NEURIPS 2026 · ORAL

Clarification as Supervision: Reinforcement Learning for Vision-Language Interfaces

John Gkountouras[1] Ivan Titov[1,2]

[1]University of Amsterdam [2]University of Edinburgh

johngou.com/acrl
Paper, code and models

Captions for a reasoner

A small vision-language model, the captioner π_θ, describes the image. A frozen text-only reasoner $\mathcal{R}$ reads the description and answers the question. The reasoner can be an API model.

Captioners are trained on descriptions written for people. These often leave out the exact values the reasoner needs, so the reasoner fails on problems it could solve.

No dataset says which details a given reasoner needs. AC-RL learns them from interaction with the reasoner.

Example

IMAGE

oval beads	$2/kilogram
rectangular beads	$3/kilogram
star-shaped beads	$2/kilogram
spherical beads	$2/kilogram
heart-shaped beads	$3/kilogram
square beads	$2/kilogram
flower-shaped beads	$2/kilogram

QUESTION

Chase wants to buy 4 kilograms of oval beads and 5 kilograms of star-shaped beads. How much will he spend? (Unit: $)

Pretrained captioner

The image appears to be a table listing various types of beads along with their respective prices per kilogram. … I can see oval beads, rectangular beads, star-shaped beads, spherical beads, heart-shaped beads, square beads, and flower-shaped beads. Each type has a corresponding price per kilogram shown in the second column. …

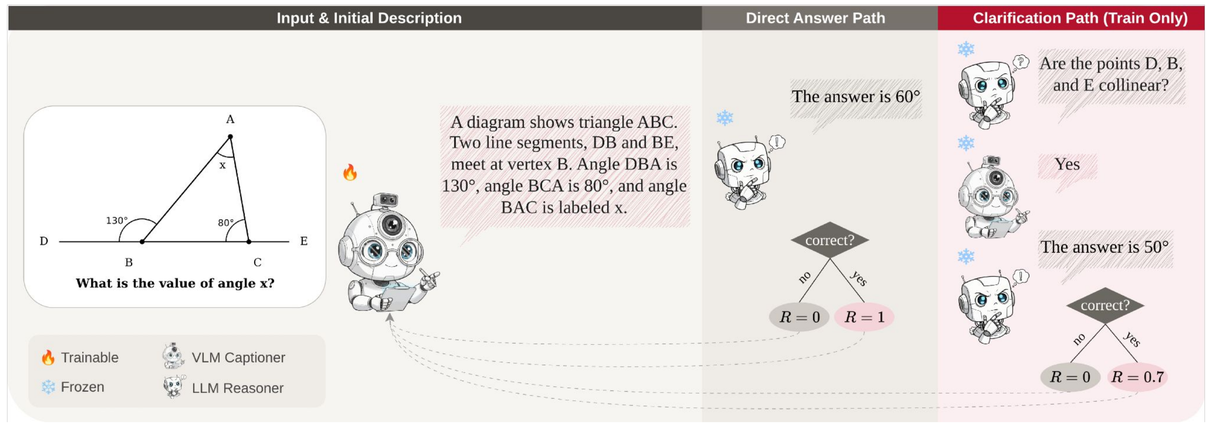
AC-RL captioner

The image appears to be a table listing various types of beads along with their respective prices per kilogram. The table is as follows:

Beads	Price/kg
oval beads	$2
rectangular beads	$3
star-shaped beads	$2
spherical beads	$2
heart-shaped beads	$3
square beads	$2
flower-shaped beads	$2

Only the AC-RL caption gives the prices of oval and star-shaped beads.

Training with clarification



$$c_0 \sim \pi_\theta(\cdot \mid I, Q)$$
$$q_1 \sim \mathcal{R}(\cdot \mid Q, c_0)$$
$$c_1 \sim \pi_{\text{ref}}(\cdot \mid I, Q, q_1)$$
$$A \sim \mathcal{R}(\cdot \mid Q, c_0, q_1, c_1)$$

$$R(\tau) = \begin{cases} 1 & \text{correct without clarification} \\ \alpha & \text{correct after clarification} \\ 0 & \text{incorrect} \end{cases}$$

$$J(\theta) = \mathbb{E}_{(I,Q)\sim\mathcal{D}}\, \mathbb{E}_{\tau\sim\pi_\theta}\big[R(\tau)\big] - \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)$$

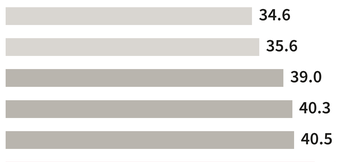
At test time the reasoner gets one caption and no clarification.

During training the reasoner answers from the first caption c_0 or asks one question q_1, and a frozen captioner π_{ref} replies with c_1.

- Gradients flow only through c_0. The clarifier is frozen, so the captioner cannot hand missing details over to it.
- Partial credit gives a learning signal for captions that are almost sufficient. The penalty $1 - \alpha$ pushes the missing details into c_0.

Where the gain comes from

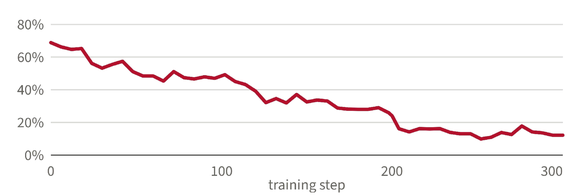
Average accuracy (%) over the seven benchmarks



The similarity reward adds image-text similarity to the binary reward, a dense signal that does not use clarification.

Clarification during training

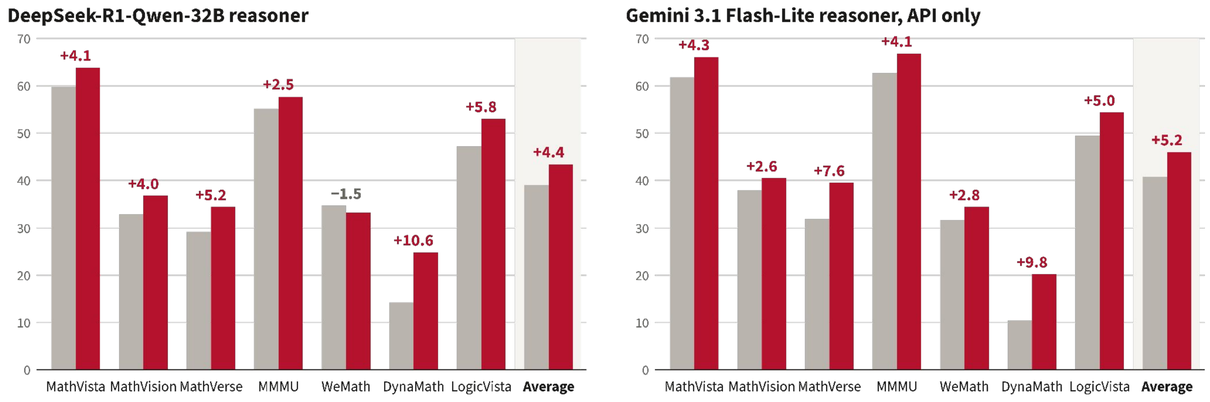
Share of rollouts in which the reasoner asks a question



With clarification allowed at test time on MathVerse, the reasoner asks on 35.0% of problems with AC-RL captions and 62.9% with binary-reward RL captions.

Results on seven benchmarks

Accuracy (%), Qwen2.5-VL-3B captioner, single-pass evaluation — pretrained captioner — AC-RL captioner



Ablating the clarifier

Accuracy (%)	LogicVista	MathVision
Pretrained, no RL	47.2	34.7
AC-RL	**55.7**	**40.2**
Weaker 2B clarifier	52.8	38.1
Rewards shuffled across rollouts	50.1	38.2
Clarifier gives no information	46.8	37.2

Shuffling keeps the reward values and breaks their link to clarification.

Transfer

Accuracy gain of the AC-RL captioner over the pretrained one (points)

Unseen reasoners		
8B, LogicVista	+8.7	44.3 → 53.0
8B, MathVision	+6.5	30.0 → 36.5
GPT-5 mini, LogicVista	+2.9	48.1 → 51.0
GPT-5 mini, MathVista	+3.9	62.2 → 66.1
Medical and science benchmarks, training reasoner		
WorldMedQA-V	+8.1	50.2 → 58.3
ScienceQA	+1.2	85.0 → 86.2