RECAP: Looking Once Is Not Enough for Vision-Language Reasoning

Zhaolu Kang[1,2], Tailong Luo, Chenxin Li[2], Yingjie He[2], Zhenyu Yu, Fengyu Zhou, Jiachen Qian, Lei Wei[2], Shuang Chen, Jiachen Li, Eric Hanchen Jiang, Rongchao Zhang[2], Zhengtao Yao, Hoi Leong Lee, Guansu Wang, Kaiyue Zhou[2†]

[1]Peking University [2]Chengdu Minto Tech †Corresponding author

Paper
Abstract & paper

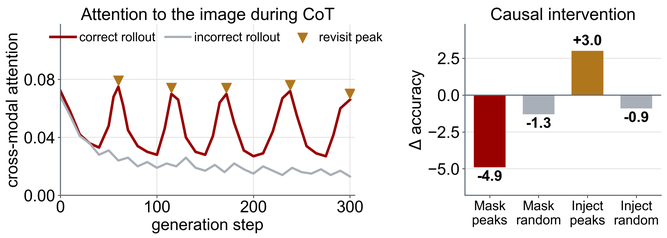
Email
zlkang25@stu.pku.edu.cn

TL;DR **Long-CoT VLMs drift from the image while RL rewards only the final answer. RECAP turns visual revisits into GRPO credit assignment and gains +3.5 to +4.1 HallusionBench points over GRPO on three backbones.**

1 Problem

- Long chain-of-thought helps VLMs, but as generation unfolds models can **drift from image evidence**.
- GRPO scores a rollout once, by its final answer, and gives **every token the same advantage**; it cannot tell which steps should have looked back at the image.
- Existing fixes reweight or re-inject visual signals statically; none asks **when** the image should re-enter the reasoning.

2 Evidence: visual revisits matter



Correct rollouts keep revisiting the image; masking revisit peaks hurts most, injecting them helps.

- **Predictive**: correct rollouts show recurring cross-modal attention peaks; failed rollouts lose or mistime them.
- **Causal**: masking peaks costs **4.9** points vs. 1.3 for random positions; injecting peaks into failed rollouts recovers 3.0.
- **Localized**: the top-8 attention heads carry at least 70% of the causal effect.

3 The signal matters

Signal	Qwen	Intern	LLaVA
Token entropy	63.2	61.6	60.1
Logit margin	63.0	61.4	59.8
Hidden norm	63.5	61.9	60.4
Revisit (ours)	**66.6**	**64.7**	**63.4**

HallusionBench; same RECAP framework, different token signal.

Heads	AUC	F1
Top-10 revisit	0.67	0.62
Random 10	0.54	0.50

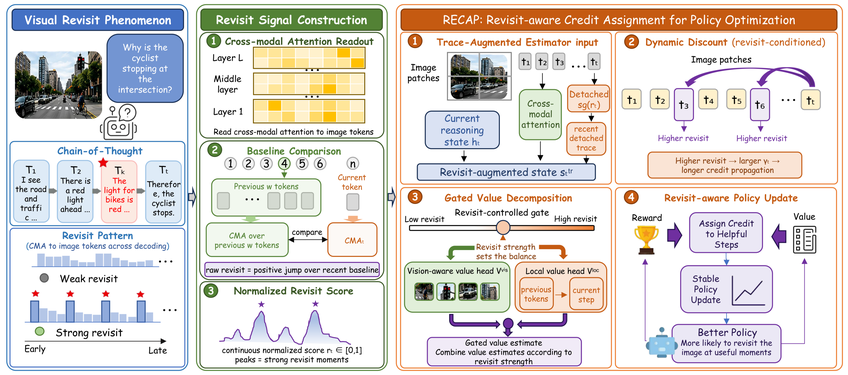
Linear probe.

- Weak correlation with other signals: ρ = 0.24 to 0.36.

Configuration	Qwen	Intern	LLaVA
SFT base	60.1	58.6	56.9
SFT + TVC (inference-time)	61.4	60.0	58.5
SFT + revisit re-injection	62.7	61.4	59.9
GRPO (RL, no revisit)	62.8	61.2	59.3
RECAP (RL + revisit)	**66.6**	**64.7**	**63.4**

HallusionBench. Revisit without RL helps a little; RL with revisit helps most.

4 RECAP: revisit-aware credit assignment



A detached revisit trace feeds three estimator slots; sampling, reward and inference stay unchanged.

$$z_t = \frac{\mathrm{ReLU}(\mathrm{CMA}_t - b_t)}{b_t + \epsilon}$$

b_t: mean CMA over the previous 16 tokens.

- **Revisit signal**: attention from each generated token to the image, averaged over middle-third layers and all heads, then squashed to [0, 1]. It is computed after sampling, so it changes how advantages are built, not how rollouts are drawn.
- **Trace-augmented input**: recent detached revisit traces enter advantage and value estimation.
- **Dynamic discount**: higher revisit raises the discount, so later revisit peaks retroactively raise credit for earlier decisions.
- **Gated value**: low-revisit tokens are often language-internal continuations, high-revisit tokens coincide with renewed visual evidence; two value heads are mixed by the revisit strength.
- **Drop-in for GRPO**: no extra rewards, rollouts or inference cost. **Theory**: the dynamic discount expresses retroactive credit that same-step reweighting cannot.

5 Twelve benchmarks (8 seeds; top rows Qwen2.5-VL-7B)

- **Training**: every method shares SFT on 50K long-CoT examples, then RL on 30K disjoint verifiable questions (group size 8, same verifier). RECAP adds no reward model; baseline hyperparameters are tuned, then frozen for RECAP.

Method	Hall.	POPE	MMHal	MathV	MMMU	MathVr	MMVet	GQA	BLINK	MMVP	MMLU	GSM8K
SFT base	60.1	86.5	3.12	59.4	41.3	33.2	51.4	63.8	48.2	54.0	72.4	78.9
TVC (no RL)	61.4	86.9	3.18	59.2	41.0	33.5	51.8	64.1	48.5	54.3	72.3	78.7
GRPO	62.8	87.1	3.27	61.2	42.5	34.7	52.8	64.3	49.1	54.8	72.1	78.5
VPPO	64.2	87.6	3.35	60.8	41.9	34.1	52.1	64.7	48.9	55.1	71.8	77.9
RECAP	**66.6**	**88.4**	**3.58**	**63.6**	**42.7**	**36.1**	**53.9**	**65.2**	**50.1**	**56.4**	**72.3**	**78.6**
Intern GRPO	61.2	87.8	3.34	63.1	45.6	36.2	54.2	65.6	50.2	54.0	73.4	79.6
Intern RECAP	**64.7**	**89.0**	**3.62**	**65.3**	**45.9**	**37.4**	**55.1**	**66.5**	**51.1**	**55.5**	**73.5**	**79.7**
LLaVA GRPO	59.3	86.1	3.19	58.0	40.9	32.8	50.8	63.0	47.6	52.9	69.7	74.9
LLaVA RECAP	**63.4**	**87.4**	**3.50**	**60.5**	**41.1**	**34.3**	**51.9**	**63.9**	**48.6**	**54.3**	**69.9**	**75.0**

Hallucination: HallusionBench, POPE, MMHal. Reasoning: MathVista, MMMU, MathVerse, MM-Vet. Grounding: GQA, BLINK, MMVP. Text-only: MMLU, GSM8K. Intern: InternVL3-8B; LLaVA: LLaVA-OV-7B.

6 Results on three backbones

Method	Qwen2.5	InternVL3	LLaVA-OV
SFT base	60.1	58.6	56.9
GRPO	62.8	61.2	59.3
VPPO	64.2	62.4	60.6
RECAP	**66.6**	**64.7**	**63.4**

HallusionBench, mean of 8 seeds; same SFT and RL data.

+3.5 to +4.1
over GRPO

+6.1 to +6.5
over SFT

- VPPO gains less, and the no-RL TVC reference stays below RECAP.

72B	Hall.	MathV
LoRA-GRPO	69.8	69.6
LoRA-VPPO	70.2	69.0
LoRA-RECAP	**72.9**	**71.5**

Qwen2.5-VL-72B with LoRA.

Chinese	GRPO	RECAP
CMMMU	46.3	49.1
MMMU	39.4	42.2
MathVista-CN	55.1	57.3

Chinese benchmarks, zero-shot; Qwen2.5-VL-7B.

7 What matters

Configuration	Hall.	MathV	Var(A_t)
Full RECAP	**66.6**	**63.6**	**1.00×**
− trace-augmented input	65.8	63.1	1.08×
− gated value	65.2	62.8	1.70×
− dynamic discount	64.3	62.4	1.09×
− stop-gradient on revisit	60.8	59.7	2.35×
GRPO (all three off)	62.8	61.2	1.81×
VPPO + RECAP	**68.0**	**63.9**	0.98×

Qwen2.5-VL-7B. Var(A_t): advantage variance relative to full RECAP.

Use of revisit	Qwen	Intern	LLaVA
advantage reweighting	64.6	63.1	61.3
dynamic discount	**66.6**	**64.7**	**63.4**

HallusionBench; same revisit signal.

- **Retroactive credit matters**: the discount beats same-step reweighting by 1.6 to 2.1 points.

Temperature η	Hall.
3	65.9
6 (default)	66.6
9	66.3
12	65.4

Window W	Hall.
8	65.6
16 (default)	66.6
32	66.2
64	65.8

- **Robust signal**: nearby settings stay within 1.2 points; the full grid spans 3.1, mostly from all-layer pooling.

8 Trust and robustness

Human study	GRPO	VPPO	RECAP
Groundedness win %	14.2	23.7	**62.1**
Coherence win %	29.4	30.1	**40.5**
Trust (1–5)	3.18	3.34	**3.87**

10 annotators, 100 samples.

- **Reformulated questions**: RECAP 66.0 ± 0.84 vs. GRPO 62.2 ± 2.81.
- **LLM judges** (GPT-4V, Claude, Qwen-Max): 4.08 vs. 3.21 for GRPO.

zlkang25@stu.pku.edu.cn