DeepSeek-R1 & PAPO

DeepSeek-AI, Nature 2025.
Wang et al, ICLR2026.

Presenters: Eric He, Frank Yang, Shimin Zhang

Georgia Tech

How to train language model to reason?

Georgia Tech

Outline

1. Problem & motivation
2. DeepSeek-R1: RLVR & GRPO
3. Multimodal RLVR gap
4. PAPO objective & stability
5. Experiments & results
6. What persisted / changed
7. Limitations, takeaways & discussion

Georgia Tech

Limitations of SFT and human-annotated CoT

High cost

OpenAI's PRM800K required 800,000 human labels on individual reasoning steps.

Cognitive biases
Human-defined reasoning patterns may limit model exploration

Data quality is capped by human's reasoning capability

Can LLMs learns to reason without SFT in post-training?

Georgia Tech

DeepSeek R1-Zero: post-training purely with RLVR

1 Generate → Sample four responses

2 Verify → Score each outcome

3 Update → Adjust the policy

4 Repeat → Sample again

LLM as an RL policy

RL role	In language generation
Policy / actor	Next-token distribution
State	Prompt + generated prefix
Action	Next token
Reward	Verifier score after completion

Verified reward example:

Attempt	Test outcome	Reward
A	Passes tests	1
B	Passes tests	1
C	Fails an edge case	0
D	Exceeds time limit	0

With the reward, how to update parameters?

Guo et al., Nature (2025), Reward design. Examples are illustrative.

PPO uses a learned critic to estimate advantage

State: the prompt and generated prefix. Action: the next token.

Actor π_θ

Generates the next token from the current prefix
Updates the policy using advantage

Critic $V_\psi(s_t)$

Predicts expected return from the current prefix
Learns by fitting observed returns

$$\hat{A}_t \approx G_t - V_\psi(s_t)$$

Observed return minus the predicted baseline (intuition)
Positive: reinforce the action. Negative: suppress it.

Why subtract a baseline? Judge performance relative to expectation and reduce gradient variance

Example: expected return at the prompt is 0.5 $\hat{A}_{\text{pass}} = 1 - 0.5 = +0.5$ $\hat{A}_{\text{fail}} = 0 - 0.5 = -0.5$

Critic costs a lot Typically a model comparable in size to the policy

Extra memory for model and training states, plus value-model training computation

Can we estimate advantage without training a critic?

Schulman et al., PPO (2017); Shao et al., DeepSeekMath (2024), §4.1.

GRPO removes the critic

Use mean reward in a group as the baseline.

$\mu_r = 0.5 \qquad \sigma_r = 0.5$

Response	r_i	$r_i - \mu_r$	A_i
A	1	+0.5	**+1**
B	1	+0.5	**+1**
C	0	−0.5	**−1**
D	0	−0.5	**−1**

Equal rewards provide no group contrast. Zero std needs handling.

$$A_i = \frac{r_i - \mu_r}{\sigma_r}$$

Subtract the mean

Compare with alternatives for the same prompt

Divide by the std

Normalize the signal scale

Group scores supply the baseline without a learned critic.

Shao et al., DeepSeekMath (2024), §4.1. Examples are illustrative.

GRPO replaces the critic, retains PPO-style optimization

Sampled-group objective to maximize

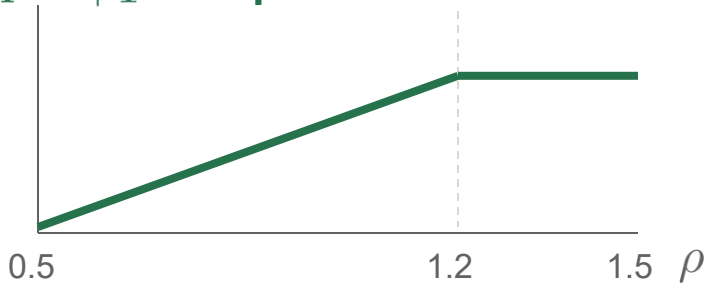
$$\mathcal{J} = \frac{1}{G}\sum_{i=1}^{G}\left[\min\left(\rho_i A_i, \mathrm{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i\right) - \beta D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}})\right]$$

$$\rho_i = \frac{\pi_\theta(o_i \mid q)}{\pi_{\mathrm{old}}(o_i \mid q)}$$

Group advantage from the previous slide

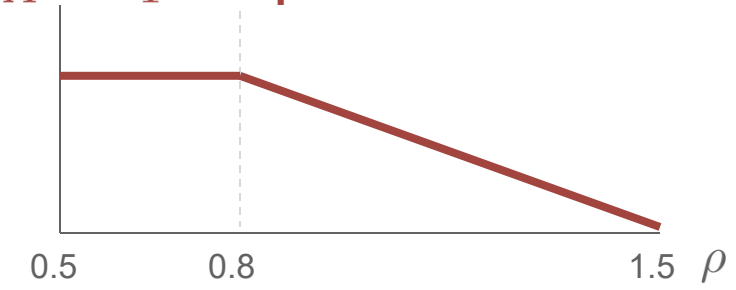
$\epsilon = 0.2$

$A = +1$ **Cap the incentive to increase**



Flat above 1.2

$A = -1$ **Cap the incentive to decrease**



Flat below 0.8

π_{old} Rollout snapshot used in the ratio

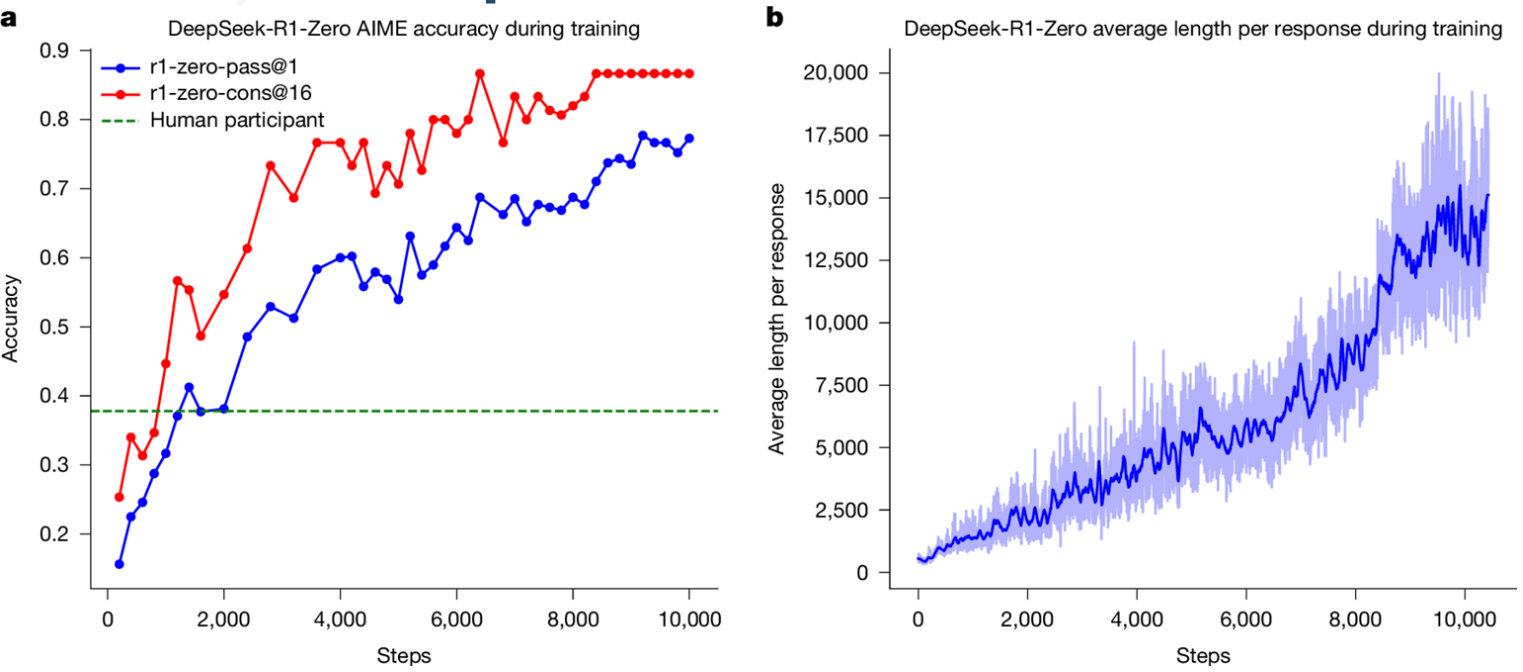
π_{ref} Reference used in the KL penalty

Clipping discourages overly large updates. KL penalty discourages drift from the reference policy. Both for stability.

Schulman et al., PPO (2017); Shao et al., DeepSeekMath (2024), §4.1.

Result of DeepSeek-R1-zero: pure RLVR post-training

AIME 2024 pass@1: 15.6% to 77.9%



Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$

Aha moment emerge

Reported behaviors: self-verification, reflection, alternative solutions

Takeaway: RLVR post-training can improve reasoning.

Guo et al., Nature (2025), Fig. 1 and Table 1. Figure reproduced, resized; CC BY 4.0.

Georgia Tech

R1-Zero's limitations motivate R1

Reported limitations

R1-Zero limitation	Example
Poor readability	"Let x be the answer… actually, use y. Then x = 2y… wait, go back to the first case…"
Language mixing	We need to find the minimum. **首先**对这个函数求导，**然后** set the derivative to zero.
Limited writing performance	Asked for a polite extension request: "I need more time. Change the deadline. Thanks."

R1: retain strong reasoning while improving readability and general assistant performance

Guo et al., Nature (2025), Main and DeepSeek-R1 sections.

Georgia Tech

RLVR can verify correctness, but is correctness enough to train a general-purpose model?

Task	What the verifier checks	What correctness alone misses
Coding / proving	Passes tests / satisfies a formal specification	Maintainability / clarity of exposition
Explaining a concept	Statements are correct	Whether this user understands
Giving advice	Claims are factually accurate	Whether the advice fits the user's goals and constraints

RLVR is not enough for a general-purpose model.

Guo et al., Nature (2025), Methods: Reward design, Helpful reward model, Safety reward model.

Georgia Tech

RLVR can verify correctness, but is correctness enough to train a general-purpose model?

Task / objective	Reward source	What it evaluates
Reasoning accuracy	Rule-based verifier (RLVR)	Final answers and code tests
General helpfulness	Learned preference model	Final summary to the user
General harmlessness	Learned safety model	Full response, including reasoning

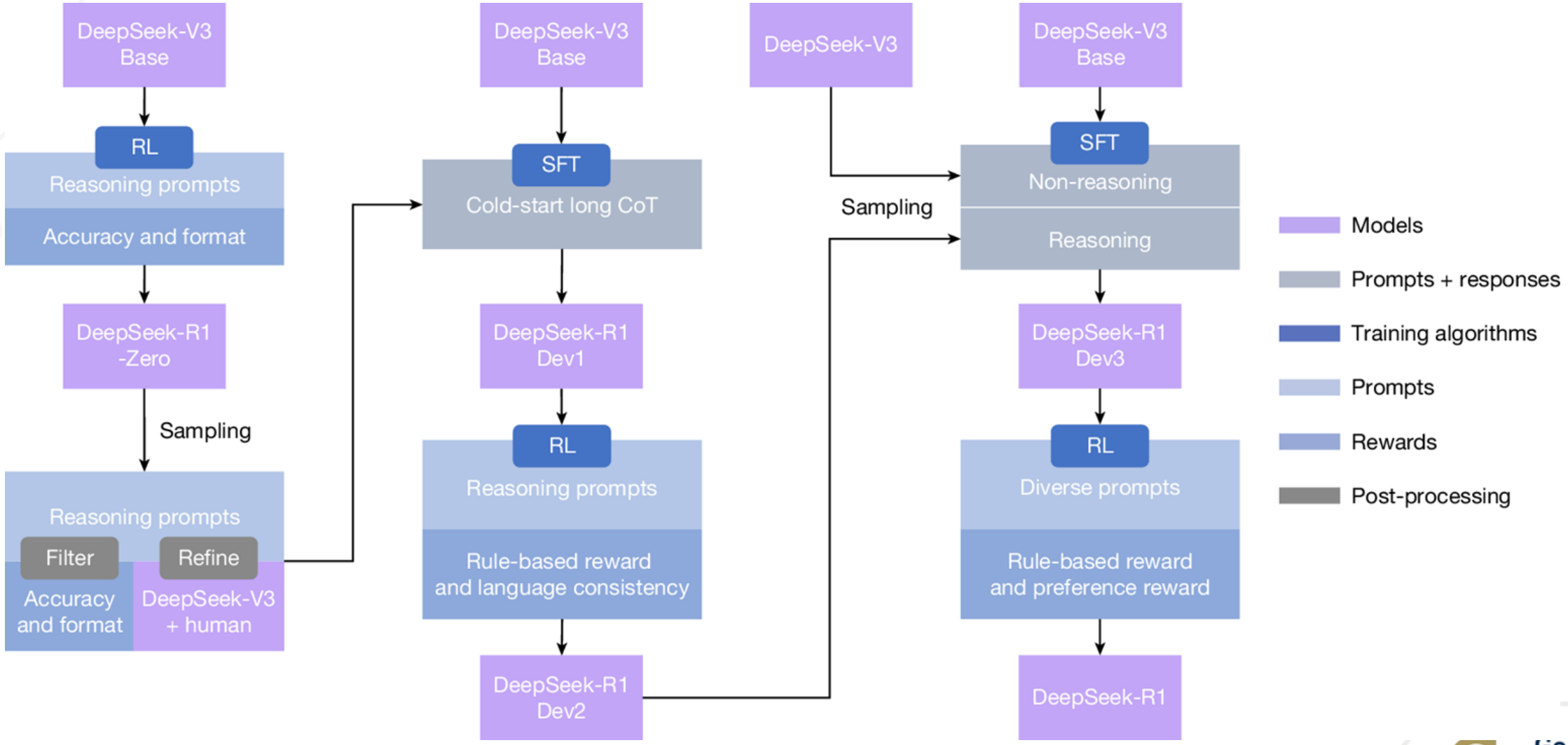
(and also auxiliary rule-based rewards: output format and language consistency)

R1 uses learned rewards for general-purpose alignment.

Guo et al., Nature (2025), Methods: Reward design, Helpful reward model, Safety reward model.

Georgia Tech

DeepSeek-R1 training pipeline



Guo et al., Nature (2025), Fig. 2. Reproduced, resized; CC BY 4.0.

Georgia Tech

R1 results: reasoning and general assistant performance

Table 2. Experimental results at each stage of DeepSeek-R1 (excerpt)

Benchmark (metric)	R1-Zero	Dev1	Dev2	Dev3	R1
AIME 2024 (pass@1)	77.9	59.0	74.0	78.1	**79.8**
LiveCodeBench (pass@1-CoT)	50.0	57.5	63.5	64.6	**65.9**
IF-Eval (prompt strict)	46.6	71.7	72.0	78.1	**83.3**
AlpacaEval 2.0 (LC win rate)	24.7	50.1	55.8	62.1	**87.6**
GPQA Diamond (pass@1)	75.8	66.1	70.7	71.2	**71.5**
CNMO 2024 (pass@1)	88.1	58.0	73.9	77.3	**78.8**

Dev1: cold-start SFT Dev2: reasoning RL Dev3: rejection sampling and SFT

Takeaway: With RLVR, R1 becomes **a more useful general-purpose model while keeping strong reasoning.**

Guo et al., Nature (2025), Table 2, selected rows. Values are percentages. Bold identifies the final R1 column.

Georgia Tech

Why RLVR training helps LLM to reason?

Georgia Tech

RLVR: Conditions for Reasoning Improvement

1. Successful strategies can be sampled.

$$p_{\text{success}} > 0$$

2. Sampled trajectories have different rewards.

$$\text{Var}(R_1, \ldots, R_G) > 0$$

3. Correct reasoning earns higher expected reward.

$$\Pr(R = 1 \mid \text{correct CoT}) > \Pr(R = 1 \mid \text{incorrect CoT})$$

RLVR can reinforce correct reasoning when these conditions hold.

Source: Wen et al. (2025), §3. Simplified GRPO mechanism.

Georgia Tech

RLVR: When the Conditions Fail

1. No successful strategy is sampled.

$$0/64$$

Correct attempts on a hard task

2. Rewards do not vary within the group.

$$R = (1, 1, 1, 1) \quad \Rightarrow \quad A_i = 0$$

An easy task gives four correct answers.

3. Shortcuts earn rewards as often as correct reasoning.

$$\Pr(R = 1 \mid \text{correct CoT}) \approx \Pr(R = 1 \mid \text{incorrect CoT})$$

A leaked answer key makes copying succeed.

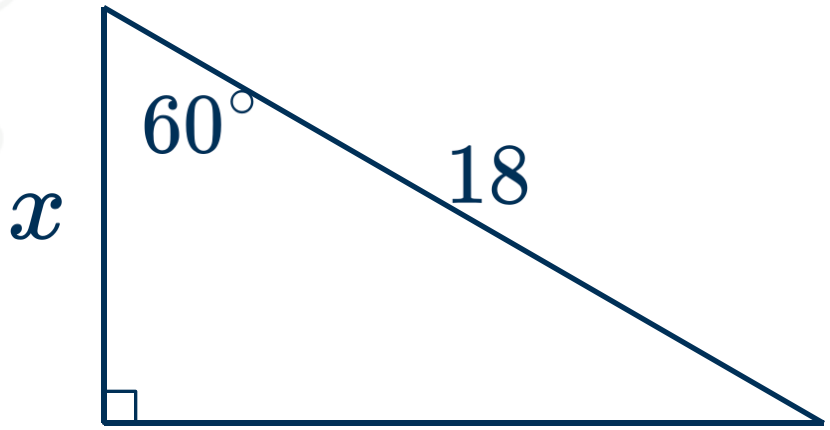
Sampling and reward design determine whether RLVR has a useful learning signal.

Illustrative cases. Sources: Wen et al. (2025), DAPO (2025).

Georgia Tech

Multimodal RLVR: What's Different

Find x.



1 Perception & fusion

Identify the labeled side and angle.

↓

2 Reason

$$x = 18\cos 60^\circ = 9$$

↓

3 Verify

$$R = \mathbf{1}[x = 9]$$

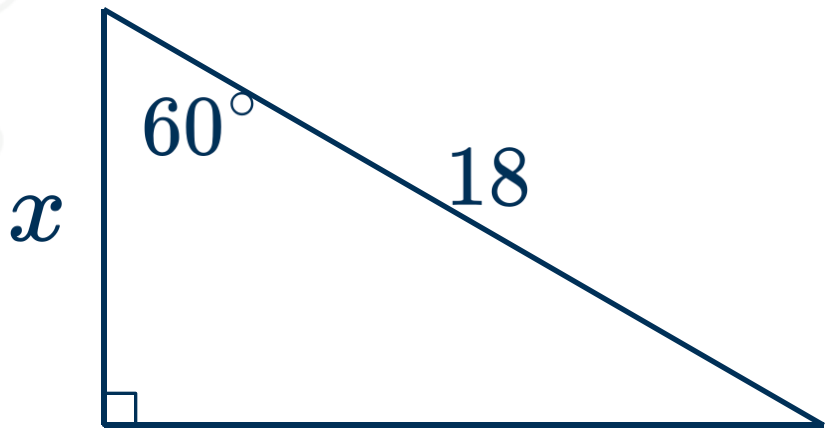
Multimodal RLVR must learn to recover and use visual evidence.

PAPO Fig. 1, diagram redrawn. Failure variants are illustrative.

Georgia Tech

Multimodal RLVR: Why This Training Can Fail

Find x.



$$x = 18 \sin 60^\circ = 9\sqrt{3}$$

Inherited problems from pure-text RLVR

1 Every rollout reuses the same wrong reading.
2 Every answer fails the verifier.
3 **A faulty rationale still gets the answer right.**

New problems from visual

A label is unreadable.
Fusion fails:x is read correctly but linked to the wrong side.
etc. (cont. later)

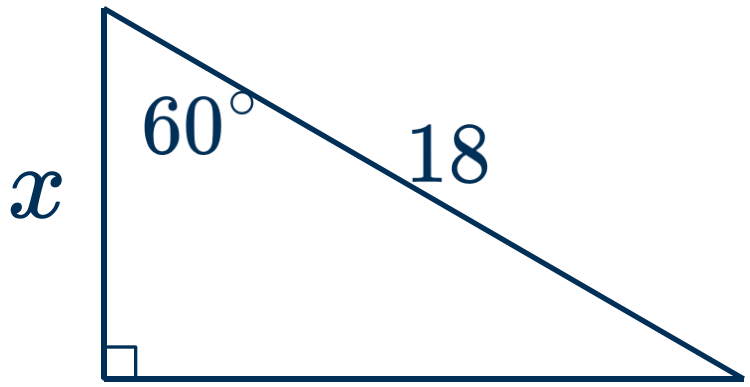
Answer accuracy alone does not reveal where the visual chain failed for credit assignment.

PAPO Fig. 1, diagram redrawn. Failure variants are illustrative.

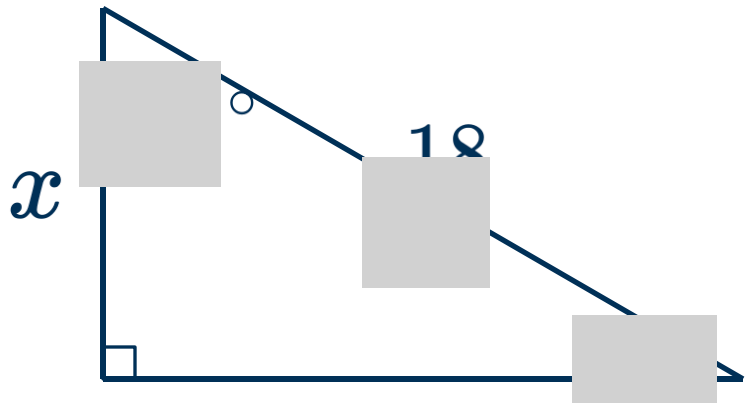
Georgia Tech

The model can give correct answer without looking at the picture.

Original image



Masked image



We should punish such action, but the correct answer gives positive advantage.

Source: PAPO, §§3 and 5.3. Masking shown schematically.

Georgia Tech

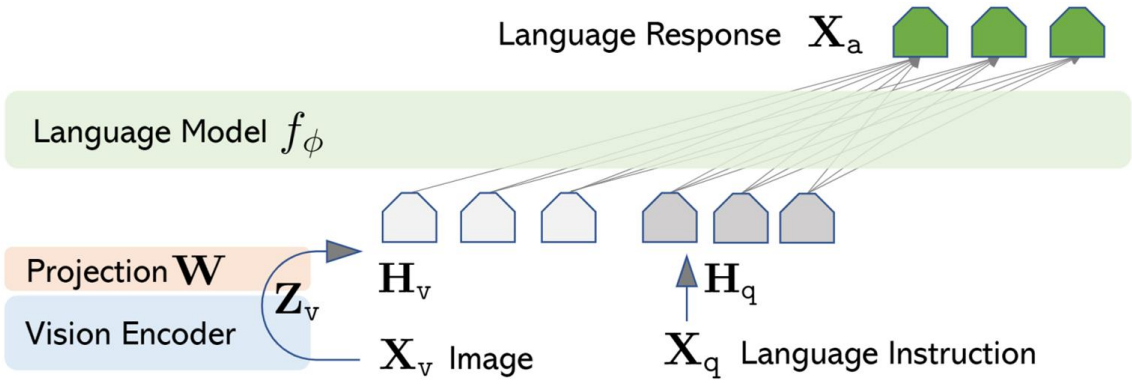
Visual Reasoning Before RLVR

VISUAL INSTRUCTION TUNING

LLaVA connects a vision encoder to a language model and trains image-conditioned responses.

MULTIMODAL CHAIN OF THOUGHT

Rationale generation and answer inference can use multimodal inputs and supervised reasoning examples.



LLaVA architecture

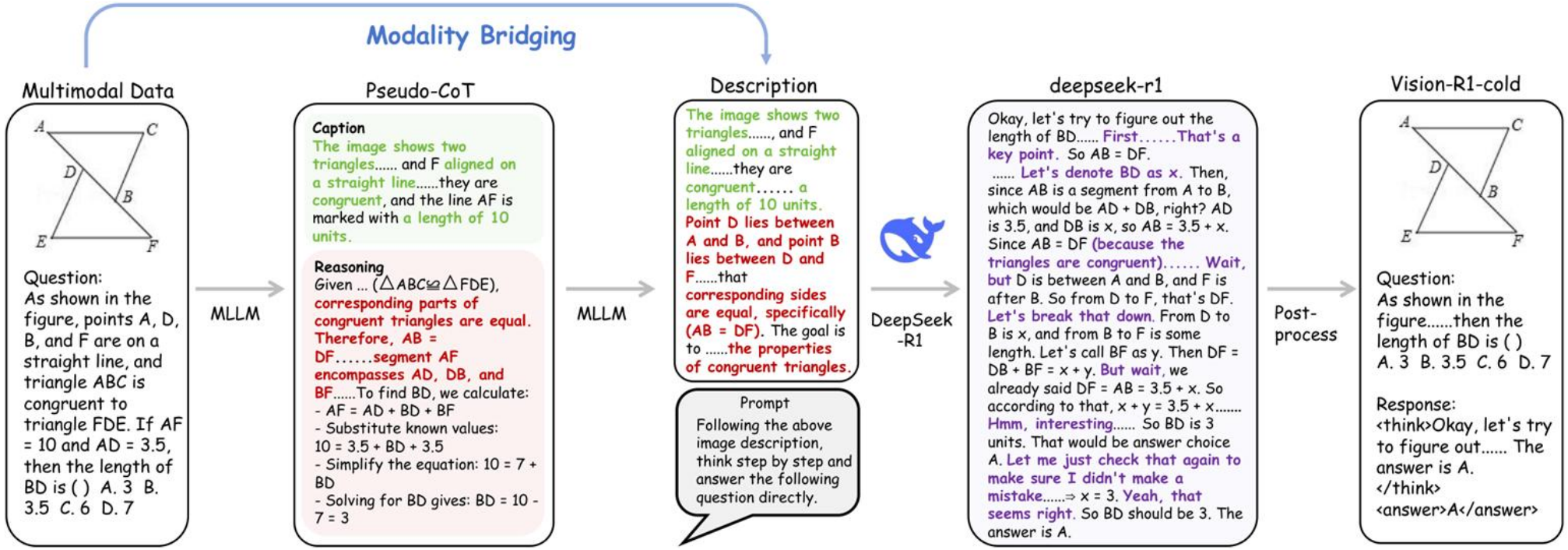
RLVR changes the training signal. It still needs reliable visual evidence.

Sources: Liu et al., Visual Instruction Tuning (2023). Zhang et al., Multimodal CoT (2023).

Georgia Tech

Vision-R1: A Visual Bridge for a Text Teacher

Intuition: transfer R1 reasoning through descriptions that preserve the image evidence.



Method: 200K cold-start CoTs, then GRPO with progressively relaxed reasoning-length limits.

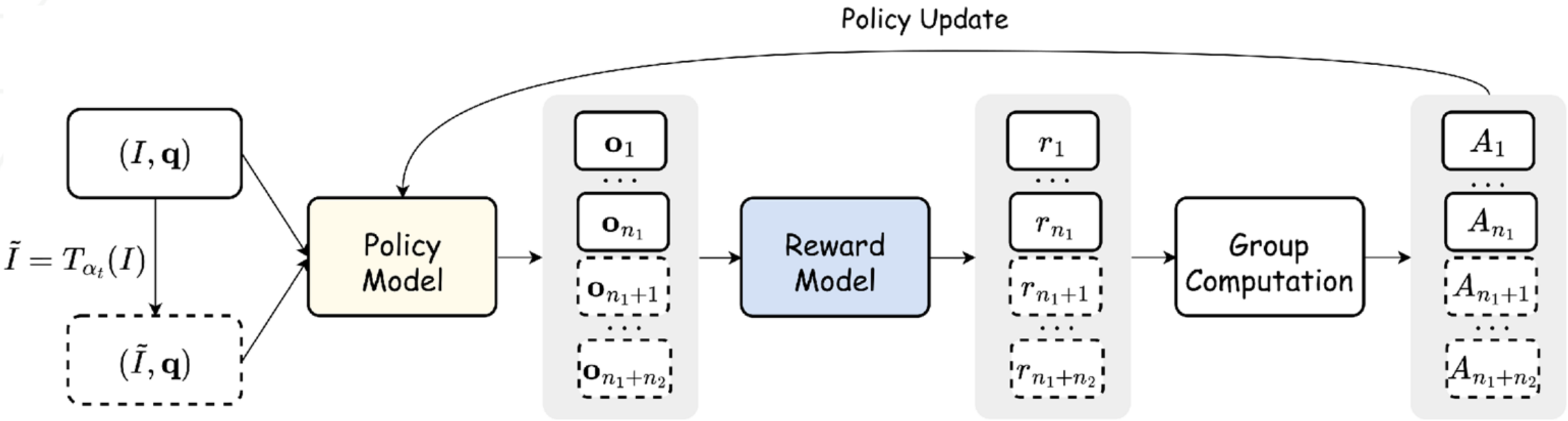
Our takeaway: strong text reasoning still depends on an accurate visual description.

Source: Vision-R1, §§3.3–3.4, Fig. 2. PTST initially restricts reasoning length and later relaxes the limit.

Georgia Tech

NoisyRollout: Exploration Through Image Noise

Intuition: text sampling can repeat the same visual mistake. Image perturbations offer different trajectories.



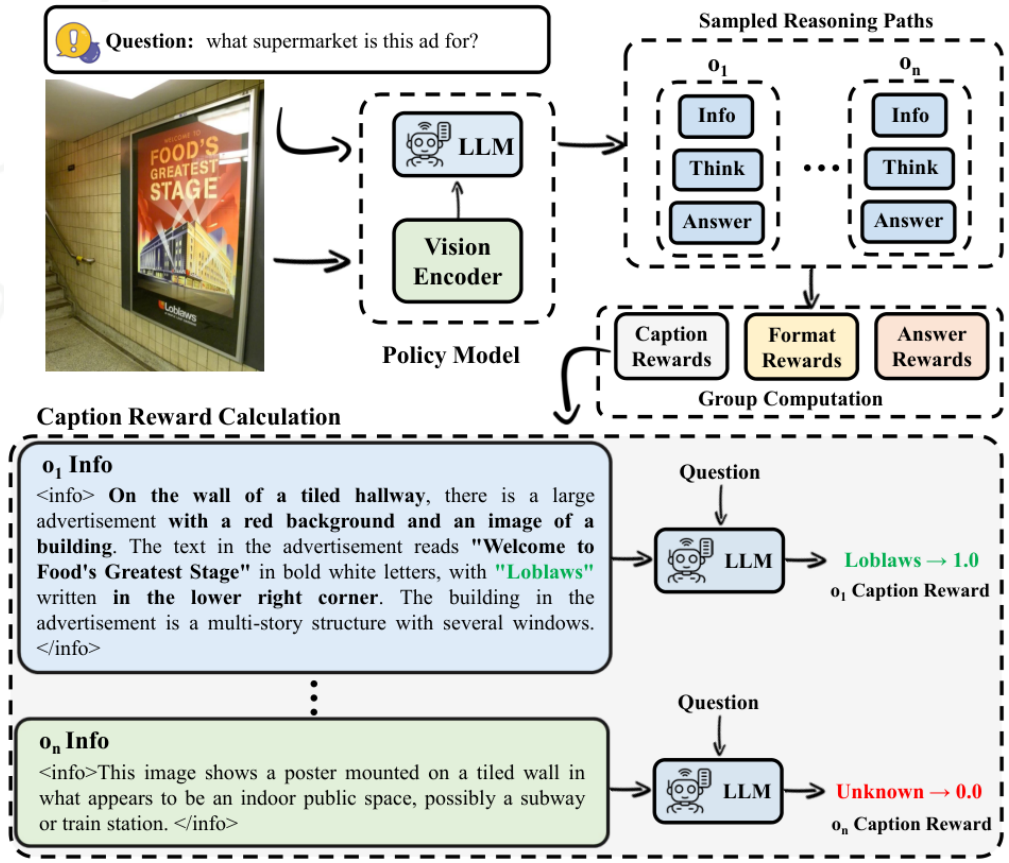
Mix clean and noisy rollouts for GRPO. Update on the clean image and gradually reduce noise.

Our takeaway: useful diversity requires noise that preserves enough evidence.

Source: NoisyRollout, §2 and Fig. 2. Noise strength decreases during training to reduce mismatch.

Georgia Tech

Visionary-R1: Rewarding Useful Captions



INTUITION

Make visual interpretation explicit before reasoning.

METHOD

Generate a caption, reasoning, and answer. Reward a caption when the policy's LLM can answer from it.

Combine caption, format, and answer rewards in GRPO.

Our takeaway: caption usefulness still leaves individual visual claims unchecked.

Source: Visionary-R1, §3 and Fig. 3. The method also anneals the reference-KL coefficient.

Georgia Tech

Failure Modes Along the Visual Evidence Chain

The same wrong answer can result from failures at different stages.

Failure	What can go wrong	What the answer reward misses
Perception and binding	Misread “8” as “3”, or attach a label to the wrong object.	Which visual fact or relation failed.
Language bias	Use a familiar question pattern without checking the image.	Whether the image was needed.
Visual hallucination	Introduce an unsupported object, value, or geometric relation.	Whether each stated premise is true.
Exploration and credit	Different CoTs reuse the same incorrect visual premise.	How to correct a shared upstream error.

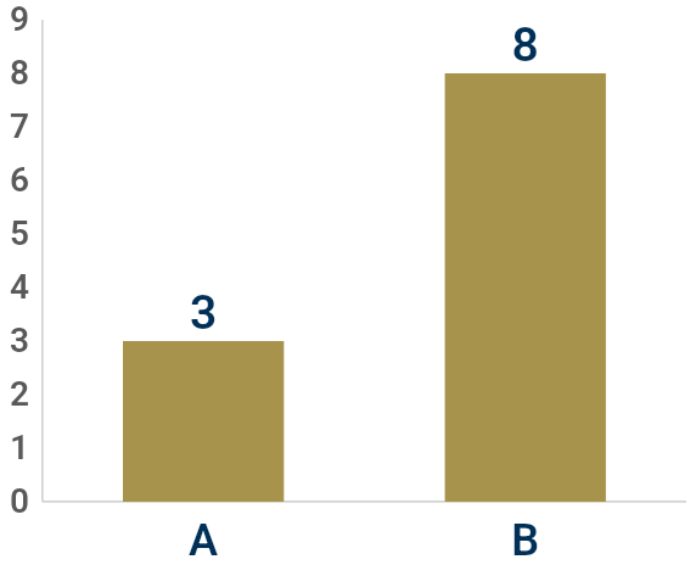
Our takeaway: more reasoning traces can preserve the same visual error.

Synthesis from Shimin’s task setup, Visionary-R1 (§3.1), NoisyRollout (§2), and PAPO (§3).

Georgia Tech

Why Answer Rewards Miss Image Use

Question: Which bar is higher, A or B?



Illustrative example

Response path	Answer	Reward
Reads A = 3 and B = 8 from the image.	B	1
Guesses B from a learned question pattern.	B	1

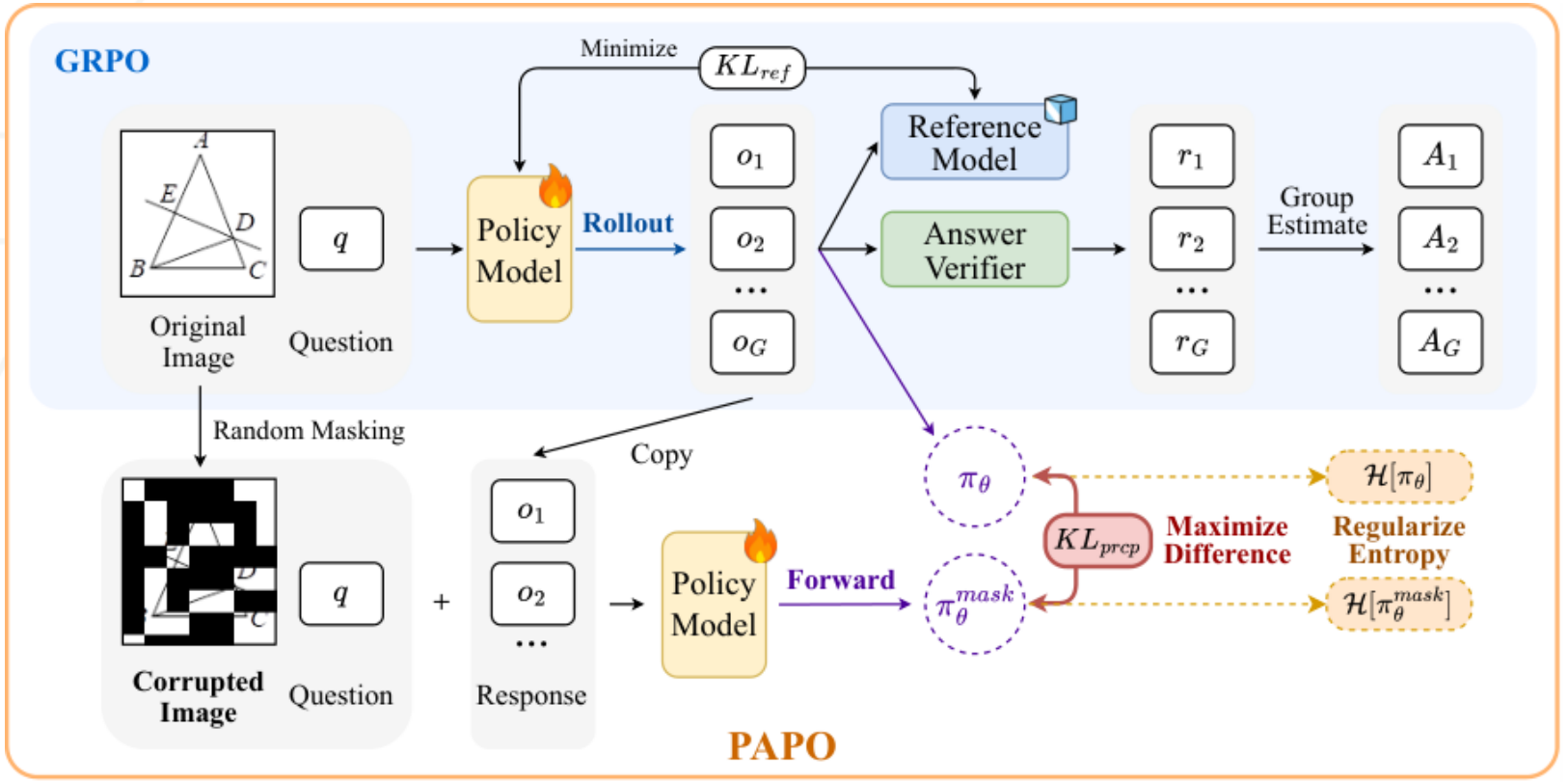
Equal outcome rewards cannot distinguish these paths in this example.

Image dependence is a proxy. Visual factual accuracy needs a separate test.

Illustration created for this talk. Motivation: PAPO §3. No empirical model outputs are shown.

Georgia Tech

PAPO: Same Rollout, Two Visual Conditions



WHAT STAYS FIXED

The question, response tokens, and preceding token prefix.

WHAT CHANGES

The image used to compute token probabilities.

One extra masked forward pass scores the existing rollout.

NoisyRollout samples new trajectories. PAPO rescores fixed response tokens.

Source: PAPO §3.2, original method overview. No extra perception labels or reward model.

Georgia Tech

The PAPO Objective

$$\mathcal{J}_{\text{PAPO}} = \mathcal{J}_{\text{GRPO}} + \gamma \mathbb{D}_{\text{KL}}\left[\pi_\theta \,\|\, \pi_\theta^{\text{mask}}\right] - \eta_1 \mathcal{H}[\pi_\theta] - \eta_2 \mathcal{H}[\pi_\theta^{\text{mask}}]$$

TASK REWARD

GRPO keeps the verified-answer signal and its policy-update constraints.

PERCEPTUAL DIFFERENCE

Maximize the original–masked divergence. γ controls its weight.

ENTROPY REGULARIZATION

Penalize the paper's entropy terms on both visual conditions.

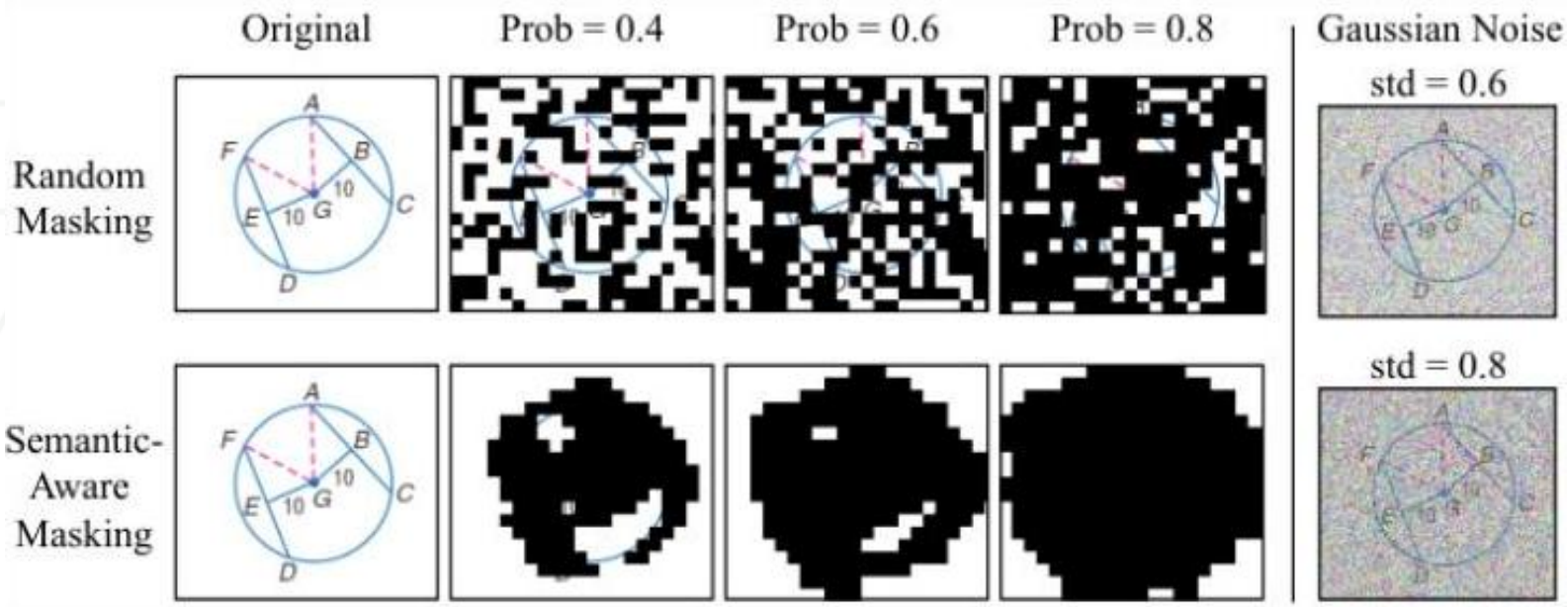
The reference KL inside GRPO pulls toward a reference policy. The perceptual KL pushes the two visual conditions apart.

A larger perceptual gap helps only if optimization preserves useful responses.

Source: PAPO §§3.2 and 5.2. The perceptual and reference-policy KL terms have different roles.

Georgia Tech

Masking Defines the Comparison



Random masking is simple and avoids selecting a particular semantic region.

3B ABLATION, γ = 0.01

Masking	Avg. (%)
GRPO	47.92
Random 0.4	48.78
Random 0.6	**49.26**
Random 0.8	48.87
Random 1.0	48.51
Semantic 0.6	48.42

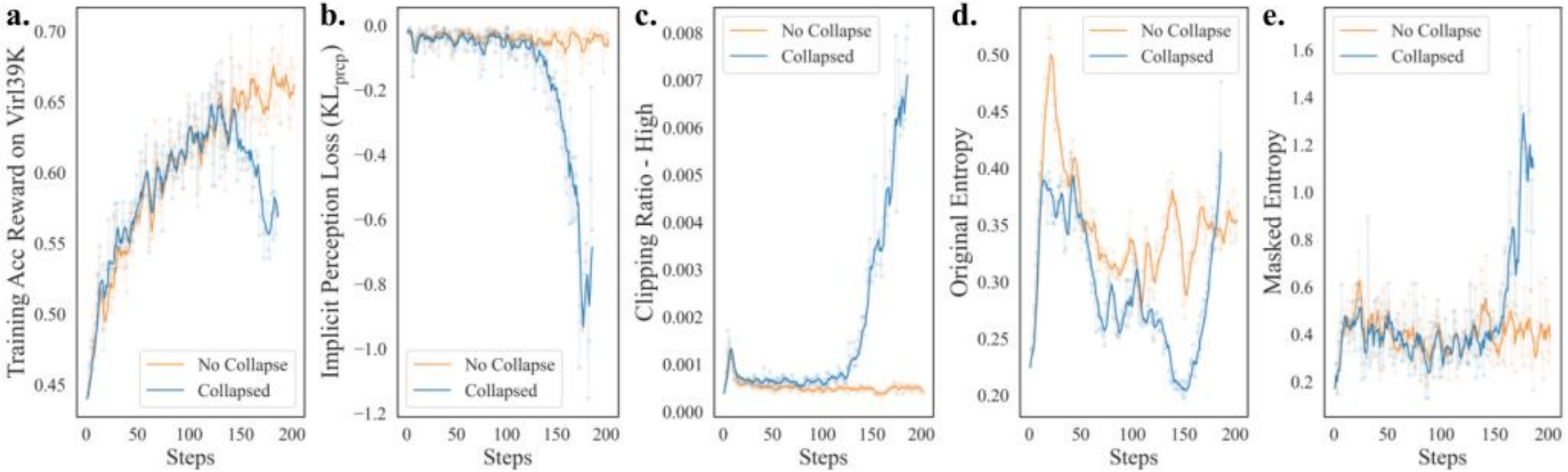
Our reading: partial masking gives a more useful contrast than a fully black image.

Source: PAPO masking ablation. This comparison does not establish a universal best ratio.

Georgia Tech

Failure Mode: Perceptual KL Hacking

A policy can increase the auxiliary divergence while its task performance deteriorates.



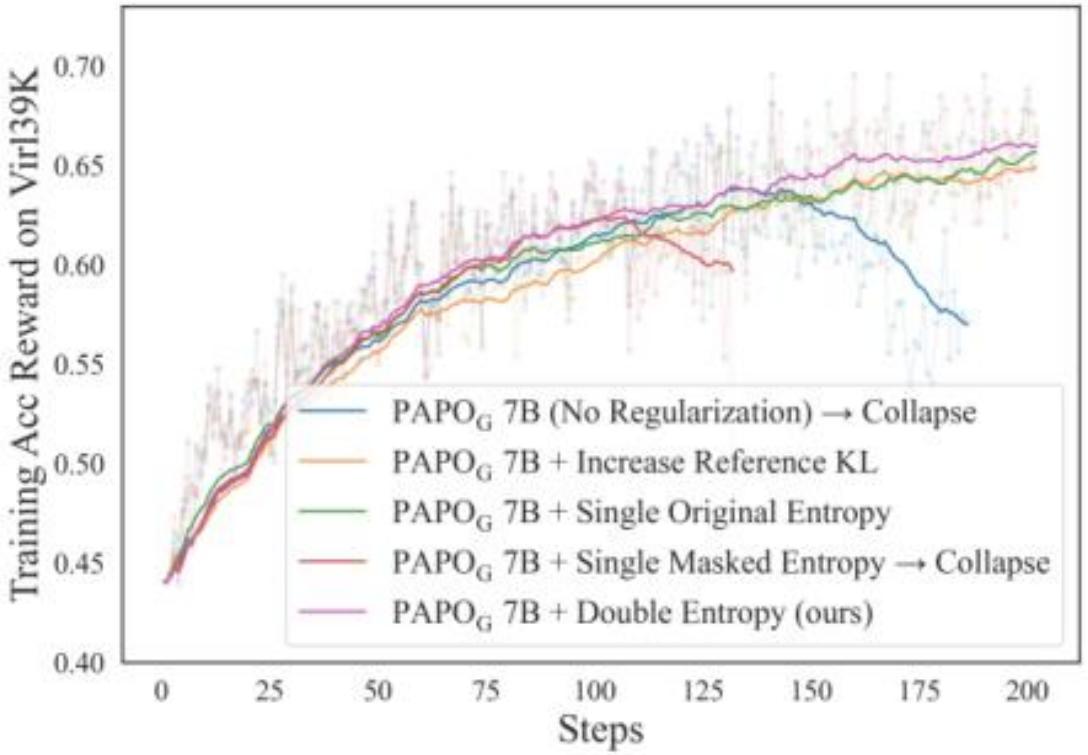
Observed pattern: answer reward falls, clipping grows, and both entropy terms become unstable.

Source: PAPO arXiv v5, Fig. 15. Compared runs differ in γ and regularization; this is a diagnostic pattern.

Georgia Tech

Double Entropy and the Evaluation Question

Regularizing both visual conditions stabilizes the shown PAPO training run.



QWEN2.5-VL-7B

Training variant	Avg. (%)
GRPO	58.78
PAPO + stronger reference KL	60.42
PAPO + original entropy	61.12
PAPO + double entropy	**61.66**

Single masked entropy delays collapse in this experiment. Double entropy controls both branches.

Next: do gains persist with stronger baselines and factual grounding tests?

Source: PAPO regularization ablation. Overall scores are averages across the nine reported benchmarks.

Georgia Tech

Experimental Setup

01	02	03	04	05
DATA	**STARTING MODELS**	**RLVR RUNS**	**REWARD**	**EVALUATION**
ViRL39K 39k verifiable image–question pairs	Qwen2.5-VL 3B / 7B Qwen3-VL 2B	Same data and checkpoints GRPO, DAPO baseline run + comparable PAPO run	Exact ground truth final answer. No human CoT labels	8 multimodal benchmarks 8 samples per question exact-answer avg@8

CONTROLLED COMPARISONS

GRPO → PAPO_G = GRPO + perception KL + double-entropy regularization

DAPO → PAPO_D = DAPO + perception KL + double-entropy regularization

Under same base model checkpoint, training data, ground truth, evaluate improvements coming from changing the RL objective.

Georgia Tech

Overall Results: PAPO Improves Accuracy on Reasoning Benchmarks

Method	General Multimodal Reasoning							Vision-Dependent Multimodal Reasoning						Overall	
	Geo3k	MathVista	We-Math	MMKI2	MathVerse	AVG	$\Delta^{\%}_{rel}$	LogicVista	Counting	MMMU-Pro	MathVerse$_V$	AVG	$\Delta^{\%}_{rel}$	AVG	$\Delta^{\%}_{rel}$
Qwen2.5-VL															
GRPO-3B	28.72	59.34	58.90	57.24	55.25	51.89	–	38.14	55.81	25.66	52.26	42.97	–	47.92	–
PAPO$_G$-3B	**30.95**	**61.38**	**60.09**	**57.39**	**57.14**	**53.39**	↑ 3.38	**38.67**	**62.56**	**27.11**	**53.95**	**45.57**	↑ 5.60	**49.92**	↑ 4.36
GRPO-7B	40.18	65.48	**68.12**	72.26	66.51	62.51	–	45.62	73.94	35.17	61.71	54.11	–	58.78	–
PAPO$_G$-7B	**40.25**	**69.53**	66.79	**72.52**	**68.43**	**63.50**	↑ 1.53	**46.07**	**89.81**	**36.63**	**64.97**	**59.37**	↑ 7.96	**61.66**	↑ 4.39
DAPO-3B	31.20	60.89	59.95	**66.83**	56.25	55.02	–	40.69	74.25	28.42	53.09	49.11	–	52.40	–
PAPO$_D$-3B	**35.65**	**62.53**	**62.67**	64.09	**60.51**	**57.09**	↑ 5.00	**41.67**	**83.56**	**28.76**	**57.72**	**52.93**	↑ 5.97	**55.24**	↑ 5.54
DAPO-7B	35.92	61.91	58.51	75.93	55.64	57.58	–	37.05	90.05	29.02	51.04	51.79	–	55.01	–
PAPO$_D$-7B	**44.11**	**67.53**	**68.30**	**80.61**	**68.58**	**65.83**	↑ 15.61	**46.70**	**91.38**	**36.34**	**64.87**	**59.82**	↑ 19.09	**63.16**	↑ 17.54
Qwen3-VL (thinking)															
GRPO-2B	39.29	53.58	57.12	47.71	47.98	49.13	–	29.84	80.13	20.51	45.41	43.97	–	46.84	–
PAPO$_G$-2B	**41.08**	**56.08**	**59.17**	**48.57**	**51.89**	**51.36**	↑ 4.52	**32.83**	**80.63**	**23.42**	**50.05**	**46.73**	↑ 6.27	**49.30**	↑ 5.25

AVG@8

avg@8 = (total correct) / (8N)
N = number of evaluation questions.

RELATIVE CHANGE

$$\Delta^{\%}_{\text{rel}} = \frac{100}{N} \sum_{j=1}^{N} \frac{A_j^{\text{PAPO}} - A_j^{\text{baseline}}}{A_j^{\text{baseline}}}$$

Interpretation: PAPO improves results across all base models sizes, and over both GRPO and DAPO on almost all 9 tested benchmarks.

Georgia Tech

Where Does PAPO Help and Where Does It Regress?

Method	General Multimodal Reasoning							Vision-Dependent Multimodal Reasoning						Overall	
	Geo3k	MathVista	We-Math	MMKI2	MathVerse	AVG	$\Delta^{\%}_{rel}$	LogicVista	Counting	MMMU-Pro	MathVerse$_V$	AVG	$\Delta^{\%}_{rel}$	AVG	$\Delta^{\%}_{rel}$
Qwen2.5-VL															
GRPO-3B	28.72	59.34	58.90	57.24	55.25	51.89	–	38.14	55.81	25.66	52.26	42.97	–	47.92	–
PAPO$_G$-3B	**30.95**	**61.38**	**60.09**	**57.39**	**57.14**	**53.39**	↑ 3.38	**38.67**	**62.56**	**27.11**	**53.95**	**45.57**	↑ 5.60	**49.92**	↑ 4.36
GRPO-7B	40.18	65.48	**68.12**	72.26	66.51	62.51	–	45.62	73.94	35.17	61.71	54.11	–	58.78	–
PAPO$_G$-7B	**40.25**	**69.53**	66.79	**72.52**	**68.43**	**63.50**	↑ 1.53	**46.07**	**89.81**	**36.63**	**64.97**	**59.37**	↑ 7.96	**61.66**	↑ 4.39
DAPO-3B	31.20	60.89	59.95	**66.83**	56.25	55.02	–	40.69	74.25	28.42	53.09	49.11	–	52.40	–
PAPO$_D$-3B	**35.65**	**62.53**	**62.67**	64.09	**60.51**	**57.09**	↑ 5.00	**41.67**	**83.56**	**28.76**	**57.72**	**52.93**	↑ 5.97	**55.24**	↑ 5.54
DAPO-7B	35.92	61.91	58.51	75.93	55.64	57.58	–	37.05	90.05	29.02	51.04	51.79	–	55.01	–
PAPO$_D$-7B	**44.11**	**67.53**	**68.30**	**80.61**	**68.58**	**65.83**	↑ 15.61	**46.70**	**91.38**	**36.34**	**64.87**	**59.82**	↑ 19.09	**63.16**	↑ 17.54
Qwen3-VL (thinking)															
GRPO-2B	39.29	53.58	57.12	47.71	47.98	49.13	–	29.84	80.13	20.51	45.41	43.97	–	46.84	–
PAPO$_G$-2B	**41.08**	**56.08**	**59.17**	**48.57**	**51.89**	**51.36**	↑ 4.52	**32.83**	**80.63**	**23.42**	**50.05**	**46.73**	↑ 6.27	**49.30**	↑ 5.25

AUTHORS' BROADER EXPLANATION

PAPO's gains are largest on high-vision-dependency tasks. We-Math and MMK12 mix in questions partly answerable from text.

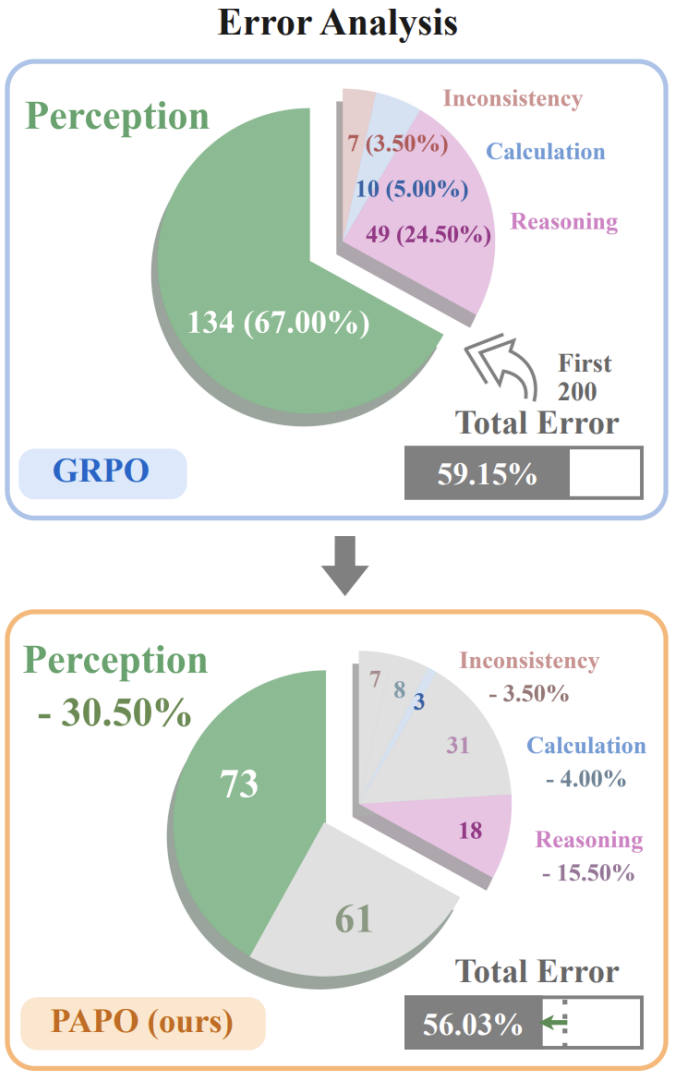
THE TWO REGRESSIONS

We-Math: 68.12 → 66.79 (−1.33 pp; GRPO-7B)
MMK12: 66.83 → 64.09 (−2.74 pp; DAPO-3B)

No task-specific cause is reported. Possible interpretation: rewarding image sensitivity can add little on text-solvable items. Without multi-seed error bars, significance is hard to measure.

Georgia Tech

Does Better Accuracy Mean Better Perception?



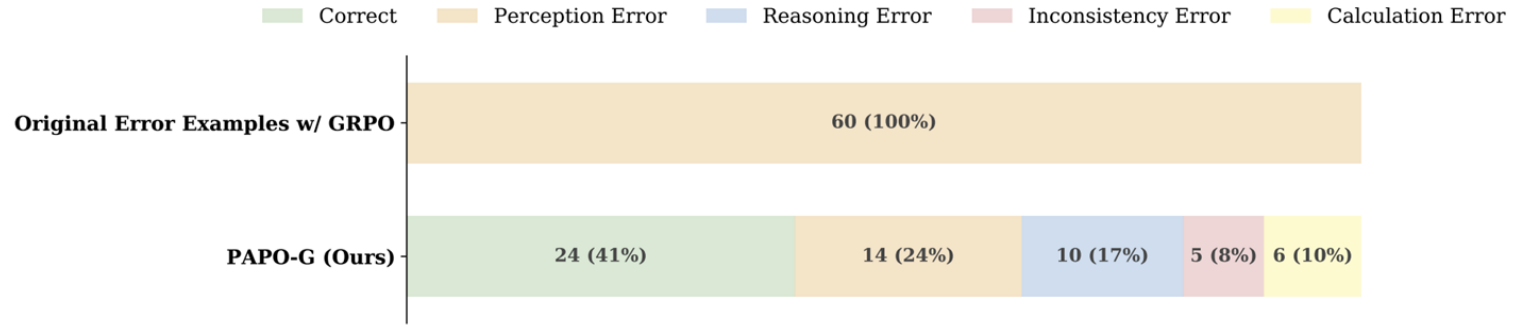


Figure 15: Detailed analysis of 60 GRPO perception-error cases and their outcomes under PAPO.

Table 12: **Results on OCR-bench-v2.** The base model is Qwen2.5-VL.

Size	Method	OCR-bench-v2 (%)
3B	Base	20.14
3B	GRPO	26.32
3B	**PAPO$_G$**	**27.22**
7B	Base	27.75
7B	GRPO	31.17
7B	**PAPO$_G$**	**32.11**

THREE DIFFERENT PIECES OF EVIDENCE

1. Fewer perception related failures (134 → 73)
2. Many prior GRPO perception error either became correct or became other types of errors.
3. OCR benchmark: better literal visual reading.

PAPO improves perception, but remaining failures reveal downstream reasoning, calculation, and consistency limits.

Georgia Tech

How the Authors Classified Errors

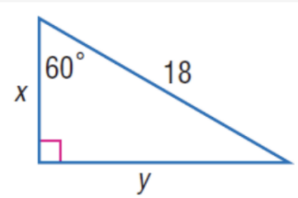
The authors manually labeled 200 GRPO errors from Geo3K, MMK12, LogicVista, and MathVerse

01	PERCEPTION	Incorrectly reads visual objects, values, labels, or spatial relations.
02	REASONING	Makes a logical-inference error, such as applying an incorrect rule or theorem.
03	CALCULATION	Uses an appropriate setup but makes an arithmetic or computation mistake.
04	INCONSISTENCY	The intermediate reasoning and final answer contradict one another.

Georgia Tech

An Example Perception Error

Inference Example



Question: Find x.

Ground Truth: 9

INPUT

GRPO: Perception Error Example

<think> To find the side x of the right-angled triangle, we can use trigono-metric ratios. The triangle has a 60-degree angle, and the hypotenuse is 18. Since x is the side opposite the 60-degree angle, we can use the sine function, which is defined as: ... Therefore, the value of x is $9\sqrt{3}$ </think> \boxed{ $9\sqrt{3}$ }

GRPO ERROR

PAPO: Corrected Perception

<think> ... In a 30-60-90 triangle, the hypotenuse is twice the length of the short leg (the side opposite the 30-degree angle). Here, the hypotenuse is 18, so the short leg, which is x, is half of 18. The long leg is y.

Therefore: $x = \frac{18}{2} = 9$ </think> \boxed{9}

PAPO CORRECTION

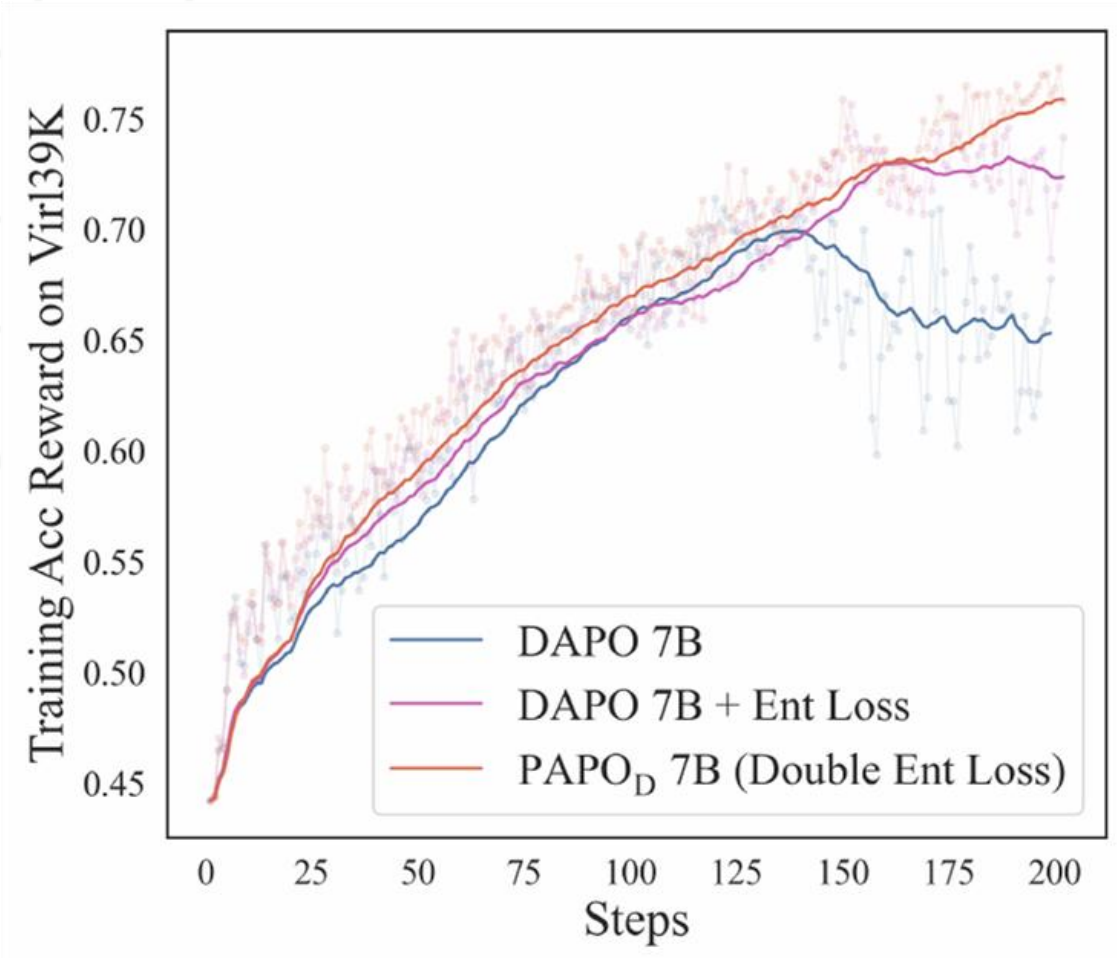
GRPO misidentifies which side x denotes; the visual error caused model to output 9√3.

PAPO maps x to the short leg and reaches the correct result, x = 9.

This illustrates perception → reasoning error propagation.

Georgia Tech

Training Stability: Preventing Late-Stage Collapse



Double entropy adds stability to training.

Method	General		Vision		Overall	
	AVG	$\Delta^{\%}_{rel}$	AVG	$\Delta^{\%}_{rel}$	AVG	$\Delta^{\%}_{rel}$
DAPO-7B	57.58	–	51.79	–	55.01	–
DAPO-7B + Ent Loss	64.77	↑ 13.52	59.14	↑ 17.73	62.27	↑ 15.86
PAPO$_D$-7B (w/ Double Ent)	**65.83**	↑ 15.61	**59.82**	↑ 19.09	**63.16**	↑ 17.54

1. Plain DAPO climbs, then its reward drops late in training.

2. A single entropy term delays and softens the drop.

3. PAPO_D + double entropy avoids the observed collapse in this run.

Why collapse? Repeated reward optimization can make the policy drift into unstable, low-quality output patterns when it is weakly constrained.

Georgia Tech

What Persisted: The RLVR Backbone

THE SHARED RLVR LOOP

SAMPLE A GROUP	VERIFY OUTCOMES	COMPARE WITHIN GROUP	UPDATE THE MODEL	REPEAT
Generate several complete answers to one prompt	Check the correctness of each answer with verifier	Responses are judged relative to one another	Favor higher-reward answers; limit large changes	Sample again from updated model

SHARED OPTIMIZATION CHOICES

NO CRITIC	SELF-GENERATED	OUTCOME-ONLY	BOUNDED UPDATE
Group rewards replace a learned value model	The model learns from its own candidate solutions	Score the final result; do not grade each step	Clipping prevents one step from changing too much

Both retain grouped, verifier-driven RL over self-generated responses.

Georgia Tech

What Changed: From Multimodal RLVR Follow-ons to PAPO

DEEPSEEK-R1

Text-only RLVR
Verifiable outcome rewards
Emergent reasoning without labeled CoT

MULTIMODAL FOLLOW-ONS

Add images and multimodal data
Often add perception rewards, CoT supervision, or explicit perception-before-reasoning stages

PAPO

Keep outcome-based RLVR
Change the optimization objective. Encourage image-dependent token probabilities

Evidence requirement also changes: final-answer benchmark accuracy is no longer enough, we must ask whether visual evidence was used correctly.

Progression: outcome-trained reasoning → image-dependent reasoning.

Georgia Tech

Why Perception KL Is Not Proof of Factual Grounding

$$\mathbb{D}_{\mathrm{KL}}[\pi_\theta || \pi_\theta^{\mathrm{mask}}] = \mathbb{D}_{\mathrm{KL}}[\pi_\theta(o|q, I) \parallel \pi_\theta(o|q, I_{\mathrm{mask}})].$$

WHAT A HIGHER SIGNAL DOES IN RL

- Adds token-level guidance beyond the final-answer reward.
- Strengthens the training incentive for responses to depend on visual evidence.
- Potential payoff: fewer mistakes on diagrams, charts, forms, or interfaces where visual details matter.

WHAT IT DOES NOT ESTABLISH

- The model read the correct object, number, or relation.
- Masking removed visual evidence that was actually needed for the answer.
- The response depended on task-relevant image content rather than the artificial masking pattern.

Training becomes more image-dependent; factual grounding and causal relevance still require separate evaluation.

Georgia Tech

How Can We Evaluate Factual Visual Grounding?

PAPO tests image sensitivity, but not correct image fact extraction or use in reasoning.
A stronger test changes one visual fact and checks whether the answer changes for the right reason.

01	EXTRACT THE FACTS	Check named objects, numbers, labels, and relations against the image.
02	EDIT RELEVANT EVIDENCE	Change the fact needed for the answer; the model's answer should update.
03	EDIT IRRELEVANT BACKGROUND	Change visual content unrelated to the question; the answer should stay stable.
04	QUESTION-CONDITIONED MASKING	Mask question-relevant evidence and expect the answer to change; mask an equal-sized irrelevant region and expect stability.

Georgia Tech

Remaining Limitations, Failure Modes, and Cost

SIGNAL VALIDITY

KL_prcp rewards sensitivity to masked pixels, not verified correctness or relevance.

EVIDENCE SCOPE

Mostly exact-answer tasks and avg@8; no multi-seed confidence intervals.

SCALE

Largest evaluated model is 7B; behavior at larger scales and in high-stakes settings is untested.

Model	Method	Per Step (s)	Additional Forward Pass (s)
3B	GRPO	360.9	-
	PAPO$_G$	428.1	48.8
7B	GRPO	258.5	-
	PAPO$_G$	367.1	49.7

TRAINING COST

PAPO raises time per training step, +18.6% for 3B and +42.0% for 7B. The masked-image branch is training-only; inference still uses one standard pass.

Trade-off: stronger multimodal RL but more training compute

Georgia Tech

Final Takeaways: What PAPO Adds—and What Remains Open

DeepSeek-R1 → multimodal credit-assignment gap → PAPO

WHAT PAPO ADDS

A masked-image perception objective encourages the policy to depend on visual evidence; double-entropy regularization keeps that optimization stable.

WHAT THE EVIDENCE SUPPORTS

PAPO improves all five matched overall comparisons, with larger gains on vision-dependent tasks. Stability explains most of the largest DAPO improvement.

WHAT REMAINS OPEN

PAPO does not verify that the model read the correct visual facts. Image dependence is evidence of visual use—not proof of factual grounding.

Bottom line: PAPO makes multimodal RLVR more image-dependent and stable—but it is not a factual-grounding verifier.

Georgia Tech

Discussion I: Rewards and Multimodal Failure

R1 / REWARD

1 If a model gets the final answer right using a hallucinated or invalid rationale, should outcome-only RLVR reinforce it?

IMAGE USE

2 Should a multimodal model always be rewarded for depending on the image—even when the image is irrelevant, ambiguous, or misleading?

MASKING

3 What intervention could test image use without rewarding sensitivity to random masking artifacts?

Georgia Tech

Discussion II: Evidence and Contribution Attribution

CONTRIBUTION ATTRIBUTION

4 DAPO rises 55.01 → 62.27 with single entropy, while PAPO_D reaches 63.16. How much of the gain is perception rather than stability?

ACTUAL GROUNDING

5 What experiment would convince you that PAPO reads visual facts more accurately—not merely that its outputs change when pixels are masked?

GENERALITY + COST

6 Given 19–42% more training time and models only up to 7B, what additional result would justify calling PAPO a general multimodal-RL method?

Georgia Tech