See More = Know More?

Mono-Anchored Advantage Normalization for Multi-Source Visual Reasoning

Fanhu Zeng[1], Zhicong Luo[2], Zefan Wang[1], You Li[3], Chi Chen[1*], Maosong Sun[1]

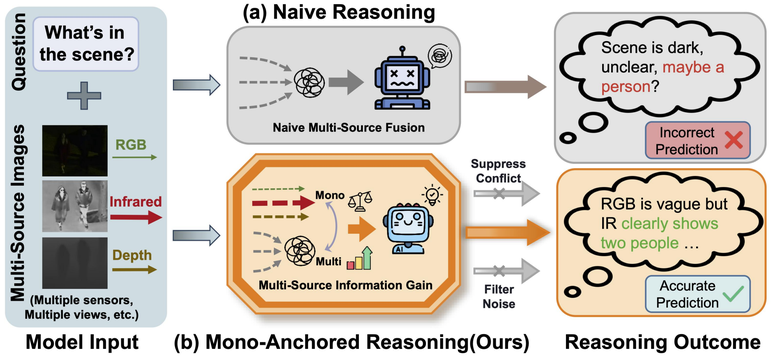
[1]Tsinghua University [2]Northwestern Polytechnical University [3]Beijing Jiaotong University

NEURAL INFORMATION PROCESSING SYSTEMS

TL;DR: An effective way for resolving conflict in multi-source visual reasoning

Background & Motivation

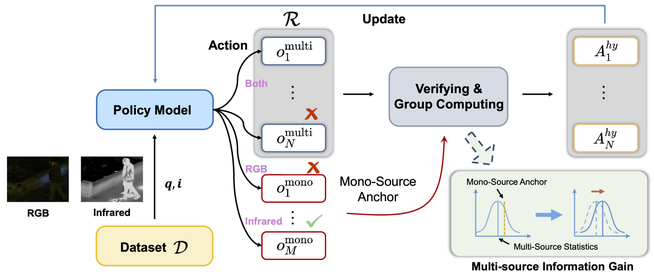
Illustration and comparison of multi-source visual reasoning



Naively incorporating more visual sources may bring conflicts with performance degradations

Method

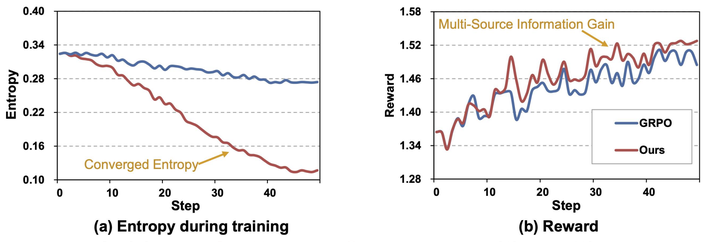
Incorporating mono-source rollouts into advantage estimation of multi-source rollouts



$$\mathcal{G}^{\text{multi}} = \{o_j^{\text{multi}}\}_{j=1}^{N} \qquad A_j = \frac{r(q,i,o_j^{\text{multi}}) - \mu^{\text{multi}}}{\sigma^{\text{multi}}}$$

$$\mathcal{G}^{\text{mono}} = \{o_j^{\text{mono}}\}_{j=1}^{M} \qquad A_j^{\text{hy}} = \frac{r(q,i,o_j^{\text{multi}}) - \mu^{\text{hy}}}{\sigma^{\text{hy}}}$$

Mono-source reasoning acts as a general **dynamic anchor** to stabilize and guide multi-source reinforcement learning in **both promotion and conflict scenarios**



Achieves **better performance** with multi-source information gain while **maintaining stability**

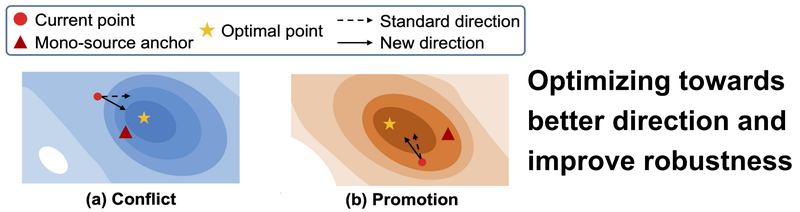
Theoretical Analysis

Theorem 1 (**Unbiasedness**). *For any measurement of gradient estimation, the expectation of our algorithm is equivalent to the expectation of on-policy optimization.*

$$\mathbb{E}_{q,i\sim\mathcal{D},\ o\sim\pi_\theta(q,i)}[A^{hy}\cdot\nabla_\theta\log\pi_\theta(\cdot|q,i)] = \nabla_\theta J(\theta)$$

Theorem 2 (**Gradient Decomposition**). *The gradient optimization based on mono-anchored advantage norm is equivalent to maximizing the multi-source information gain while optimizing the standard multi-source reward.*

$$\nabla_\theta J^{hy}(\theta) = \nabla_\theta J(\theta) + (1-\alpha)\Delta_{IG}\nabla_\theta J^{reg}(\theta)$$



Optimizing towards better direction and improve robustness

Experiment

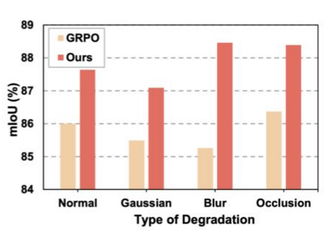
Model	Infrared		Depth		Multi-View		Text-Rich		Avg.	
	Union	Multi	Union	Multi	Union	Multi	Union	Multi	Union	Multi
Previous Methods										
Qwen2.5-VL-3B [3]	63.0	38.8	86.3	73.8	82.0	74.0	69.0	51.0	75.1	59.4
LLaVA-OV-7B [16]	4.9	2.1	80.0	56.3	88.0	44.0	81.0	67.0	63.5	42.3
R1-Onevision-7B [31]	19.7	11.1	87.5	72.5	90.0	81.0	77.0	58.0	68.5	55.6
VisionReasoner-7B [20]	82.5	75.2	85.0	78.8	85.0	70.0	62.0	45.0	78.6	67.2
Supervised Post-Training										
SFT	87.1	84.5	85.0	65.0	90.0	78.0	83.0	78.0	86.2	76.4
Chain-of-Thought	88.2	84.2	90.0	77.5	90.0	75.0	77.0	76.0	86.3	78.2
Reinforcement Post-Training										
GRPO [22]	89.7	86.0	80.0	70.0	88.0	72.0	77.0	71.0	83.6	74.7
GRPO+MARS	89.9	87.7	80.0	73.8	90.0	75.0	**79.0**	**75.0**	**84.7 (+1.1)**	**77.9 (+3.2)**
DAPO [32]	88.8	85.5	76.3	73.8	**92.0**	74.0	75.0	70.0	83.0	75.8
DAPO+MARS	**92.3**	**89.4**	**82.5**	**77.5**	88.0	**81.0**	78.0	**75.0**	**85.2 (+2.2)**	**80.7 (+4.9)**

Effective across various multi-source visual reasoning **datasets and RLVR strategies**

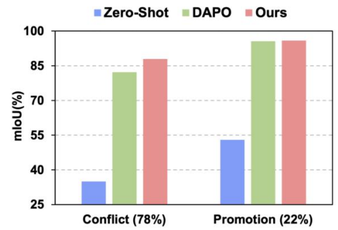
Method	Single-Image			Multi-Image			Avg.
	MMStar	HalluBench	TextVQA	Mantis	Muirbench	Blink	
Qwen2.5-VL-3B	54.5	**60.6**	**79.1**	59.4	**49.0**	53.8	59.4
GRPO	55.2	57.6	75.5	57.1	47.3	81.7	62.4
MARS	**55.9**	58.4	76.0	**62.7**	48.9	**82.5**	**64.1 (+1.7)**

Improve performance **general multimodal benchmarks**

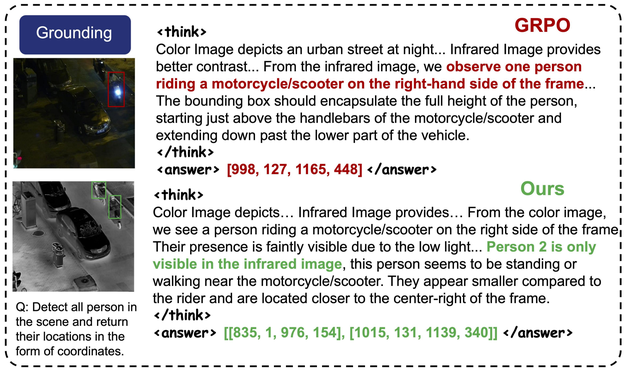
Further Analysis



Robustness to different visual degradations



Gain across modality conflict and promotion



Qualitative results of visual reasoning

Conclusion

- We identify that naively integrating multiple sources where specific source contains dominant signal can degrade performance.
- We design a novel visual reasoning method to measure multi-source information gain, enabling adaptive regulation of sources in RLVR training.

Paper Dataset Code Rednote