Does Seeing More Mean Knowing 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

Abstract

Visual reasoning through reinforcement learning with verifiable rewards (RLVR) has achieved remarkable progress. However, when dealing with multi-source inputs, existing approaches tend to treat them as a mere accumulation of information, lacking explicit mechanisms to distinguish whether integrating additional sources yields information gain or introduces interference. Therefore, they struggle to effectively model dynamic interaction when integrating multiple sources, particularly when they differ significantly in physical properties and semantics, *e.g.*, infrared and depth, leading to inferior performance to mono-source reasoning when a certain source holds the dominant signal. To address this issue, we propose MARS, a novel mono-anchored multi-source reasoning framework that models each visual modality as an independent information source. Specifically, by treating mono-source rewards as dynamic anchors, our method explicitly incorporates the information gain introduced by multi-source fusion into advantage normalization and adaptively emphasizes mutual promotion between sources while suppressing potential noise or conflicts during RLVR. From theoretical analysis, our method effectively quantifies information gain introduced by multi-source integration in gradient estimation, enabling consistent modality regulation. Empirical results also show impressive 3.2% and 4.9% performance gains on GRPO and DAPO across diverse datasets, confirming the effectiveness of our method. Code is available at https://github.com/AI9Stars/MARS.

1 Introduction

Recent advances in multimodal large language models (MLLMs), which align representations across vision and language modalities [2], have demonstrated strong capabilities in multimodal perception and understanding [19]. More recently, visual reasoning [18, 29] has been introduced to encourage deeper thinking through reinforcement learning with verifiable rewards (RLVR), allowing models to generate structured responses with self-reflection through explicit reasoning rather than direct prediction, thereby fostering the emergence of chain-of-thought (CoT) reasoning [27] and enhancing the ability of complex understanding, multi-step reasoning, and logical consistency.

Despite the progress of visual reasoning, current methods largely optimize for aligned representations, and the complementary strengths of different sources are often assumed and overutilized, *i.e.*, seeing more modalities means knowing more, but potential interference or conflicts are seldom explicitly explored. In particular, existing RLVR frameworks optimize multi-source rewards directly, without explicitly assessing whether integrating additional sources yields positive information gain or instead introduces interference relative to strong mono-source reasoning, especially when their attributes

*Correspondence to: Chi Chen <chenchithu@gmail.com>

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and semantics have significant differences, such as medical imaging [1], autonomous driving [7], remote sensing [36], and so on. In these scenarios, ***naively integrating multiple sources can even lead to performance inferior to strong mono-source reasoning***, when a specific source contains the dominant and reliable signal. Therefore, there is a lack of principled mechanisms to model dynamic interaction, whether complementary or conflicting, between multiple sources, which may become particularly evident in complex scenarios where a critical cue resides in a specific source. As shown in Figure 1, in tasks where inherent physical limitations and degradation are caused by illumination variation, occlusion, and adverse weather conditions, relying solely on RGB imagery or the relationship between sources is often inadequate. In contrast, different sources such as infrared, depth, or multi-view can provide crucial and robust information with more reliable scene understanding, which requires handling multi-source data in a comprehensive manner.

In this paper, we aim to enhance the ability of visual reasoning when dealing with multi-source data. Based on the analysis, we uncover that a core reason for this limitation lies in the naive way current visual reasoning frameworks handle source integration. Specifically, they treat all inputs equally contributive to the task, and fail to explicitly model the performance interactions between specific source and multi-source data, which is a key factor for effective multi-source regulation. From an optimization perspective, these interactions correspond to whether multi-source reasoning improves or degrades performance relative to mono-source baselines, a distinction that remains invisible to advantage estimation in existing RLVR frameworks. This gap motivates the need for a general approach that can dynamically regulate contributions from certain sources.

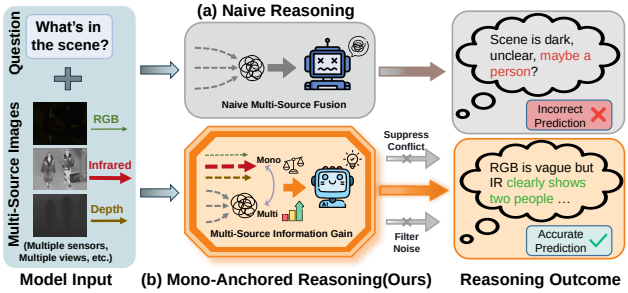


Figure 1: Illustration of multi-source visual reasoning. (a) Existing methods struggle to model dynamic interaction in multi-source data; (b) our method explicitly uses mono-source rewards as anchors to measure the information gain from multi-source integration and enhances reasoning.

To this end, we propose MARS, a novel multi-source reasoning framework that explicitly incorporates each visual modality as an individual information source and models the information gain introduced by multi-source integration. Concretely, by treating mono-source rewards as the anchors, it computes advantages based on information gains between multi-source and mono-source[2] rewards. We theoretically analyze that our method guarantees and enables dynamic optimization that emphasizes promotion while suppressing noisy or conflicting information during training by maximizing multi-source information gain. Notably, our algorithm enhances multi-source utilization from inherent capability without architectural redesign, offering a general and effective solution for improving visual reasoning performance.

We conduct experiments on various multi-source tasks, including depth, infrared, multi-view and text-rich understanding. Extensive results with notable **3.2%** and **4.9%** improvements on GRPO and DAPO and in-depth analyses strongly validate the effectiveness and generalizability of our method. Our contributions are summarized as follows:

- We reveal that existing multi-source visual reasoning can systematically degrade performance, and identify relative information gain over mono-source reasoning as the key factor for effective multi-source integration from theoretical derivation.
- We design a novel visual reasoning method that introduces mono-source rewards as anchors to quantitatively measure multi-source information gain from integration in advantage normalization, enabling adaptive regulation of different sources during RLVR training.
- We conduct extensive experiments on various multi-source visual reasoning tasks, and the consistent and significant performance improvements on different RL algorithms validate the effectiveness and generality of our approach.

[2]They specifically describe features of different visual modalities that differ in physical properties and semantics.

2 Related Work

Reinforcement Learning with Verifiable Rewards has made substantial progress in recent years, with pioneering systems such as DeepSeek-R1 [12] and Kimi [25] demonstrating that complex reasoning patterns can emerge through optimization with verifiable rewards, where outcome reward signals are used to guide the learning of long reasoning chains. Within this paradigm, some approaches focus on enhanced optimization strategies [39, 38], such as regularization, stabilized policy updates, and refined reward designs, to improve consistency and robustness. Building on these foundations, visual reasoning incorporates images into reasoning by coordinating linguistic reasoning with perceptual states. It achieves strong performance in vision-centric tasks such as grounding [4] and image understanding [31], highlighting it as a promising paradigm for complex multimodal understanding and deduction. In this paper, we focus on the capability of visual reasoning with multi-source data.

Multi-Source Visual Reasoning refers to tasks that require a joint understanding of images from multiple sources, potentially captured from different sensors, times or viewpoints [37]. This is crucial for real-world intelligent systems, where a single source is often insufficient for achieving completeness and reliable decisions in complex environments. Early studies focus on multi-source fusion [6, 34], where extracted features from different cameras are explicitly fused to enhance robustness and geometric consistency. More recently, multimodal large language models have reframed multi-image reasoning as unified and aligned comprehension with implicit correspondences and shared representations. Nevertheless, current attention is paid to general domain enhancement and evaluation [10, 33], which overlooks the complementarity and contradictions of multi-source data in the reasoning process. In practical applications, it is crucial for models to determine the reliability of multiple visual inputs, rather than relying on manually selecting the most salient modality, which is valuable in critical fields such as medical diagnosis [13] and robotic manipulation [21].

3 Methodology

3.1 Motivation

Visual reasoning has exhibited strong understanding capabilities under multi-image inputs. However, we observe a consistent and non-trivial phenomenon: as depicted in Figure 1, when handling images from multiple sources, *e.g.*, infrared, depth and so on, and only one image among multiple sources is truly informative for the task, typical multi-source reasoning often fails to capture and concentrate on the critical visual scene and therefore underperforms the upper bound of mono-source reasoning, even if all available sources are provided jointly. This contradicts the cognition of humans that integrating more information always brings more knowledge, and naturally raises an open question:

Does seeing more mean knowing more in multi-source visual reasoning? If not, how can we solve it?

We attribute the issue to potential modality interference in multi-source reasoning. Specifically, a typical visual reasoning model is normally trained under the assumption of complementary data integration, *i.e.*, seeing more images brings more knowledge, and only learns the positive guidance of multi-image fusion with implicitly unified and shared representations. Without explicitly identifying which image is causally responsible for correct decisions, it therefore struggles to capture the dynamic interaction, *i.e.*, promotion or interference between modalities. This is especially severe in multi-source scenarios where images have different properties and semantics, resulting in unstable or noisy learning dynamics. Therefore, advantage estimation becomes unreliable under such conflict, where standard advantage normalization estimates statistics solely from multi-source trajectories, and may be dominated by spurious correlations introduced by non-informative sources.

At this point, specific mono-source reasoning often provides a significant and stable inductive signal in these scenarios. When the key image is present, a specific mono-source rollout tends to produce more consistent reward with semantic information, effectively guiding the optimization direction.

To this end, we propose to incorporate mono-source rollouts into the advantage estimation of multi-source rollouts. Intuitively, mono-source reasoning acts as a general dynamic anchor to stabilize and guide multi-source reinforcement learning: (1) if it underperforms with modality conflicts, the algorithm softly regularizes trajectory updates toward the more reliable mono-source behavior; (2) moreover, if multi-source reasoning outperforms mono-source reasoning with modality mutual promotion, the algorithm also encourages exploration beyond mono-source cues.

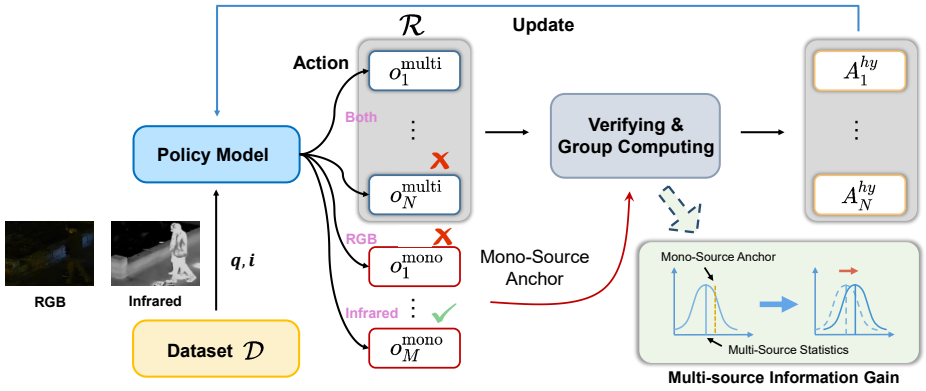


Figure 2: Structure of the proposed mono-anchored advantage normalization for multi-source visual reasoning. Mono-source rewards serve as dynamic anchors to quantify the influence of source integration with multi-source information gain in on-policy optimization.

3.2 MARS: Mono-Anchored Advantage Normalization for Multi-Source Reasoning

Preliminary. For each instance consisting of question q and multiple images i in training dataset $\mathcal{D}$, multi-source trajectories (rollouts) are generated through policy π_θ parameterized with θ, where multiple images are jointly provided as input:

$$\mathcal{G}^{\text{multi}} = \{o_j^{\text{multi}}\}_{j=1}^{N}. \tag{1}$$

The reward r is exploited to measure the output in response to input and each rollout is normalized by group-wise mean and variance of reward to obtain advantage for stability:

$$A_j = \frac{r(q, i, o_j^{\text{multi}}) - \mu^{\text{multi}}}{\sigma^{\text{multi}}}. \tag{2}$$

The standard policy gradient algorithm optimizes the expected advantage function $J(\theta)$, and its policy gradient estimator [24] has the following form:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\{q,i\}\sim\mathcal{D},\, o\sim\pi_\theta(q,i)}[A \cdot \nabla_\theta \log \pi_\theta(o|q,i)], \tag{3}$$

where $\{q, i\}$ is the question and image from dataset $\mathcal{D}$, and the policy π_θ generate the trajectories for verifiable reward.

Advantage Normalization with Mono-Source Anchor. As illustrated in Figure 2, in terms of reasoning with multi-source visual tasks, motivated by the function of mono-source rewards in advantage estimation, we additionally generate mono-source rollouts, where each image is individually paired with the textual input to produce rewards with the same policy model:

$$\mathcal{G}^{\text{mono}} = \{o_j^{\text{mono}}\}_{j=1}^{M}. \tag{4}$$

In terms of advantage estimation, it is performed for multi-source rollouts only, while leveraging hybrid statistics of both multi and mono-source rewards for gradient estimation to stabilize the normalization:

$$A_j^{\text{hy}} = \frac{r(q, i, o_j^{\text{multi}}) - \mu^{\text{hy}}}{\sigma^{\text{hy}}}, j = 1, \cdots, N. \tag{5}$$

Specifically, mono-source rollouts are not used to directly update the multi-source policy. Instead, their role is to adjust the normalization statistics as an adaptive reference. Intuitively, as illustrated in the right of Figure 2, when multi-source reward outperforms mono-source reward, the introduced estimation would lower the mean for multi-source enhancement. Conversely, if a particular modality plays a decisive role, our algorithm will inhibit the model from learning contradictory multi-source rewards and instead drive it toward better modality-specific learning.

Verifiable Rewards. The verifiable reward is a key component in reinforcement learning to align the preferences of models, which may include simple verification functions [22] that check whether predictions match the correct answers in contents and formats. Applying this concept to visual

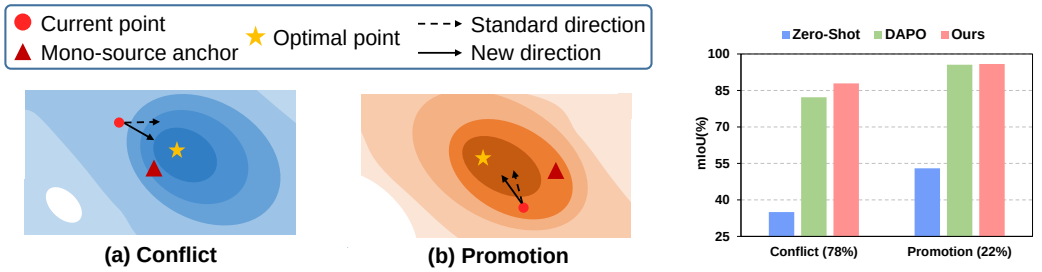


Figure 3: Illustration of mono-source anchor in optimization. It provides a consistent facilitation in both modality conflict and promotion. This results in significant gain across scenarios.

tasks requires adaptation of specific rule-based verifiable reward functions. **For grounding tasks**, grounding reward is directly formulated by calculating the average Intersection-over-Union (IoU) between predicted and ground truth bounding boxes:

$$r_{\text{iou}}(q, i, o) = \frac{1}{K} \sum_{i=1}^{K} IoU_i, \tag{6}$$

where K is the number of objects in the scene, and the grounding reward consists of the iou reward and the format reward:

$$r_{\text{grounding}} = r_{\text{iou}} + r_{\text{format}}. \tag{7}$$

For visual question answering tasks, the accuracy reward is determined by whether the output matches the ground truth:

$$r_{\text{acc}}(q, i, o) = \begin{cases} 1 & \text{, if ground truth in } o, \\ 0 & \text{, otherwise.} \end{cases} \tag{8}$$

The final reward is a combination of accuracy reward and format reward:

$$r_{\text{vqa}} = r_{\text{acc}} + r_{\text{format}}. \tag{9}$$

Remark. In the case of multi-source visual reasoning, instead of estimating the baseline from multi-source rewards alone, our algorithm computes hybrid statistics over the union of mono-source and multi-source rewards:

$$\mathcal{R} = \left\{ o_1^{\text{mono}}, \ldots, o_M^{\text{mono}}, o_1^{\text{multi}}, \ldots, o_N^{\text{multi}} \right\}, \tag{10}$$

from the same policy model. This can be seen as on-policy optimization with a hybrid distribution. Concretely, by leveraging mono-source rewards as anchors, our method precisely utilizes the difference between trajectories, thereby enhancing performance with exact information gain from mono-source to multi-source rewards shown in Figure 4b.

3.3 Theoretical Analysis

We present several theoretical analyses of the proposed mono-anchored advantage normalization algorithm from policy optimization perspective to construct key properties for stability and rationality.

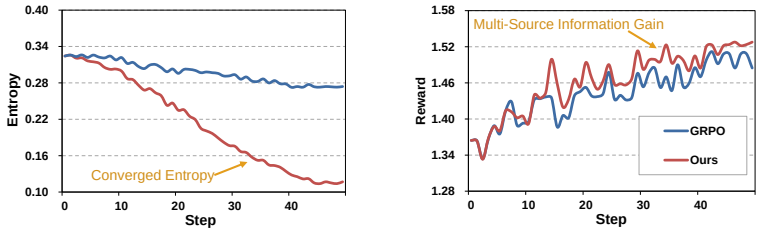
Theorem 3.1 (Unbiasedness)**.** *For any measurement of gradient estimation, the expectation of our algorithm for is equivalent to the expectation of on-policy optimization:*

$$\mathbb{E}_{q,i\sim\mathcal{D},\, o\sim\pi_\theta(q,i)}[A^{hy} \cdot \nabla_\theta \log \pi_\theta(\cdot|q,i)] = \nabla_\theta J(\theta). \tag{11}$$

This provides a theoretical guarantee that the proposed algorithm introduces no bias for gradient estimation under the condition of on-policy optimization from the perspective of expectation, ensuring stability. Appendix A.2 provides the detailed proof.

Theorem 3.2 (Gradient Decomposition)**.** *The gradient optimization based on mono-anchored advantage normalization is equivalent to maximizing the multi-source information gain while optimizing the standard multi-source reward:*

$$\nabla_\theta J^{hy}(\theta) = \nabla_\theta J(\theta) + (1-\alpha)\Delta_{IG}\nabla_\theta J^{reg}(\theta), \tag{12}$$



(a) Entropy during training (b) Reward

Figure 4: Learning statistics of (a) entropy and (b) reward when training vanilla GRPO and our algorithm. Our method achieves better performance with multi-source information gain from a mono-source anchor while maintaining stability.

where conventional advantages have a zero mean. $\alpha = N/(M+N)$ is the proportion of multi-source rollouts representing the strength of guidance. $\Delta_{IG} = \mu^{\text{multi}} - \mu^{\text{mono}}$ is the expectation of the reward increment of multi-source trajectories compared to mono-source trajectories. It measures the relative information gain from multiple image fusion relative to mono-source reasoning, and a negative value indicates conflict between modalities that perform inferior to mono-source reasoning.

The derivation holds for any $\alpha \in [0, 1]$ and justifies the practical effectiveness of the advantage estimation scheme combining both sample types through unified standardization. This reveals that the practical effect of MARS arises from the finite-sample gradient estimator used by RL algorithm that our algorithm is theoretically optimizing a weighted multi-source reward with a multi-source information gain regularization, which emphasizes trajectories consistent with reliable sources, while suppressing those affected by cross-source interference, thereby reweighting the sampled gradients. Specifically, by leveraging the mono-source reward as an explicit and systematic signal, it dynamically adjusts the standard gradient direction according to multi-source information gain during the optimization procedure, which guides the optimization towards an optimal point with faster convergence and better multi-source performance. Figure 3 gives an intuitive illustration of mono-source anchor in optimization. In conflict where certain source performs well, it is around the optimal point, and pull the optimization direction close by information gain. This greatly improves the performance, where standard multi-source reasoning undergoes a severe performance drop. The process is similar in promotion that pushes the direction away from mono-source anchor towards the optimal point, providing a consistent facilitation in multi-source visual reasoning. The quantitative results show the commonalities of conflict and the utility of our method in resolving conflict rather than a general effect. The proof is shown in Appendix A.3.

Remark. Inherently, we study policy optimization with a mono-anchored advantage for multi-image reasoning. Instead of universally increasing rewards or advantages, it enforces a principled cross-modal comparison: multi-source rollouts receive positive updates if and only if they outperform mono-source reasoning. This is consistent with the motivation that multi-source reasoning improves when it provides complementary information, while being regularized otherwise. Therefore, MARS reduces the blind exploration of multi-source policies, accelerates convergence, and improves robustness against possible visual inconsistency, yielding a more stable and interpretable optimization trajectory for multi-source reasoning models as shown in Figure 4.

Simplicity and Stability. Our algorithm only requires one policy model for optimization and does not introduce additional storage for models or samples, as opposed to experience sampling [35] or off-policy correction [30] methods. In practical implementation, we generate the mono-source samples by modifying image inputs and obtain the normalized statistics without computing gradients, thereby maintaining algorithmic efficiency and stability as shown in Table 3.

4 Experiments

4.1 Experimental Setup

Datasets. Regarding the datasets, we employ diverse multi-source datasets. For visual modalities, in addition to typical RGB images, we incorporate four different modalities, including depth SpatialQA [8], infrared LLVIP [14], multi-view nuScenes [5] and text-rich OCR-VQA [17]. Detailed explanations are shown in Appendix B.2.

Table 1: Overall performance of four different multi-source tasks based on Qwen2.5-VL-3B. We compare with previous methods, supervised and reinforcement post-training methods, respectively. Union denotes the best performance among all mono-source results. The best results within a comparable group are in **Bold**.

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)

Baselines. For previous algorithms, we compare with various methods with [31, 20] and without [16, 3] reinforcement post-training. In addition, for supervised post-training, SFT and CoT are also incorporated for comprehensive comparison. Regarding reinforcement post-training, we employ GRPO [22] and DAPO [32], which are two typical group-based reinforcement learning algorithms for visual reasoning. Since our method does not rely on a specific training framework, unless otherwise stated, all comparative experimental results are conducted within the same basic structure.

Implementation details. We use Qwen2.5-VL-3B [3] as the base model for supervised and reinforcement post-training. We mainly conduct the experiments on visual question answering (VQA) and grounding, and the evaluation metrics are accuracy and mIoU, respectively. The concrete calculation is similar to that for verifiable rewards. For a comprehensive understanding, in addition to standard multi-source visual reasoning that jointly take all images as inputs (**Multi**), we furthermore perform mono-source reasoning (**Union**), *i.e.* reason with each single source and then obtain the best result as final performance, as the upper bound to showcase the utility of information gain in performance. Concretely, we separately take every single source as input for visual reasoning during inference, and consider it to be correct if any single source correctly answers. We uniformly generate one trajectory for each visual source, *i.e.*, $M = \texttt{image_num}$ and $N = 12$. More implementation details can be found in Appendix B.1.

4.2 Main Results

We mainly conduct the experiment on multi-source datasets. Furthermore, we extend our algorithm to the application scenarios, the reinforcement post-training strategies and the scale of the model to comprehensively validate the effectiveness of the proposed method. More results are in Appendix C.

MARS is effective across various multi-source visual reasoning datasets. We evaluate the proposed advantage algorithm on diverse visual reasoning tasks with multi-source datasets. It can be seen in Table 1 that: **(1)** the performance of Union exhibits better performance than Multi by a substantial margin across all tasks, which can be seen as the multi-source information gain, showing great potential to utilize for performance enhancement in multi-source tasks; **(2)** higher Union does not mean better performance in Multi, especially in supervised post-training methods, including SFT and CoT and naive RLVR methods. This strongly demonstrates the dynamic interaction between different sources is not well modeled by existing methods, leading to performance inferior to mono-source upper bound, which is consistent with our topic that *seeing more does not mean knowing more in multi-source visual reasoning*; **(3)** compared to supervised post-training, better reinforcement learning strategy does not necessarily achieve better performance for Multi, revealing the difficulty

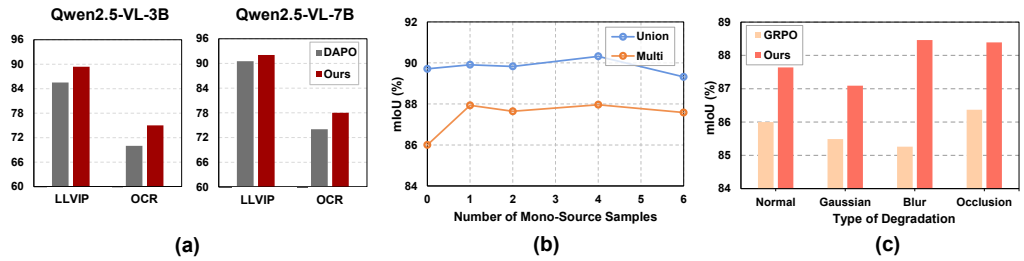


Figure 5: (a) Performance across different model sizes. (b) Influence of different numbers of mono-source samples in grounding. (c) Performance comparison under different visual degradations.

of the task and highlighting the necessity for designed algorithms; **(4)** in contrast to standard visual reasoning that even achieves no gains in some scenarios, our method consistently outperforms baselines and obtains substantial improvements on multi-source visual reasoning. Specifically, on the most important multi-source reasoning metric, Multi, our approach significantly boosts all tasks, with notable **3.8%** and **7.0%** improvements on infrared and multi-view datasets, respectively. Also, our method further facilitates the upper bound of Union, bringing in coherent improvements. This confirms that dynamically leveraging mono-source rewards as anchors effectively guides the model to focus on informative sources while suppressing noise from less relevant images, and therefore achieves an impressive performance enhancement, firmly validating the effectiveness of our method.

Different training strategies. Also, we validate the versatility of our approach by integrating it into two distinct reinforcement learning frameworks, *i.e.*, GRPO and DAPO. As indicated in bottom of Table 1, our method obtains substantial improvements under both settings. Concretely, our method achieves an average performance gain of **3.2%** in GRPO and **4.9%** in DAPO on visual reasoning with various vision sources. Moreover, our algorithm also boosts the performance of mono-source reasoning by **1.1%** and **2.2%**, respectively. It highlights that our method is independent of the reinforcement learning framework, and only requires additional mono-source reward signals during advantage estimation without modifying the core training objectives, which shows great generalization ability and strong promise to serve as a plug-and-play integration to enhance multi-source performance in different training strategies.

Generalizability across different model sizes. To further assess the scalability, we validate our method with model sizes comprising 3B and 7B parameters, respectively. As shown in Figure 5a, we can conclude that (1) larger models benefit from larger knowledge capacity and typically achieve a better performance; (2) the proposed advantage algorithm brings consistent and substantial improvement across scales, *e.g.*, a 4.4% gain for small model and 2.8% for large model on average. This shows that the stable performance promotion stems from the inherent adaptability of the method by advantage estimation, making it agnostic to variation of model size, which suggests that the approach is not limited by model capacity and generalizes effectively.

4.3 Ablation Study and Further Analysis

We conduct extensive ablation studies and in-depth analyses to validate the effectiveness of our method and provide insight into the contribution of the core components under different conditions.

Table 2: Detailed statistics of reward in different algorithms.

Statistics	Qwen2.5-VL	GRPO	Ours
$\max(r(o^{\text{multi}}))$ (↑)	1.67	1.69	**1.68**
$\text{mean}(r(o^{\text{multi}}))$ (↑)	1.49	1.51	**1.62**
$\max(r(o^{\text{mono}}))$ (↑)	1.55	1.56	**1.63**

Table 3: performance and efficiency balance with different numbers of trajectory generation.

Number of Trajectory	N	$M+N$	Ours
GPU Hours	13×8	17×8	14×8
mIoU (%)	86.0	86.2	87.7

Incorporating more mono-source samples. In the main results, we take one mono-source reward from every single source as the dynamic anchor. We investigate how the number of available mono-source samples, *i.e.*, M, influences overall performance as the number of mono-source anchors employed in advantage normalization indicates the strength of the multi-source information gain towards robust visual reasoning. Results are shown in Figure 5b, where $M = 0$ represents the vanilla



Figure 6: Qualitative results of visual reasoning on multi-source datasets. In grounding with RGB and infrared images and VQA with RGB and depth images, GRPO would excessively rely on the RGB image, resulting in improper predictions. Our method benefits from multi-source information gain and is capable of adaptively focusing on the key images with correct responses.

algorithm. It reveals that visual reasoning accuracy saturates as M increases and additional samples yield diminishing returns. It implies that a small set of mono-source trajectories is adequate to provide diverse information gain. Conversely, a larger M does not necessarily bring gains, as mono-source samples may dominate the advantage estimation with unstable direction, leading to higher dynamics against multi-source reasoning in policy updates. Thus, we employ $M = 2$, which is the number of sources, for stable training. More importantly, guiding multi-source learning without requiring exhaustive sampling brings negligible computational or storage overhead, showing the efficiency.

Impact of reward statistics. We compare different statistics of reward in Table 2 to investigate the changes in intuitive indicators. The results reveal that instead of increasing the best trajectory (maximum multi-source reward), MARS significantly improves the average quality of multi-source reasoning (1.49 to 1.62), approaching the upper bound of mono-source reward. This shows that our method indeed exploits multi-source information gain to guide the optimization towards a better direction rather than stochastic exploration with diversity, which demonstrates the effectiveness of mono-source anchor in modeling the dynamic interaction. Also, the simultaneous improvement in mono-source trajectory (1.55 to 1.63) during training uncovers the efficiency of the on-policy generation strategy without introducing additional model overload.

Performance-efficiency trade-off. Our algorithm requires additional trajectory generation in practical implementation during reinforcement post-training. To certify the key component in effectiveness, we additionally perform GRPO with rollout to be $M + N$ and compare the training efficiency in Table 3. The results show that our method, which includes only N samples in advantage normalization, introduces moderate training overhead. By contrast, incorporating more actual trajectory in policy update brings marginal improvement (0.2%) with substantial overhead (30%). This strongly suggests that the key to this approach is the mono-source anchor rather than introducing more rollouts, showcasing both effectiveness and efficiency.

Robustness to visual degradation. As our method dynamically integrates interaction between modalities, and concentrates on a specific source if another displays no information gain, it shows great promise in exhibiting robustness to possible visual degradations of certain sources, *i.e.*, noise, illumination, occlusion and so on. To simulate real-world conditions where visual inputs may be corrupted, we randomly degrade image quality by adding Gaussian noise, motion blur, and occlusion to the input images. As is illustrated in Figure 5c, our method maintains superior performance under all degradation types, *e.g.*, only **0.5%** drop in accuracy under severe Gaussian noise. Moreover, in the case of motion blur and occlusion, our method even obtains **0.8%** performance gain, which outperforms the baseline by a substantial **1.3%** improvement, showcasing its strong robustness and effectiveness. This can be attributed to our way of advantage normalization, which explicitly reduces the influence from unreliable source and relies more on stable, informative modality, thereby exhibiting adaptability and generalization ability.

Case Study. We visualize the reasoning process on several representative scenarios in Figure 6 to qualitatively validate the effectiveness of our method. Specifically, in grounding task, our method successfully prioritizes key information in infrared image and detects the specific person that is hard to see in RGB, yielding more precise bounding boxes than GRPO. Moreover, in VQA task with RGB and depth images, while vanilla GRPO is uncertain about the answer without clearly analyzing depth

image, our approach is confident about the inference relying solely on the RGB image and generates correct responses accordingly. The intuitive visualizations clearly show how our algorithm guides the model to weigh different sources appropriately, showcasing the effectiveness and superior capability.

5 Conclusion

In this paper, we revisit visual reasoning from a multi-source perspective, where visual modalities differing significantly in physical properties and semantics exhibit both promotion and conflict characteristics that are insufficiently modeled by existing methods. To address this issue, we propose a multi-source visual reasoning framework that adaptively emphasizes informative interaction while suppressing conflicting ones using mono-source rewards as dynamic anchors under RLVR training. This theoretically formulates the multi-source information gain of integration to guide the optimization towards stable and advanced trajectories. Comprehensive experiments across diverse datasets and training strategies with significant and consistent performance improvements strongly validate the effectiveness, efficiency and generalizability of the proposed approach.

Acknowledgments and Disclosure of Funding

The work is initiated and supported by AI9Stars Team.

References

[1] Muhammad Adeel Azam, Khan Bahadar Khan, Sana Salahuddin, Eid Rehman, Sajid Ali Khan, Muhammad Attique Khan, Seifedine Kadry, and Amir H Gandomi. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers in biology and medicine*, 144:105253, 2022.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025.

[5] Ankan Bansal, Yuting Zhang, and Rama Chellappa. Visual question answering on image sets. In *European Conference on Computer Vision*, pages 51–67. Springer, 2020.

[6] Martin Brenner, Napoleon H Reyes, Teo Susnjak, and Andre LC Barczak. Rgb-d and thermal sensor fusion: A systematic literature review. *IEEE Access*, 11:82410–82442, 2023.

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[8] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498. IEEE, 2025.

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[10] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[11] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385, 2024.

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[13] Heyu Huang, Wanran Sun, Chi Chen, Bo Chen, Zonghao Guo, Yuhua Li, Ruixuan Li, Kunlun He, and Maosong Sun. Echomllm: Incentivizing echocardiographic video understanding with keyframe grounding and report generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 20053–20071, 2026.

[14] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021.

[15] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024.

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[17] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[18] You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. Imagination helps visual reasoning, but not yet in latent space. *arXiv preprint arXiv:2602.22766*, 2026.

[19] You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, et al. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. *arXiv preprint arXiv:2501.05767*, 2025.

[20] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025.

[21] Fei Ni, Zhuo Chen, Yifu Yuan, Zibin Dong, Xianze Yao, Shan Luo, Jianye Hao, Jiankang Deng, and Stefanos Zafeiriou. Semanticvla: Towards semantic reasoning over action memorization via synergistic explicit trace and latent action planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12237–12247, 2026.

[22] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[23] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[24] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[25] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[26] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*, 2025.

[27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[28] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[29] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098, 2025.

[30] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *Advances in Neural Information Processing Systems*, 38:117157–117186, 2025.

[31] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[32] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[33] Youngjoon Yu, Sangyun Chung, Byung-Kwan Lee, and Yong Man Ro. Spark: Multi-vision sensor perception and reasoning benchmark for large-scale vision-language models. *arXiv preprint arXiv:2408.12114*, 2024.

[34] Maoxun Yuan, Xiaorong Shi, Nan Wang, Yinyan Wang, and Xingxing Wei. Improving rgb-infrared object detection with cascade alignment-guided transformer. *Information Fusion*, 105:102246, 2024.

[35] Runzhe Zhan, Yafu Li, Zhi Wang, Xiaoye Qu, Dongrui Liu, Jing Shao, Derek Wong, and Yu Cheng. Exgrpo: Learning to reason from experience. In *International Conference on Learning Representations*, volume 2026, pages 12810–12846, 2026.

[36] Jixian Zhang. Multi-source remote sensing data fusion: status and trends. *International journal of image and data fusion*, 1(1):5–24, 2010.

[37] Lili Zhang, Yuxiang Xie, Luan Xidao, and Xin Zhang. Multi-source heterogeneous data fusion. In *2018 International conference on artificial intelligence and big data (ICAIBD)*, pages 47–51. IEEE, 2018.

[38] Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, et al. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*, 2025.

[39] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Proofs

In this section, we provide a rigorous theoretical justification for the proposed method. We demonstrate that incorporating mono-source rewards as a dynamic baseline within the policy gradient framework yields an unbiased gradient estimator. Furthermore, we analyze how this mechanism effectively guides the optimization by adaptively weighting the objective based on multi-source information gain.

A.1 Notations and Problem Formulation

Consider a certain input with question q and multiple images i. We define two distinct sets of rewards independently and identically distributed (*i.i.d.*) from the same policy π_θ: (1) multi-source rollouts $\{o_j^{\text{multi}}\}_{j=1}^N$, which follow standard reasoning procedure with joint images; and additionally (2) mono-source rollouts $\{o_j^{\text{mono}}\}_{j=1}^M$, which take only a subset of visual information,*i.e.*, single-image to generate trajectories and subsequent rewards. Crucially, the reward depends solely on the input q, i and is independent of the current trajectory o generated by the policy.

Define the empirical means of the reward:

$$\mu^{\text{multi}} = \frac{1}{N}\sum_{j=1}^{N} r(q, i, o_j^{\text{multi}}),\ \mu^{\text{mono}} = \frac{1}{M}\sum_{j=1}^{M} r(q, i, o_j^{\text{mono}}). \tag{13}$$

Moreover, the proposed algorithm constructs a dynamic baseline by concatenating mono-source and multi-source rollouts:

$$\mu^{\text{hy}} = \frac{M\mu^{\text{mono}} + N\mu^{\text{multi}}}{M+N}. \tag{14}$$

The standard advantage function is obtained by performing normalization on the reward to reduce variance and regularize the training:

$$A_j = \frac{r(q, i, o_j^{\text{multi}}) - \mu^{\text{multi}}}{\sigma^{\text{multi}}}, \tag{15}$$

where σ is the variance of the reward for multi-source trajectories. In our algorithm, the advantage function is computed by replacing reward statistics with hybrid ones:

$$A_j^{\text{hy}} = \frac{r(q, i, o_j^{\text{multi}}) - \mu^{\text{hy}}}{\sigma^{\text{hy}}}. \tag{16}$$

The objective is to maximize the advantage function:

$$\max_\theta J(\theta) = \mathbb{E}_{o\sim\pi_\theta}(A(o)). \tag{17}$$

Here, we simplify the notation that the expectation of the advantage function performs on all samples and possible rollouts. Following the Policy Gradient Theorem [24], the gradient estimation of the advantage function is:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\{q,i\}\sim\mathcal{D},\ o\sim\pi_\theta(q,i)}[A \cdot \nabla_\theta \log \pi_\theta(o|q,i)]. \tag{18}$$

Denote the multi-source information gain $\Delta_{IG} = \mu^{\text{multi}} - \mu^{\text{mono}}$ as the reward signal improvement brought by multiple visual modality input compared to mono-source input. $\Delta_{IG} > 0$ underlines positive benefits from multi-source visual reasoning and $\Delta_{IG} < 0$ suggests that multi-source reasoning introduces noise.

Lemma A.1 (Policy Gradient Identity)**.** *For any differentiable policy $\pi_\theta(\tau)$, we have:*

$$\mathbb{E}_{o\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(o)\right] = 0. \tag{19}$$

Proof. By definition,

$$\begin{aligned}
\mathbb{E}\left[\nabla_\theta \log \pi_\theta(\cdot)\right] &= \int \pi_\theta(\tau)\nabla_\theta \log \pi_\theta(\tau)\, d\tau \\
&= \int \pi_\theta(\tau)\frac{\nabla_\theta \pi_\theta(\tau)}{\pi_\theta(\tau)}\, d\tau \\
&= \nabla_\theta \int \pi_\theta(\tau)\, d\tau = \nabla_\theta 1 = 0.
\end{aligned} \tag{20}$$

The first step expresses the objective in integral form over the distribution space. The second equation applies the *log-derivative trick*, defined as $\nabla_\theta \log \pi_\theta(\cdot) = \frac{\nabla_\theta \pi_\theta(\cdot)}{\pi_\theta(\cdot)}$. The third equation follows the standard regularity conditions that for any differentiable policy, we can interchange the gradient and the integral operations. And finally, given by definition that π_θ is a valid probability distribution, its integral over the distribution space is always a constant.

A.2 Proof Sketch of Theorem 3.1

Proof. We proceed by expanding the expectation of policy gradient using the linearity of expectation:

$$\begin{aligned}\nabla_\theta J^{\text{hy}}(\theta) &= \mathbb{E}_{q,i\sim\mathcal{D},\, o\sim\pi_\theta(q,i)}[A^{\text{hy}} \cdot \nabla_\theta \log \pi_\theta(o|q,i)] \\ &= \mathbb{E}_{q,i\sim\mathcal{D},\, o\sim\pi_\theta(q,i)}[r(q,i,o^{\text{multi}}) \cdot \nabla_\theta \log \pi_\theta(o|q,i)] \\ &\quad - \mathbb{E}_{q,i\sim\mathcal{D},\, o\sim\pi_\theta(q,i)}[\mu^{\text{hy}} \cdot \nabla_\theta \log \pi_\theta(o|q,i)].\end{aligned} \tag{21}$$

By the definition of gradient estimation and Lemma. A.1, the first term corresponds exactly to the policy gradient of the standard advantage function $\nabla_\theta J(\theta)$ and the second term equals zero [28]. Thus, the expectation of the gradient with a dynamic mono-source anchor remains equal to the original gradient, which completes the proof. □

We prove that for any baseline function that depends on the input sampled from the current policy π_θ does not introduce bias to gradient estimation, and also holds for the proposed advantage normalization in our algorithm. This provides a theoretical guarantee for the stability of our algorithm.

A.3 Proof Sketch of Theorem 3.2

Proof. The standard gradient estimation (ignoring constant scaling σ) can be formulated as:

$$\nabla_\theta J(\theta) = \frac{1}{N}\sum_{j=1}^{N}(r(q,i,o_j^{\text{multi}}) - \mu^{\text{multi}})\nabla_\theta \log \pi_\theta(o_j^{\text{multi}}|q,i), \tag{22}$$

where standard advantages have a zero mean:

$$\sum_{j=1}^{N}(r(q,i,o_j^{\text{multi}}) - \mu^{\text{multi}}) = 0, \tag{23}$$

that guarantees reduced gradient variance. Consider the gradient under hybrid advantage normalization:

$$\nabla_\theta J^{\text{hy}}(\theta) = \frac{1}{N}\sum_{j=1}^{N}(r(q,i,o_j^{\text{multi}}) - \mu^{\text{hy}})\nabla_\theta \log \pi_\theta(o_j^{\text{multi}}|q,i). \tag{24}$$

In the sense of expectation, the mean of the proposed algorithm can be regarded as a mixture distribution over mono-source and multi-source rollouts:

$$\mu^{\text{hy}} = \alpha\mu^{\text{multi}} + (1-\alpha)\mu^{\text{mono}}, \tag{25}$$

where $\alpha = N/(M+N)$ is the proportion of multi-source trajectories. Therefore, expanding the normalization term:

$$\begin{aligned}&r(o_j^{\text{multi}}) - \mu^{\text{multi}} \\ &= r(o_j^{\text{multi}}) - \left(\alpha\mu^{\text{multi}} + (1-\alpha)\mu^{\text{mono}}\right) \\ &= r(o_j^{\text{multi}}) - \left(\mu^{\text{multi}} - (1-\alpha)\mu^{\text{multi}} + (1-\alpha)\mu^{\text{mono}}\right) \\ &= (r(o_j^{\text{multi}}) - \mu^{\text{multi}}) + (1-\alpha)(\mu^{\text{multi}} - \mu^{\text{mono}}) \\ &= A_j + (1-\alpha)\Delta_{IG}.\end{aligned} \tag{26}$$

Substituting the simplified term into the original expression, the hybrid gradient can be decomposed into two parts:

$$\nabla_\theta J^{\text{hy}}(\theta) = \underbrace{\frac{1}{N}\sum_{j=1}^{N} A_j \nabla_\theta \log \pi_\theta(o_j^{\text{multi}}|q,i)}_{\text{Original GRPO gradient}} + (1-\alpha)\underbrace{\Delta_{IG}\Big(\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta \log \pi_\theta(o_j^{\text{multi}}|q,i)\Big)}_{\text{Information gain regularization}}. \tag{27}$$

It can be concluded that the gradient estimation of our algorithm is to optimize the standard gradient weighted with information gain regularization during training, where the first term refers to the original GRPO gradient estimation, and the second term is a regularization term in proportion to the relative information gain. This completes the proof. □

We reveal that the method amplifies the relative performance signal between multi-source and mono-source reasoning by anchoring a mono-source baseline, preserves unbiased policy gradients, and inherits standard convergence guarantees. Concretely, (1) if the mono-source baseline performs well, μ^{hy} increases. This sets a higher bar for the gradient policy. If the multi-source model generates a response with a reward lower than the mono-source one, the advantage A^{hy} becomes negative. Consequently, the optimization suppresses the probability of this trajectory; (2) if the mono-source anchor is low, μ^{hy} decreases. The model is effectively rewarded for any marginal improvement over mono-source performance, encouraging effective utilization of multi-source cues.

B More Details of Experiments

B.1 Implementation details.

We build the codebase for visual reasoning based on verl [3]. The main experiments are trained on NVIDIA A800 for 2 epochs to get a fair comparison. In evaluation, we separately reason with every single source and get the best result for each question as the final mono-source performance to see the effectiveness of the mono-source anchor. The results are constructed by extracting the prediction from the answer part `<answer>`$\cdots$`<answer>`.

B.2 Datasets.

LLVIP is an object detection dataset. It is composed of RGB and infrared images corresponding to the same scene. It contains a large number of scenes with poor light conditions at night with rich pedestrian annotations, validating the robustness of cross-modal perception under challenging illumination scenarios. We therefore use it as the RGB & Infrared dataset.

SpatialQA is a robotic manipulation dataset that aims to precisely comprehend spatial information through depth images. It is composed of corresponding RGB-Depth image pairs with multi-level questions across various embodiment tasks and scenarios, and we exploit it to evaluate RGB & Depth polar questions.

nuScenes is a multimodal VQA dataset for autonomous driving scenarios. For each sample in the driving scenarios, it consists of six surrounding images and question-answer pairs based on visual scenes and question templates. Its different perspectives can be seen as different sources of information, so we employ it to evaluate the performance of visual reasoning models in multi-view scenarios.

OCR-VQA is a visual question answering dataset focusing on scene text understanding. It contains real-world images with rich embedded text, which requires accurate reasoning over the recognized text. We use the dataset to evaluate text-centric visual reasoning with multiple-choice questions.

[3] https://github.com/volcengine/verl

C More Experimental Results

Results on general multimodal benchmarks. We also comprehensively evaluate the general capabilities of different methods including single-image benchmarks MMstar [9], HalluBench [11] TextVQA [23] and multi-image benchmarks Mantis [15], Muirbench [26] Blink [10]. It can be seen from Table 4 that our method also preserve the capability in general multi-modal benchmarks with average **1.7%** improvement, showing the generalizability.

Table 4: Comparison on general multimodal benchmarks.

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)

More analysis of computational overhead. In our method, mono-source trajectory is generated for each visual source. Therefore, the efficiency when incorporating more visual sources should be carefully discussed. To better illustrate this, we increase the number of mono-source samples M in nuScenes, where mono-source rollouts are generated for six visual sources in training sample and report wall-clock time and GPU memory overhead to analyze its impact on computational cost.

The results in Table 5 show that adding more mono-source samples brings performance saturation (75.0→76.0), negligible increase in GPU memory (52 GB→54 GB) and training time (11.1×8 h→11.5×8 h) without additional GPU memory consumption in backpropagation (80 GB). This indicates that mono-source samples are not computational bottleneck, which can be attributed to the anchor function of mono-source rewards that are not actually engaged in loss backpropagation. The major computational cost is still backward-pass of N multi-source rewards, strongly validating the effectiveness and efficiency of the proposed method.

Table 5: Detailed analysis of computational overhead.

M	0	6	12
Performance	72.0	75.0	76.0
GPU Memory (rollout generation)	50 GB	52 GB	54 GB
GPU Memory (backpropagation)	80 GB	80 GB	80 GB
Wall-Clock Training Hours	10.6×8	11.1×8	11.5×8

Table 6: Influence of different anchors.

Method	GRPO		DAPO	
	Union	Multi	Union	Multi
Fixed	88.1	84.8	88.9	85.2
Weighted	86.9	79.7	86.2	80.6
Ours	89.9	87.7	92.3	89.4

Comparison to alternative anchors. To validate the effectiveness of the proposed anchor, we replace our dynamic mono-source anchor with a fixed value function or a weighted average reward of different modalities. In Table 6, both alternatives lead to a significant performance drop from 2% to 10%. By contrast, our approach that integrates concurrent single- and multi-source rewards with simple average theoretically ensures consistent and context-aware advantage estimation, underscoring the importance of using an on-policy, rollout-based anchor.

Per-source results. Figuring out which source is strong and the relation with multi-source reasoning is important for understanding the effectiveness. Therefore, we conduct experiments on infrared and depth and report the performance of each source and to show the influence of different sources. Results are shown in Table 7. It can be concluded that (1) infrared and depth images can provide reliable evidence for certain scenes, so they are strong and can obtain better performance than RGB, which is consistent with the motivation of the paper; (2) our method consistently captures reliable modality, achieves better performance in each modality, and as a result gets better performance in Multi, showing the effectiveness.

Table 7: Per-source results on infrared and depth.

Method	Infrared				Depth			
	RGB-only	Infrared-only	Multi	Union	RGB-only	Depth-only	Multi	Union
Qwen2.5-VL-3B	46.7	51.8	38.8	63.0	65.0	52.7	73.8	86.3
DAPO	71.9	82.0	85.5	88.8	62.4	68.7	73.8	76.3
DAPO+MARS	73.1	83.5	89.4	92.3	65.1	71.2	77.5	82.5

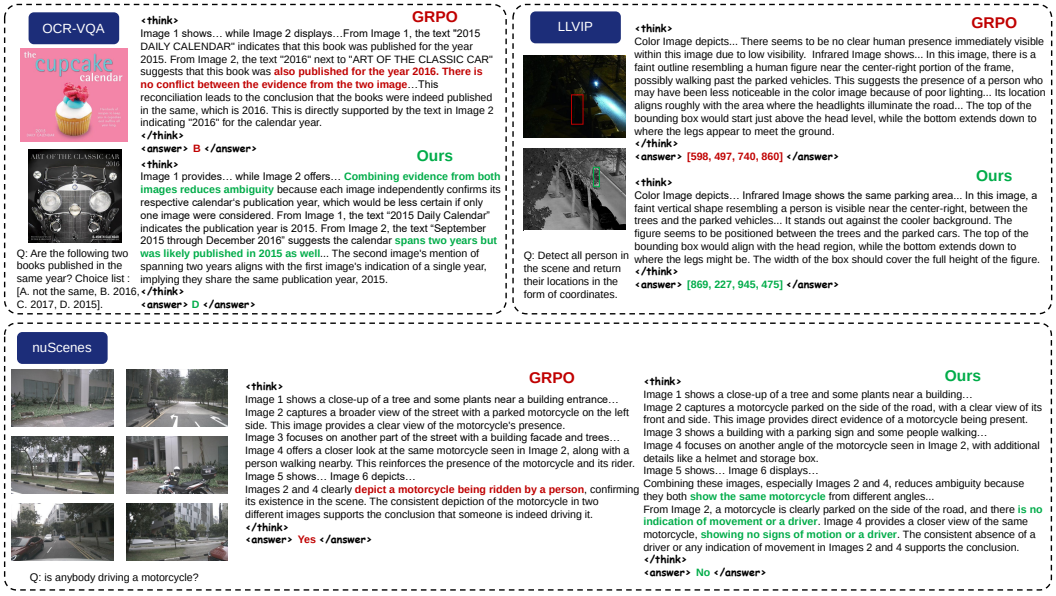


Figure 7: More qualitative results of visual reasoning on OCR-VQA, LLVIP and nuScenes.

More qualitative results. We visualize more comparison examples in Figure 7 to show that our method effectively focuses on specific source and gives proper responses accordingly in different kinds of occasions, which further validates the effectiveness of our method.

D Limitations and Future Directions

This work proposes a multi-source visual reasoning framework for MLLMs that explicitly models and quantifies multi-source information gain from modality integration and effectively improves the performance. Despite the promise, several issues are to be addressed. One primary limitation is that due to resource constraints, our evaluation is mainly conducted on moderately sized multimodal models and does not extend to more types of foundation models and reinforcement learning algorithms. Nevertheless, since our approach is general and architecture-agnostic, we believe it can generalize well. Moreover, our study focuses on different visual modalities and does not directly address broader multimodal scenarios. Real-world reasoning tasks would also involve additional modalities such as video, audio, and 3D point clouds, which introduce temporal and structural complexity. We believe the core principle of dynamic anchor can generalize beyond multi-source visual reasoning and would treat adapting our framework to these richer multimodal settings as a future direction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See in Sec. 4 and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and data are released.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See in Sec. 4 and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors have reviewed the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: No societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We correctly cite the original paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: No LLMs used in the core methods in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.