$$F(z_s) := \frac{R(z_{1:s} \top)}{q_{\theta}(\top | z_{1:s})}, \quad \tilde{F}(z_s) := \frac{\tilde{R}(z_{1:s} \top)}{q_{\theta}(\top | z_{1:s})}, \quad \varepsilon_s := \tilde{R}(z_{1:s} \top) - R(z_{1:s} \top).$$

From Eq. (3) we have $|\varepsilon_s| \leq \frac{M\lambda^2}{8}$, so

$$\tilde{F}(z_s) = F(z_s) \left(1 + \frac{\varepsilon_s}{R(z_{1:s} \top)} \right).$$

Which cube is identical to the unfolded net?



Reasoning: Okay, so I'm trying to figure out which cube is identical to the unfolded net based on the given image. Let me break this down step by step. The top row has ... Looking at the cubes mentioned: Cube A: Blue, green, yellow sides. Cube E: Purple, yellow, green sides. The image shows that the top row (boxes 1-4) has purple and cyan boxes, ... Cube A represents the top-left box (1),... Cube D would represent the middle-bottom box (4), and Cube E would represent the bottom-right box (5).... **Wrong reasoning.**
Final output: E **Wrong answer.** SFT

Reasoning: Okay, so I'm trying to figure out which cube is identical to the unfolded net based on the image itself. Let me break it down step by step. ... The hint says that the correct answer is C, but let's verify this directly from the image. Looking at Cube C, we can see that it has the numbers "2" and "5" on its visible faces. This matches the arrangement of the numbers in the grid where "2" and "5" ... Let's check the other cubes: Cube A has "2" and "3". Cube B has "4" and "1". ... Cube C must be the correct choice. **Wrong reasoning.**
Final output: C **Wrong answer.** GRPO

Reasoning: Okay, so I'm trying to figure out which cube is identical to the unfolded net based on the image itself. Let me start by understanding the problem.

First, the image shows an unfolded net of a die, which means it's a ...

Looking at the image:

Cube A has faces showing 2, 5, and 3. Cube B has faces showing 4, 1, and 3. Cube C has faces showing 2, 5, and 9. Cube D has faces showing 2, 3, and 5. Cube E has faces showing 2, 5, and 9.

The net in the image has the following arrangement:

Face 1 is in the top center. Face 2 is in the top left corner. Face 3 is in the top right corner. Face 4 is in the bottom left corner. Face 5 is in the bottom center. Face 6 is in the bottom right corner. ...

Now, let's check each cube against the net:

Cube A: Face 2, 5, and 3. ... Cube D: Face 2, 3, and 5. If face 2 is in the top left, then face 3 would be in the top right, and face 5 would be in the top center. This fits the net's arrangement where face 5 is at the top center...

Final output: D **Correct.** LaCoT

Figure F1: Qualitative results of visual reasoning. We **highlight** the important reasoning steps.

- 18 Denote $R_{\min} := \min_{s \leq t + \lambda} R(z_{1:s}^\top) > 0$ (positivity follows from likelihoods). Then
 19 $|\varepsilon_s / R(z_{1:s}^\top)| \leq \frac{M\lambda^2}{8R_{\min}}$.
 20 Consider any $i < j \leq t + \lambda$. Applying Eq. (2) to both F and \tilde{F} , we obtain

$$\tilde{F}(z_i) \prod_{k=i+1}^j P_F(z_k | z_{k-1}) = \tilde{F}(z_j) \prod_{k=i+1}^j P_B(z_{k-1} | z_k) (1 + \mathcal{O}(\lambda^2)),$$

- 21 where the $\mathcal{O}(\lambda^2)$ term accumulates at most $(j - i)$ relative perturbations bounded by $M\lambda^2 / (8R_{\min})$.
 22 Hence, the forward and backward product flows still match up to a multiplicative factor that vanishes
 23 quadratically as $\lambda \rightarrow 0$. Therefore, the interpolated rewards preserve flow consistency to arbitrary
 24 precision for sufficiently small segment lengths. \square

2 Experiments

2.1 Qualitative results

- 27 In Fig. F1, we provide qualitative results of a comparison between Qwen2.5-VL-7B (SFT), Qwen2.5-
 28 VL-7B (GRPO), and LaCoT-Qwen-7B. As can be seen, LaCoT-Qwen-7B can provide a more accurate
 29 reasoning chain, leading to the correct answer. Meanwhile, due to limited generalizability, SFT and
 30 GRPO samples show the wrong visual CoT. In Fig. F2 and Fig. F3, our LaCoT model can sample
 31 more straightforward and accurate reasoning chains, demonstrating the effectiveness and robustness
 32 of the proposed training and inference algorithm.

2.2 Experiments compute resources

- 34 This work utilizes an 8*80GB GPU-node for training. We set the Deepspeed Zero-3 stage and
 35 gradient-checkpointing to reduce memory costs during optimization. It takes around 30 hours
 36 for supervised fine-tuning on 250k reasoning data samples, and 120 hours for GRPO and RGFN
 37 fine-tuning on 3k data samples.



Question: Subtract all green rubber cubes. Subtract all red matte blocks. How many objects are left?

Reasoning: Okay, so I need to figure out how many objects are left after subtracting all green rubber cubes and red matte blocks from this image. Let me start by understanding what each term means.

First, "green rubber cubes." The image shows three cubes: one blue with a reflective surface, one cyan with a glossy finish, and one yellow with a shiny appearance. **There is no green cube in the image. Therefore, no green rubber cubes exist to remove.**

Next, "red matte blocks." The image shows two small red objects, both of which appear to be cubes based on their shape. They do not look matte; they have a glossy finish. Therefore, there are **no red matte blocks to subtract either.** ❌ *Wrong reasoning.*

Given that neither green rubber cubes nor red matte blocks are present in the image, we can conclude that **no objects need to be removed from the total count.**

Looking at the initial setup described: there's a large blue sphere with a reflective surface, a smaller red cylinder, a medium-sized red matte cube, a blue-tinted metallic cube, a small yellow cylinder, another small yellow cylinder, and a small purple sphere. Let's count them:

1. Blue sphere (reflective)
2. Red cylinder (cylinder)
3. Red matte cube (cube)
4. Blue-tinted metallic cube (cube)
5. Yellow cylinder (cylinder)
6. Another yellow cylinder (cylinder)
7. Purple sphere (sphere)

That makes seven objects in total. Since neither green rubber cubes nor red matte blocks were present in the image, all seven objects remain.

Therefore, the answer should be 7.

Final output: 7. ❌ *Wrong answer.* SFT

Reasoning: The image contains several objects: a blue sphere, two red cylinders, a red cube, a teal cube, a yellow cylinder, a purple sphere, and another yellow cylinder.

First, I need to identify and remove any green rubber cubes. As seen in the image, there are no green rubber cubes present.

Next, I will look for any red matte blocks. The image shows a red cube, which appears to be matte. I will subtract this red cube from the total count. ✓

Now, let's count the remaining objects:

- Blue sphere
- Two red cylinders
- Red cube (removed)
- Teal cube
- Yellow cylinder
- Purple sphere
- Another yellow cylinder

After removing the red cube, the remaining objects are:

1. Blue sphere
2. Two red cylinders
3. Teal cube
4. Yellow cylinder
5. Purple sphere
6. Another yellow cylinder

This totals to six objects.

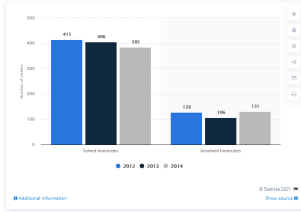
Therefore, after removing the red block, six objects remain ✓

Final output: 6 ✓ LaCoT

Figure F2: Qualitative results of visual reasoning. We **highlight** the important reasoning steps.

2.3 Hyperparameter

We detail the hyperparameters used for training the reward model and LaCoT in our experiments in Table T1. During LaCoT training, we randomly sample (mini-batch size) Z s for every (X, Y) as exploration.



Question: What is the average of longest light blue bar and shortest gray bar?

Reasoning: Okay, so I need to figure out the average of the longest light blue bar and the shortest gray bar from the given image. Let me start by understanding what each part means.

First, the image shows a bar chart comparing solved and unsolved homicides over three years: 2012, 2013, and 2014. The bars are color-coded—light blue for 2012, dark blue for 2013, and gray for 2014.

Looking at the data provided in the Python code, there's a list called `homicide_data` which includes two arrays: `solved` and `unsolved`. The `solved` array has values [415, 406, 385] corresponding to the years 2012, 2013, and 2014 respectively. As seen in the image, the light blue bars represent 2012, the dark blue bars represent 2013, and the gray bars represent 2014. So, the light blue bar is for 2012 with a value of 415.

Next, the `unsolved` array has values [128, 106, 131]. The first value (128) is for 2012, the second (106) is for 2013, and the third (131) is for 2014. However, since the question mentions solving or unsolving homicides, it seems we're focusing on the solved ones here because the user referred to "longest light blue bar." But wait, **the light blue bar is actually the solved homicides for 2012, which is 415.** ✓

Now, the gray bar is for 2014, which is the shortest among all the gray bars. As seen in the image, the gray bars represent 2014, and the shortest one is 385, but the note says it's 385? Wait, no—the note mentions that 385 is for 2014, but looking back, the `unsolved` array has 385 as its last value, which corresponds to 2014. So perhaps the gray bar is indeed the shortest unsolved homicides, which would be 385. But let me check again. Wait, maybe I'm mixing up something. Let me clarify. The solved homicides are shown in light blue, and their values are [415, 406, 385]. The unsolved homicides are shown in gray, with values [128, 106, 131]. So for 2014, both solved and unsolved have numbers, but the gray bar specifically refers to unsolved homicides, which is lower than the solved ones. So the shortest gray bar is 106 (for 2013).

But hold on, the user mentioned the longest light blue bar and the shortest gray bar. So the light blue bar is 415, which is the longest. The gray bar varies between 131, 385, and 106. The smallest number here is 106, which is in 2013. So **the shortest gray bar is 106.** ✗ *Wrong reasoning.*

So now, averaging these two: 415 and 106. To find the average, I add them together and divide by 2. That gives $(415 + 106)/2 = 521/2 = 260.5$. Since the user asked for an integer, I should round this. 260.5 rounds to 261. So the average is 261.

Final output: 261. ✗ *Wrong answer.*

SFT

Reasoning: Okay, so I need to figure out the average of the longest light blue bar and the shortest gray bar in the given image.

First, let's look at the image itself. The image shows a bar chart with two main categories: "Solved homicides" and "Unsolved homicides." Each category has three bars representing the years 2012, 2013, and 2014. The colors used are light blue for 2012, dark blue for 2013, and gray for 2014.

Now, focusing on the light blue bars, which represent the year 2012. The values for solved homicides are 415, and for unsolved homicides, it's 128. So, the light blue bars show 415 for solved and 128 for unsolved.

Next, looking at the gray bars, which represent the year 2014. The values here are 385 for solved homicides and 131 for unsolved. So, the gray bars show 385 and 131.

The question asks for the average of the longest light blue bar and the shortest gray bar. **The longest light blue bar would be the one with the highest value, which is 415 for solved homicides. The shortest gray bar would be the one with the lowest value, which is 131 for unsolved homicides.** ✓

So, we have the numbers 415 (longest light blue) and 131 (shortest gray). To find the average, we add these two together and divide by 2.

Calculating that: $(415 + 131 = 546)$, then $(546 / 2 = 273)$. ✓

Therefore, the average should be 273.

Final output: 273 ✓

LaCoT

Figure F3: Qualitative results of visual reasoning. We **highlight** the important reasoning steps.

Table T1: Hyperparameters for training.

LoRA dropout	0.05
Batch size (SFT)	2
Batch size (RGFN)	1
Gradient accumulation (SFT)	16
Learning rate	0.00001
Optimizer	AdamW
Weight decay	0.05
Temperature max	1.0
Temperature min	0.5
Reward temperature start	1.0
Reward temperature end	0.7
Reward temperature horizon	50
mini-batch size	6
τ_{max}	1.5
τ_{min}	1.0
Maximum rationale length	700
Minimum rationale length	64