PaLMR: Towards Faithful Visual Reasoning via Multimodal Process Alignment

Yantao Li[1,2,3], Chenyang Yan[1], Qiang Hui[2,3], Fang Zhao[2,3,†], Kanzhi Cheng[1],
Chao Tan[2,3], Huanlin Gao[2,3], Jianbing Zhang[1,*], Kai Wang[2,3], Xinyu Dai[1], Shiguo Lian[2,3,*]

[1] National Key Laboratory for Novel Software Technology, Nanjing University
[2] Data Science & Artificial Intelligence Research Institute, China Unicom
[3] Unicom Data Intelligence, China Unicom

li_yantao@smail.nju.edu.cn {zjb,daixinyu}@nju.edu.cn {huiq,zhaof50,liansg}@chinaunicom.cn

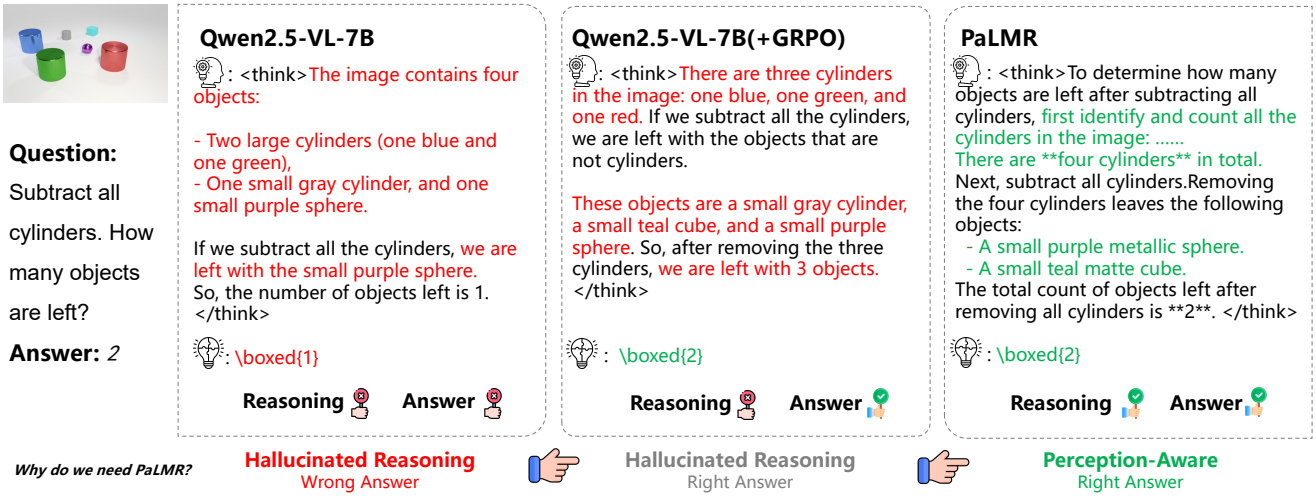


Figure 1. **Comparison of reasoning behaviors among baseline models, baseline models(+GRPO) and PaLMR on a visual reasoning sample.** As shown, PaLMR demonstrates perception-aware reasoning and produces faithful answers by process-level perception alignment, addressing the hallucinated reasoning issue in prior models.

Abstract

*Reinforcement learning has recently improved the reasoning ability of Large Language Models (LLMs) and Multimodal LLMs (MLLMs), yet prevailing reward designs emphasise final-answer correctness and consequently tolerate process hallucinations—cases where models reach the right answer while misperceiving visual evidence. We address this process-level misalignment with **PaLMR** (Process Alignment for Multimodal Reasoning), a framework that aligns not only outcomes but also the reasoning process itself. **PaLMR** comprises two complementary components: a perception-aligned data layer that constructs process-aware reasoning data with structured pseudo-ground-truths and verifiable visual facts, and a process-aligned optimisation layer that constructs a hierarchical reward fusion scheme with a process-aware scoring function to encourage visually faithful chains-of-thought and improve training stability. Experiment on Qwen2.5-VL-7B shows that our approach substantially reduces reasoning hallucinations and improves visual reasoning fidelity, achieving state-of-the-art results on HallusionBench while maintaining strong performance on MMMU, MathVista, and MathVerse. These findings indicate that PaLMR offers a principled and practical route to process-aligned multimodal reasoning, advancing the reliability and interpretability of MLLMs.*

1. Introduction

Multimodal large language models(MLLMs) have recently achieved remarkable progress in visual reasoning and perception-language understanding. Building upon reinforcement learning(RL) successes in large language models(LLMs) such as DeepSeek-R1 [15] and R1-Zero [62], researchers have extended outcome-based opti-

* Corresponding authors. † Project leader.

mization to multimodal settings, outcome many promissing models, like MM-Eureka [32], OpenVLThinker [12] and Perception-R1 [54], which improve answer accuracy across benchmarks such as MMMU [55] and MathVista [28]. However, most existing reward mechanisms focus solely on what the model answers, while overlooking how it reasons. This often leads to reasoning hallucinations — models obtain correct answers through visually inconsistent reasoning. For example, it may claim "three cups are on the table" on the chain-of-thought(CoT) [47] when four are clearly visible, yet still predict the right answer based on textual priors.

Such hallucinations reveal a core limitation of current multimodal reinforcement learning: rewards are primarily *outcome-oriented*, offering supervision only for final correctness while neglecting the *faithfulness* of process reasoning steps. This observation highlights the need for *process-level alignment*: ensuring that the reasoning trajectory remains consistent with visual evidence at every step. Previous reinforcement learning with visual reasoning methods [12, 32] mainly optimize textual reasoning results, while visual reward models such as VisualPRM [44] and VRPRM [8] rely on human preference comparisons between visually correct and incorrect samples.

In contrast, our method links perceptual accuracy to process-level reasoning via a verifiable binary visual-textual consistency metric. We propose **PaLMR** (**P**rocess **a**L**i**gnment for **M**ultimodal **R**easoning) to enforce visual faithfulness throughout the reasoning process, moving beyond outcome-only correctness. PaLMR comprises two components: a perception-aligned data layer (**PaDLayer**) that generates visually grounded reasoning samples, and a process-aligned optimization layer (**PaOLayer**) that reinforces trajectory coherence. To stabilize policy learning, **PaOLayer** employs a hierarchical reward fusion scheme integrating perception and outcome scores. We implement this via **Vision-Guided Group Relative Policy Optimization (V-GRPO)**, embedding visual-consistency rewards into the *RL* objective. V-GRPO prioritizes process-level perception feedback, effectively unifying perception fidelity and reasoning quality.

Extensive experiments on multiple benchmarks demonstrate that **PaLMR** consistently improves reasoning faithfulness and visual consistency while maintaining competitive answer accuracy. Notably, our method significantly reduces hallucination rates in reasoning and achieves state-of-the-art visual alignment performance among multimodal reasoning models.

Our main contributions are summarized as follows:

- We introduce the **PaLMR** framework, a faithful multimodal process alignment framework that enforces reasoning-process faithfulness by unifying perception-aligned data construction layer and process-aligned optimization layer.
- We propose **V-GRPO** training paradigm that enhances visual faithfulness by incorporating a perception-aware scoring strategy and integrating visual cues into the **GRPO** framework, forming a hierarchical reward mechanism that jointly optimizes reasoning accuracy and perception consistency.
- Experiments show that **PaLMR** significantly outperforms baseline models in both alignment faithfulness and reasoning quality.

2. Related Work

2.1. Multimodal Chain-of-Thought Reasoning

The Chain-of-Thought (CoT) [47] paradigm has become central to reasoning enhancement in both textual and multimodal large language models [49]. Early works such as *Visual Thoughts* [11] and *CoT-VLA* [61] demonstrated that decomposing multimodal reasoning into explicit intermediate steps significantly improves interpretability and performance on benchmarks like ScienceQA [25] and A-OKVQA [35]. *Kam-CoT* [34] and *LLaVA-CoT* [51] extended this paradigm to vision–language models by introducing explicit visual reasoning chains. Subsequent works such as *Self-Consistency* [45] and *Tree-of-Thought* [53] encouraged multi-path exploration and reflection. While test-time compute methods [10, 38, 39] incorporated self-reflection signals with CoT to improve reasoning performance. Together, these studies established CoT reasoning as a fundamental approach for stepwise multimodal understanding.

2.2. Reinforcement Learning for Multimodal Reasoning

Reinforcement-learning frameworks have been widely used to improve reasoning robustness and factuality. *DeepSeek-R1* [15] and *R1-Zero* [62] introduced group-relative policy optimization to incentivize structured reasoning behavior in large language models. Extending this idea to vision–language models, *Vision-R1-Zero* [17] and *R1-VL* [58] applied rule-based and visual-feedback-guided optimization to enhance multimodal reasoning consistency. These methods established a strong foundation for post-training alignment of reasoning behavior but primarily relied on outcome-level feedback: judging only the correctness of the final answer without explicitly supervising intermediate reasoning or perceptual faithfulness.

2.3. Process Reward Models for Multimodal Reasoning

Recent work has shifted from outcome-based RL to process-level supervision, rewarding models for faithful intermediate reasoning steps. *VRPRM* [8] pioneered vi-

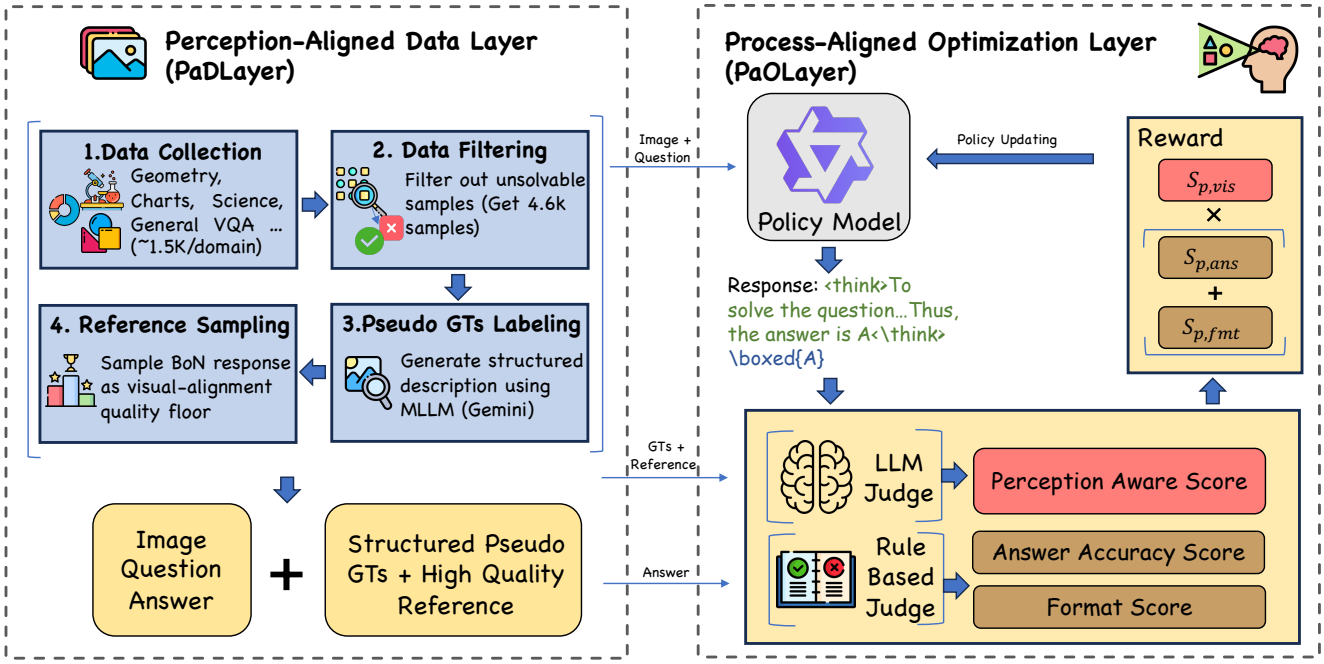


Figure 2. **Overview of the proposed PaLMR framework.** The model adopts a two-layer architecture: (a) the **Perception-Aligned Data Layer (PaDLayer)** builds process-aware multimodal data with structured pseudo ground truths and verifiable visual facts; and (b) the **Process-Aligned Optimization Layer (PaOLayer)** integrates perception-aware, answer, and format rewards into GRPO to enforce visually faithful and logically coherent reasoning.

sual reasoning process reward modeling with a two-stage training strategy, achieving over 100% relative improvement in Best-of-N(BoN) selection. *ViLPRM* and its benchmark *ViLBench* [43] evaluated visual–language process rewards at scale, while *VisualPRM* [44] provided fine-grained step-level evaluation using a BoN selector. Further variants such as *Mm-PRM* [13], *PRM-BAS* [16], and *GM-PRM* [57] supervise each reasoning step to reduce error accumulation; *GM-PRM* even treats the PRM as an active collaborator that can refine reasoning chains via the Refined-BoN strategy. Other models, including *Dream-PRM* [5], *ER-PRM* [56] and *EDU-PRM* [4], enhance calibration through entropy regularization or ranking-based supervision.Weakly and self-supervised approaches such as *FreePRM* [40] explore pseudo-label and self-reward signals to reduce annotation costs. Unified frameworks—*Unified Multimodal Chain-of-Thought Reward Model* [46] and *Unified Reward Model* for Multimodal Understanding and Generation [60]—generalize process reward modeling across reasoning, understanding, and generation tasks. These studies highlight a growing shift toward process-aware reward modeling, in which models are evaluated and optimized based on the quality, coherence, and perceptual grounding of their intermediate reasoning steps.

3. Method

We propose **PaLMR**, a unified framework that enhances visual faithfulness by aligning perception and reasoning via process-level scoring and reinforcement optimization. As in Fig. 2, PaLMR has two interdependent layers: a **PaDLayer** that constructs verifiable multimodal data, and a **PaOLayer** that enforces process alignment through hierarchical scoring and vision-guided GRPO.

3.1. Preliminaries

A visual reasoning task often provides a textual question along with an image and asks an MLLM to provide a response sequence for the question. Formally, given a query sequence $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ and corresponding image $\mathbf{I}$, the MLLM model $\mathcal{M}$ generates sequence $\mathbf{y} = [y_1, y_2, \ldots, y_m]$, where x_i and y_i are individual tokens. The whole responce $\mathbf{y}$ is sampled from conditional distribution $p_\theta(\cdot|\mathbf{x}, \mathbf{I})$, as $p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{I}) = \prod_{j=1}^{m} p_\theta(y_j|\mathbf{x}, \mathbf{I}, y_1, y_2, \ldots, y_{j-1})$.

Group Relative Policy Optimization(GRPO)[36]: GRPO is a widely used policy gradient RL algorithm that leverages intra-group relative performance to optimize the policy model. During Training, for each query x, the model samples a group of G responses, and GRPO computes the relative advantage of the group based on its rewards.

$\{r_1, r_2, \ldots, r_G\}$ as follows:

$$A^i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^G)}{\text{std}(\{r_i\}_{i=1}^G)}$$

Using the advantage values, GRPO optimizes the model with PPO-clip loss in terms of:

$$J_{\text{GRPO}}(\theta) = -\mathbb{E}_{\substack{x\sim\mathcal{D} \\ y\sim\pi_\theta}} \left[\text{Avg}(\min\left(\psi_i A_i,\ \text{clip}(\psi_i, 1\pm\epsilon)A_i\right)\right] \tag{1}$$

where ψ_i and A_i donate per-token policy ratio and advantage on response y_i, and $\text{Avg}(\cdot)$ means the average over group and response length. By adjusting the components of the reward function, practitioners can steer the policy toward behaviors that yield higher rewards. In RLVR (reinforcement learning from verifiable rewards), it is common to use rule-based, verifiable signals such as exact-answer correctness to provide reliable rewards and mitigate the risk of reward hacking.

However, when exact-answer correctness is used as the sole reward signal during optimization of Multimodal Large Language Models (MLLMs), the resulting optimization tends to prioritize textual accuracy at the expense of visual perception integrity. This text-centric reward fails to adequately capture the importance of visual information in multimodal reasoning tasks. To address this limitation, we propose a pairwise, visually aware score that explicitly aligns the model's responses with visual fidelity during reinforcement learning.

3.2. Perception-Aligned Data Layer (PaDLayer)

PaDLayer establishes a verifiable basis for process-level alignment through a four-step procedure (see Figure 2). We begin by uniformly sampling 1,500 instances from FineVision [48] across multiple reasoning domains, including geometry, charts, science, and OCR, to mitigate distributional bias.

We then apply a data filtering strategy based on learnability to select informative examples. Specifically, we perform stochastic rollouts under the **PaOLayer** policy and remove outliers that are consistently incorrect or unstable, as well as trivial cases with excessively high accuracy that do not contribute meaningful reinforcement learning signals. We further exclude formatting-incompatible problems using rule-based matching. This process results in approximately 4,700 retained instances.

To establish verifiable targets, we use *Gemini* to generate structured pseudo ground-truths for each sample. These captions enumerate objects, spatial relations, and visual attributes, thereby converting images into symbolic representations. We then sample a semantically coherent reasoning trajectory from the policy model using a Best-of-N selection, which serves as the reference-response baseline for subsequent visual-faith optimization.

3.3. Process-Aligned Optimization Layer (PaO-Layer)

After the training sets are ready, we use them to enhance visual process reasoning by coupling perception-aware scoring with **GRPO**, forming a **Vision-Guided GRPO (V-GRPO)** training strategy. **V-GRPO** are constructed by assessing the visual faithfulness of reasoning trajectories through pairwise comparison and producing reliable binary visual fidelity scores. This integration ensures that models learn to reason accurately, perceiving not only the final correctness.

Perception-Aware Scoring: A common approach to quantifying visual consistency is first to extract visual claims $\mathbf{Z}$ from the model's chain of thought and then compute a consistency score by comparing them against a set of ground-truth facts. This score, denoted as the visual-aligned score, can be integrated into the reward function to penalize hallucinated or missing visual details. However, our preliminary experiments revealed a significant limitation of this point-wise evaluation paradigm: it is susceptible to the intrinsic biases of the "LLM-as-judge". We observed that this method tends to yield high accuracy only for reasoning paths that are already essentially correct, but struggles to provide a reliable judgment when imperfect or partially correct trajectories appear. To mitigate evaluator bias while retaining the flexibility of LLM-as-judge, we replace point-wise scoring with pairwise comparisons, using the LLM to judge which trajectory demonstrates more faithful and coherent reasoning. To ensure stable, consistent comparisons throughout the RL training process, we first sample a target trajectory τ_{target} during training. With a given high-quality reference trajectory τ_{ref} sourcing from *PaDLayer*, LLM determines whether the former is superior to the latter. This finally yields a binary visual fidelity score:

$$S_{p,vis}(\tau) = \mathbb{I}(\tau_{\text{target}} \succ \tau_{\text{ref}}) \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function and the relation $\succ$ is a preference induced by the LLM-as-judge given the ground-truth facts. $S_{v,vis}(\tau)$ equals 1 if the target is preferred and 0 otherwise.

Figure 3 illustrates that the pairwise evaluation paradigm achieves a significantly higher human alignment ratio compared to the point-wise method. Furthermore, utilizing powerful models, such as Qwen3[52], can lead to over 88% model-human alignment. Such a high level of concordance between LLM-as-judge and human evaluators provides a robust and reliable binary signal for optimizing the model's perceptual grounding.

This approach effectively combines the scalability of LLM-as-judge with the enhanced accuracy of comparative judgment, providing a robust, human-aligned signal for perceptual fidelity. However, it remains an isolated evaluation metric unless tightly coupled with the optimization pro-

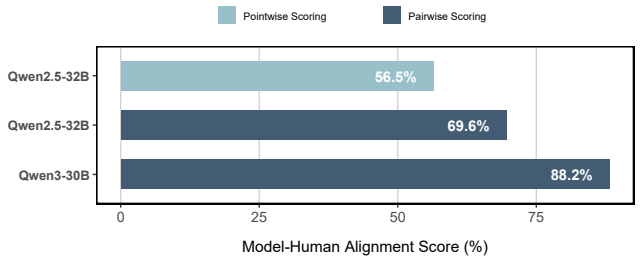


Figure 3. **Model-Human alignment ratio in identifying visual perception errors.** The evaluation assesses visual perception errors in 100 randomly sampled responses generated from the Geo3K dataset with Qwen2.5-VL-7B, and using Qwen2.5-32B, Qwen3-30B as judgement model.

cess. So we embed S_p^{pair} directly into the reward structure of GRPO, forming a Vision-Guided GRPO (V-GRPO).

V-GRPO: To enforce visual alignments as a prerequisite for successful task completion, we design a hierarchical reward function for our **V-GRPO**. It prioritizes visual cues, following the principle of Perception-Aware Scoring, e.g., S_p. A trajectory deemed visually unfaithful ($S_p = 0$) receives no reward for its final answer, regardless of its accuracy. So it strictly penalizes any reasoning that contains visual hallucinations or misinterpretations. The total reward $R_{V-GRPO}(\tau)$ is formulated as a hierarchical combination of scores:

$$R_{V-GRPO}(\tau) = S_{p,\text{vis}}(\tau) \cdot (\alpha S_{p,\text{ans}}(\tau) + (1-\alpha) S_{p,\text{fmt}}(\tau)) \tag{3}$$

where $S_{p,\text{vis}}(\tau)$ is the binary visual fidelity score, $S_{p,\text{ans}}(\tau)$ and $S_{p,\text{fmt}}(\tau)$ are rule-based scores for answer-accuracy and format-correctness and α is a balancing coefficient. Specifically, $S_{p,\text{vis}}(\tau)$ enforces visual fidelity and holds the highest priority—if a trajectory contains perceptual errors, the entire reward is set to zero. $S_{p,\text{ans}}(\tau)$ encourages task-level correctness, and $S_{p,\text{fmt}}(\tau)$ ensures that the model's outputs remain structurally and syntactically consistent. By integrating this reward mechanism into the GRPO algorithm, we propose V-GRPO. This approach explicitly forces the model first to learn to "see correctly" before learning to "reason correctly," thereby promoting the development of more faithful and reliable models.

4. Experiments

4.1. Experimental Setup

We briefly introduce the training dataset, implementation details, baselines, and evaluation settings in this section.

Training Dataset: As stated in Secion 3.2, we curated data from 19 distinct domains within the FineVision [48] dataset, spanning areas such as geometry, Chart, Science, OCR, and general VQA. For each domain, we first sampled approximately 1.5K instances and discarded samples with

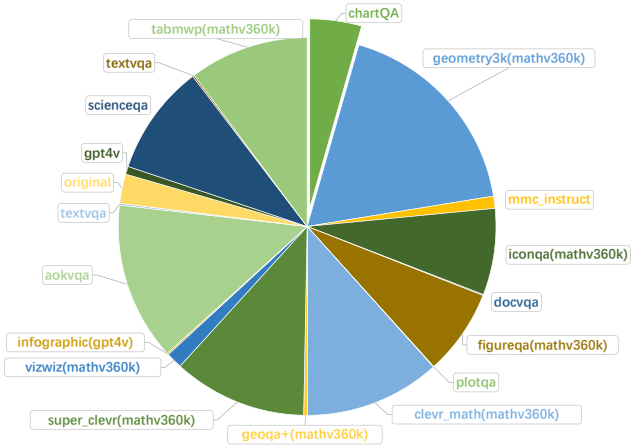


Figure 4. **Data distribution after our PaDLayer data filtering**. 19 distinct sub-domains are selected, and 4728 samples are finally used to generate the training dataset.

low image-question relevance. The remaining candidates underwent a learnability-based data filtering step, resulting in a final training set of 4,728 high-quality instances. After that, we augment the standard VQA triplet(image, question, answer) with structured pseudo visual ground-truth and a reference response, providing richer guidance for process alignment training.

Implementaion Details: We select a widely used multimodal LLM, Qwen2.5-VL-7B [1], as **PaLMR**'s backbone and trained it with the VeRL [41]. In each reinforcement learning loop, we use Qwen3-30B-A3B [52] as the percept-score judging LLM. All experiments were conducted on a cluster with 8 NVIDIA H100-80 GPUs. We use EasyR1 default parameters, but set the learning rate to 1e-6, the batch size to 128, the rollout batch size to 512, and the temperature to 1.0 during rolling. **V-GRPO** group size(G) is set to 16, and the model is trained for 20 epochs with KL loss disabled. For reward calculation, we set the score coefficient α to 0.9 to emphasize the visual score of the final answer.

Baseline approaches: To build a comprehensive evaluation of our method, we compare it with a vanilla GRPO training baseline and basemodel without training. Additionally, we include state-of-the-art MLLMs and similar open-source R1-style MLLMs: (1) Proprietary MLLMs: GPT4o [18], Gemini [42], (2) Open-source general MLLMs: Qwen2.5-VL series, InternVL 2.5 series [9], (3) Open-source Reasoning MLLMs: MM-Eureka [32], OpenVLThinker [12], and Perception-R1 [50], All of which use Qwen2.5-VL-7B as the MLLM backbone, similar to our setting.

Evaluation Benchmarks: To demonstrate the enhanced reasoning performance of **PaLMR** across different domains, we employ a comprehensive set of visual math reasoning tasks and general multi-domain perception-centered

Model	#Data	MMMU_{val}	HallusionBench	MathVerse*	MMStar	MathVista
GPT-4o [18]	-	60.0	68.0	-	-	63.8
Gemini2-Flash [42]	-	70.6	69.4	-	-	70.4
Qwen2.5-VL-72B [1]	-	68.2	71.4	-	70.8	74.8
Qwen2.5-VL-32B [1]	-	63.7	72.1	54.3	67.3	74.7
InternVL2.5-8B [9]	-	56.2	67.4	-	62.9	64.4
MM-Eureka-7B [32]	15K	55.4	69.5	46.6	64.6	73.0
OpenVLThinker-7B [12]	12K	56.3	66.9	40.4	62.1	70.2
Perception-R1-7B [50]	2K	56.3	70.0	46.1	66.3	73.6
Qwen2.5-VL-7B [1]	-	56.4	63.8	42.6	64.3	68.2
+ GRPO	4.7K	57.8	66.7	45.9	66.0	**74.1**
PaLMR-7B	4.7K	**59.3**	**70.9**	**47.5**	**67.1**	73.8

Table 1. **Performance comparison of MLLMs on a suite of out-of-domain benchmarks.** Accuracy are reported for all benchmarks. MathVerse use vision only subset. Open-source models are evaluated using VLMEvalkit, and R1-style reasoning models are used the same template as it training if provided.

reasoning tasks. These benchmarks include HallusionBench [14] and MMStar [7], which assess the model's robustness and performance on perception-intensive tasks that require fine-grained visual grounding on massive domains. Additionally, MMMU [55], MathVista [28], and the MathVerse [59] vision-only subset are focus problems demanding complex logical deductions from diagrams and images.

4.2. Main Results

Table 1 shows that **PaLMR** achieves the best results on most out-of-distribution benchmarks among 7B-scale MLLMs. It not only surpasses general-purpose open-source MLLMs such as InternVL-8B across all benchmarks, but also outperforms reinforcement-learning-based reasoning models, including MM-Eureka and Perception-R1.

The PaDLayer data curation pipeline demonstrates strong data efficiency. Without visual-aware scores, the GRPO baseline (second-to-last row in Table 1) outperforms OpenVLThinker across all benchmarks, requiring only 4.7K training examples compared to 12K for OpenVLThinker, a 2.5-fold reduction under identical GRPO training conditions. When a process-level perception-aware score is incorporated into V-GRPO, the PaOLayer achieves state-of-the-art performance among 7B-scale models. PaLMR surpasses MM-Eureka-7B on MathVerse (47.5 vs 46.6) and HallusionBench (70.9 vs 69.5), even though vanilla GRPO does not outperform Eureka. These findings highlight the critical role of process-level visual alignment in GRPO: process-level visual scoring enforces consistency and reduces hallucinated reasoning (Figure 1).

PaOLayer achieves strong scalability through noise robustness. We compare against Perception-R1, which also integrates visual-aware scores into RL. Perception-R1 relies on point-wise scoring and requires Gemini to generate complex, question-relevant visual content. In contrast, PaOLayer employs pairwise reranking and only requires Gemini to produce a generic image caption. Despite relying on these less specialized data, PaLMR outperforms Perception-R1 on all benchmarks. This confirms that pairwise scoring, by leveraging self-generated reference responses, reduces reliance on high-quality, task-specific annotations and offers better scalability potential.

Qualitative comparisons (Figure 5) further validate that PaLMR consistently produces visually aligned reasoning chains while maintaining answer correctness. Ultimately, process-level perception alignment improves not only benchmark performance but also reasoning stability, as further analyzed in Figure 6.

4.3. Generalizability Across Model Scales and Architectures

To validate the robustness and scalability of PaLMR, we extend our evaluation beyond the base 7B model to different parameter scales and newer architectural generations. The comprehensive results are summarized in Table 2.

Across the evaluated model scales ranging from 3B to 32B on the Qwen2.5 family, PaLMR consistently outperforms the GRPO method on out-of-domain benchmarks. In particular, on the 32B model, PaLMR improves the baseline MMMU performance from 64.3 to 66.8, while GRPO slightly degrades to 64.0. This confirms that PaLMR generalizes well to models within the same family.

While PaLMR yields substantial improvements on the Qwen2.5-VL series, its performance-boosting effect diminishes when applied to the more advanced Qwen3-VL-8B architecture. Specifically, on the HallusionBench and MathVerse Vision Only benchmarks, PaLMR achieves scores of 75.2 and 60.0, marginally underperforming the standard GRPO baselines of 75.3 and 60.8. This performance saturation can be attributed to the degradation of the visual

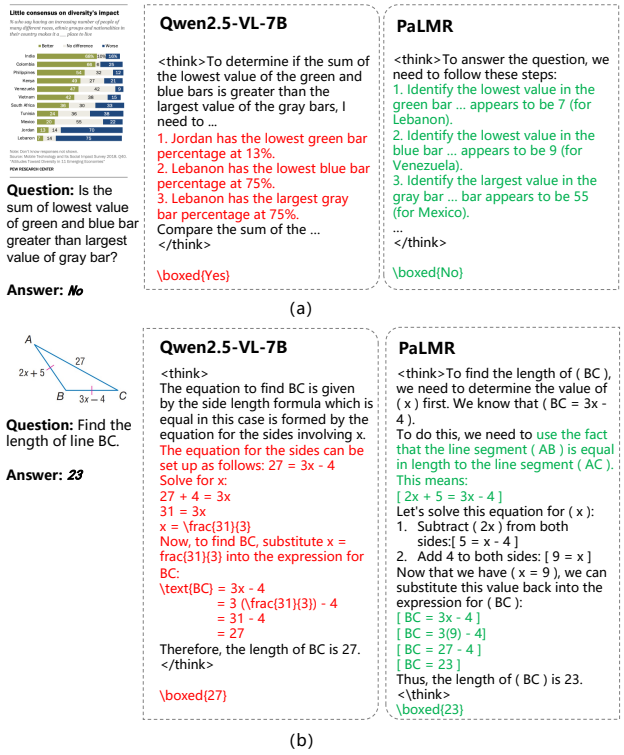


Figure 5. **Qualitative comparison of reasoning chains between baseline models and PaLMR across different domains.**

Model	MMMU	Hallusion	MathVerse	MMStar	MathVista
Qwen2.5-VL-3B	49.3	-	32.4	55.0	63.3
+ GRPO	53.3	-	34.0	56.7	62.8
PaLMR-3B	**53.9**	-	**38.8**	**57.2**	**64.0**
Qwen2.5-VL-7B	56.4	63.8	42.6	64.3	68.2
+ GRPO	57.8	66.7	45.9	66.0	**74.1**
PaLMR-7B	**59.3**	**70.9**	**47.5**	**67.1**	73.8
Qwen2.5-VL-32B†	64.3	69.1	49.4	65.8	75.1
+ GRPO†	64.0	70.1	49.7	66.9	72.9
PaLMR-32B†	**66.8**	**71.5**	**51.3**	**67.6**	74.5
Qwen3-VL-8B†	61.3	73.5	-	70.9	77.2
+ GRPO†	64.6	**75.3**	**60.8**	72.6	77.5
PaLMR-8B†	**65.3**	75.2	60.0	72.6	**78.3**

Table 2. **Comparison of model performance across different scales on out-of-domain benchmarks.** † indicates inference via vLLM.

reward mechanism. PaLMR utilizes a pair-wise scoring paradigm based on reference data annotated by the less capable Qwen2.5-VL-7B model. As the target model's intrinsic capabilities eclipse those of the annotator, the judge model inherently loses its discriminative precision. Consequently, the visual gate mechanism fails to accurately quantify performance gains, causing the optimization framework to degenerate into standard GRPO.

4.4. Effectiveness of the Perception-Aware Score

To systematically evaluate the integration of visual perception signals with the correctness of the final answer, we

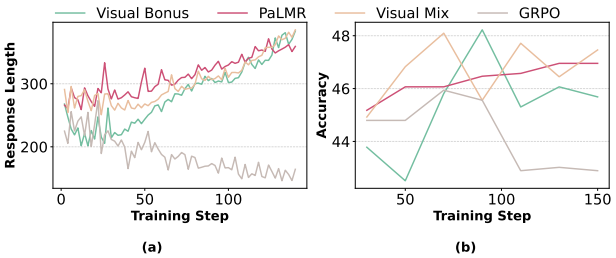


Figure 6. **(a) Comparison of average response length on the training set across different reward settings. (b) Comparison of accuracy on the MathVerse vision-only benchmark.**

Model	MMMU_{val}	MathVerse*	MMStar	MathVista
Qwen2.5-VL-7B	56.4	42.6	64.3	68.2
+ GRPO	57.8	45.9	66.0	**74.1**
Visual Bonus-7B	59.3	**48.2**	65.0	73.9
Visual Mix-7B	58.6	48.1	66.5	72.5
PaLMR-7B	**59.3**	47.5	**67.1**	73.8

Table 3. **An ablation study on the formulation strategies of visual-aware rewards.** Utilizing the identical configuration for training as the primary experiment, we compare the proposed **PaLMR** against the strategies of Visual Bonus and Visual Mix across standard benchmarks.

unify and compare multiple strategies for reward formulation. These strategies combine the visual alignment score $S_{p,vis}$, the answer correctness score $S_{p,ans}$, and the formatting score $S_{p,fmt}$ in distinct configurations. Specifically, we design three representative variants within a standard GRPO training framework:

1. *Vanilla GRPO* (baseline) and **PaLMR** adopt the settings detailed in Sec. 3.3.
2. *Visual Bonus*. This strategy introduces a supplementary reward bonus when the model achieves both correct answers and visual alignment with the ground truth. The reward is formulated as $R = \alpha S_{p,ans} + (1-\alpha) S_{p,fmt} + C$, where $C = 0.5$.
3. *Visual Mix*. This strategy integrates the visual score as a weighted component of the overall reward, formulated as $R = \alpha S_{p,vis} + \beta S_{p,ans} + \gamma S_{p,fmt}$, where $\alpha + \beta + \gamma = 1$. For our experiments, we set $\alpha = 0.2$, $\beta = 0.7$, and $\gamma = 0.1$.

This unified design facilitates a rigorous evaluation of different paradigms for integrating visual consistency with answer correctness within a single training pipeline. Figure 6 illustrates the training dynamics of the four aforementioned strategies, evaluated by the average length of responses and the accuracy on the vision-only subset of MathVerse.

Impact on Response Generation: As shown in Figure 6(a), the three strategies that incorporate the visual-aware score, namely **PaLMR**, Visual Mix, and Visual

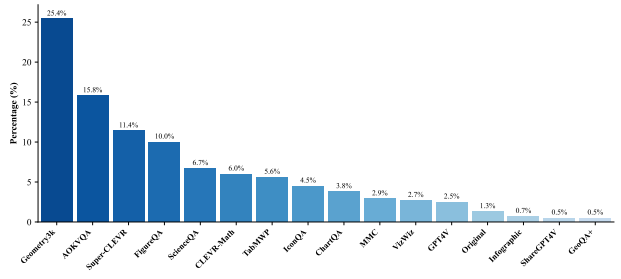


Figure 7. **Distribution of persistent errors across data sources on the training set.**

Bonus, consistently drive the model to generate longer responses than the baseline of vanilla GRPO. This observation indicates that providing rewards for visual alignment encourages the model to initiate the reasoning process with visual analysis, even in the absence of explicit prompts. Conversely, the baseline of GRPO tends to generate shorter responses that directly target the final answer.

Performance and Stability Trade-off: As shown in Figure 6(b), PaLMR maintains a stable, monotonically increasing accuracy on MathVerse. This stability stems from its hierarchical reward, which serves as a rigorous gating mechanism that requires coherent visual perception before rewarding the final answer ($S_{p,ans}$). In contrast, treating the visual score as a mere auxiliary objective (Visual Mix/Bonus) leads to volatile training, as models bypass visual grounding in favor of spurious, result-oriented solutions.

Table 3 highlights this critical trade-off: for base models with limited capacity, visual faithfulness and outcome correctness often diverge, especially on reasoning-heavy tasks. While PaLMR's strict visual gating enforces alignment to ensure superior robustness on comprehensive benchmarks (MMMU, MMStar), auxiliary objectives (Visual Mix/Bonus) exploit this divergence. They achieve marginally higher peak accuracy on specific math tasks (MathVerse) by bypassing visual verification.

4.5. Error Analysis and Reward Robustness

To understand the boundaries of PaLMR, we analyze its persistent training failures. As shown in Figure 7, over half of these errors originate from three datasets: Geometry3k [23] (25.4%), AOKVQA [35] (15.8%), and Super-CLEVR [20] (11.4%). Manual inspection categorizes these failures into two primary mechanisms:

First, **modality limitations** (e.g., Geometry3k): pure textual reasoning inherently struggles to capture complex, implicit spatial and geometric constraints. Second, **overly strict evaluation** (e.g., AOKVQA): exact-matching introduces false penalties for semantically valid but lexically distinct descriptions (e.g., "blue" vs. "cyan").

However, this strict exact-matching acts as a double-edged sword, serving as a **robust filter against label noise**. Synthetic datasets like Super-CLEVR and CLEVR-Math [21] frequently contain visually inconsistent ground truths. While the standard GRPO baseline hallucinates incorrect reasoning paths to overfit these flawed labels, PaLMR's intrinsic consistency gating resists this deceptive optimization pressure. By actively rejecting corrupted synthetic samples, PaLMR effectively preserves the mathematical and logical soundness of its outputs.

4.6. Limitations

While PaLMR establishes a robust foundation for process-aligned multimodal reasoning, we identify two primary directions for ongoing exploration. First, PaLMR's reasoning upper bound is constrained by the base model's foundational capabilities; future work could explicitly integrate auxiliary visual experts to improve this perceptual baseline. Second, our exact-matching reward inadvertently penalizes lexically distinct but semantically valid descriptions. Formulating robust, semantic-aware consistency metrics is necessary to reduce these false penalties.

5. Conclusion

We present PaLMR (Process Alignment for Multimodal Reasoning), a unified framework for faithful multimodal process-level reasoning. In contrast to previous reinforcement learning approaches that focus solely on final-answer correctness, PaLMR enforces alignment across perception, reasoning, and optimization, ensuring visually consistent and interpretable reasoning. The perception-aligned data construction pipeline, PaDLayer, programmatically generates multimodal datasets with structured and verifiable textual descriptions, providing a reliable basis for evaluating process-level faithfulness. Building on this, we introduce a process-aligned optimization layer, PaOLayer, instantiated with Vision-Guided GRPO (V-GRPO), which incorporates hierarchical perception-aware scores into the reward to promote visual consistency and reasoning coherence during reinforcement learning. Empirical results across multiple multimodal reasoning benchmarks show that PaLMR, trained with approximately 4.7K high-quality multimodal samples, consistently improves visual faithfulness and reasoning stability while maintaining competitive accuracy. These findings demonstrate that aligning the reasoning process itself, rather than optimizing only for outcome correctness, is essential for developing more reliable and interpretable multimodal large language models.

Acknowledgement

This work was supported by the National Natural Science Foundation of China Enterprise Innovation and Development Joint Fund Project U24B20181. The authors extend their sincere gratitude to Wenjie Qiu and Wenpo Song for their constructive discussions and continuous support throughout this project.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 6

[2] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 12

[3] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. 12

[4] Lang Cao, Renhong Chen, Yingtian Zou, Chao Peng, Wu Ning, Huacong Xu, Qian Chen, Yuxian Wang, Peishuo Su, Mofan Peng, et al. Process reward modeling with entropy-driven uncertainty. *arXiv preprint arXiv:2503.22233*, 2025. 3

[5] Q. Cao et al. Dreamprm: Domain-reweighted process reward model for multimodal reasoning. *arXiv preprint arXiv:2505.20241*, 2025. 3

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 12

[7] 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? *arXiv preprint arXiv:2403.20330*, 2024. 6

[8] Xinquan Chen, Bangwei Liu, Xuhong Wang, Yingchun Wang, and Chaochao Lu. Vrprm: Process reward modeling via visual reasoning. *arXiv preprint arXiv:2508.03556*, 2025. 2

[9] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 5, 6

[10] Kanzhi Cheng, Li YanTao, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025. 2

[11] Z. Cheng, Q. Chen, X. Xu, J. Wang, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*, 2025. 2

[12] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025. 2, 5, 6

[13] Lingxiao Du, Fanqing Meng, Zongkai Liu, Zhixiang Zhou, Ping Luo, Qiaosheng Zhang, and Wenqi Shao. Mm-prm: Enhancing multimodal mathematical reasoning with scalable step-level supervision. *arXiv preprint arXiv:2505.13427*, 2025. 3

[14] 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. 6

[15] 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. 1, 2

[16] Pengfei Hu, Zhenrong Zhang, Qikai Chang, Shuhang Liu, Jiefeng Ma, Jun Du, Jianshu Zhang, Quan Liu, Jianqing Gao, Feng Ma, et al. Prm-bas: Enhancing multimodal reasoning through prm-guided beam annealing search. *arXiv preprint arXiv:2504.10222*, 2025. 3

[17] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 2

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5, 6

[19] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 12

[20] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 8, 12

[21] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional lan-

guage, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 8, 12

[22] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1287–1310, Mexico City, Mexico, 2024. Association for Computational Linguistics. 12

[23] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 8, 12

[24] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 12

[25] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 2022. 2

[26] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 12

[27] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 12

[28] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 2, 6

[29] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 12

[30] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 12

[31] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 12

[32] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 2, 5, 6

[33] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536, 2020. 12

[34] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *AAAI Conference on Artificial Intelligence*, 2024. 2

[35] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, 2022. 2, 8, 12

[36] 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. 3

[37] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 12

[38] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 2

[39] Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. 2

[40] Lin Sun, Chuang Liu, Xiaofeng Ma, Tao Yang, Weijia Lu, and Ning Wu. Freeprm: Training process reward models without ground truth process labels. *arXiv preprint arXiv:2506.03570*, 2025. 3

[41] ByteDance Seed Team. verl: Volcano engine reinforcement learning for llms, 2025. 5

[42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5, 6

[43] Haoqin Tu, Weitao Feng, Hardy Chen, Hui Liu, Xianfeng Tang, and Cihang Xie. Vilbench: A suite for vision-language process reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. 3

[44] W. Wang et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025. 2, 3

[45] X. Wang et al. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 2

[46] Y. Wang et al. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025. 3

[47] 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. 2

[48] Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025. 4, 5

[49] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, 2023. 2

[50] Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. *arXiv preprint arXiv:2506.07218*, 2025. 5, 6

[51] 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*, 2025. 2

[52] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 4, 5

[53] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 2023. 2

[54] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025. 2

[55] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 6

[56] Hanning Zhang, Pengcheng Wang, Shizhe Diao, Yong Lin, Rui Pan, Hanze Dong, Dylan Zhang, Pavlo Molchanov, and Tong Zhang. Entropy-regularized process reward model. *arXiv preprint arXiv:2412.11006*, 2024. 3

[57] J. Zhang et al. Gm-prm: A generative multimodal process reward model for mathematical reasoning. *arXiv preprint arXiv:2508.04088*, 2025. 3

[58] J. Zhang et al. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025. 2

[59] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 2024. 6

[60] X. Zhang et al. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. 3

[61] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2

[62] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025. 1, 2

Appendix

A. Training Data Distribution

Table 4 delineates the domain-specific distribution of the filtered training dataset. Our filtering protocol was designed to remaining samples learnable for the reinforcement learning stage. First, because we employ a rule-based direct matching method to evaluate response correctness, tasks that inherently require open-ended or verbose optical character recognition (OCR) outputs were removed. Consequently, the representation of OCR-heavy datasets (e.g., DocVQA) is intentionally minimized. Furthermore, we filtered samples based on model solvability to prevent hallucinations on tasks exceeding the capacity of our base model, Qwen2.5-VL-7B. Specifically, we excluded instances from PlotQA (complex plot reasoning), VizWiz (fine-grained grounding for visually impaired users), and human-annotated MMC-Instruct that required advanced real-world chart understanding, etc. This ensures that the training data remains within the learnable manifold of the base model while maintaining diversity across geometry, science, and chart domains.

Category	Dataset	Number of Samples	Percentage (%)
Chart VQA	ChartQA[29]	209	4.42
	FigureQA[19]	348	7.36
	PlotQA[33]	2	0.04
	TabMWP[27]	483	10.22
	Subtotal	**1042**	**22.04**
GeoVQA	GeoQA+[3]	16	0.34
	Geometry3K[23]	852	18.02
	Subtotal	**868**	**18.36**
Science VQA	ScienceQA[26]	449	9.50
	Subtotal	**449**	**9.50**
Math VQA	CLEVR-Math[21]	549	11.61
	Super-CLEVR[20]	542	11.46
	IconQA[24]	350	7.40
	Subtotal	**1441**	**30.48**
OCR VQA	DocVQA[30]	4	0.08
	TextVQA[37]	7	0.15
	InfographicVQA[31]	6	0.13
	Subtotal	**17**	**0.36**
General VQA	A-OKVQA[35]	640	13.54
	VizWiz[2]	66	1.40
	MMC-Instruct[22]	49	1.04
	Original	119	2.52
	GPT-4V	32	0.68
	ShareGPT4V[6]	5	0.11
	Subtotal	**911**	**19.27**
Total		**4728**	

Table 4. Detailed Dataset Composition. The distribution represents the training data after applying solvability and evaluability filtering heuristics. The 'Subtotal' rows indicate the aggregate number of each domain.

B. Prompt Templates for PaLMR

Pseudo-Visual Ground Truth Generation. Obtaining large-scale, human-annotated visual descriptions as ground truth is too expensive. To address this, we leverage Gemini-2.5-Flash for its optimal trade-off between inference speed and visual perception capability. As illustrated in Figure 8, Gemini is prompted to generate structured descriptions of the visual content directly. Crucially, this generation is performed in a question-agnostic manner. By decoupling the description from the specific query, we ensure that the visual ground truth captures a comprehensive representation of the image states. To ensure scalability and computational efficiency, particularly for datasets where a single image is associated with multiple question-answer pairs (one-to-many mapping)

Prompt for Gemini2.5-Flash:
Please give a detail caption for this image, including all objects, attributes, and relationships, in no less than 150 words. Output them in list form.

Figure 8. Prompt template for Visual Ground Truth Generation. The prompt template used to instruct Gemini-2.5-Flash to generate structured, question-agnostic visual descriptions.

Visual-Aware Scoring Mechanism. To verify visual fidelity during the reasoning process, we implement an LLM-as-a-Judge metric on the training set. We employ Qwen3-30ba3b as the judge model to balance evaluation accuracy with computational efficiency. To align the reward signal with human judgment, we formulate the evaluation as a pairwise re-ranking task. As shown in Figure 9 and 10, the judge is provided with the question, the generated pseudo-visual ground truth (from Gemini), the current model rollout, and a pre-selected reference response from the base model. The judge evaluates the current rollout against the reference, explicitly conditioning its decision on the structured visual ground truth. This pipeline significantly improves the consistency of the scoring and the alignment rate with human preference.

Thinking prompt. We design our prompt template following the format in EasyR1, wherein the user prompt explicitly specifies the required output structure, including the use of `<think><\think>` and `\boxed{}` tags to separate the reasoning process and the final answer as in Figure 11. This prompt is appended to all queries for training samples, not set as a system prompt.

C. Implementation Details for Visual-Aware Scoring

Inference Optimization and Verdict Extraction. We observe that the thinking token used for PaLMR (e.g., `<think>`) will trigger Qwen3's thinking mode. While useful for complex reasoning, this introduces significant latency, rendering it computationally prohibitive for online re-

System prompt for Qwen3-30ba3b:
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Due to the question is a Visual reasoning question, your evaluation should consider which assistant's response have less mis-alignment to the Caption ground truth.
Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format:
" [[A]]" if assistant A is better, "[[B]]" if assistant B is better.

Figure 9. System prompt template for Visual-Aware Scoring. The LLM-as-a-Judge prompt used by Qwen3-30ba3b to perform pairwise ranking between the model rollout and a reference response, conditioned on the pseudo-visual ground truth.

User prompt for Qwen3-30ba3b:
[The start of User Question]
QUESTION
[The End of User Question]

[The Start of Caption Ground Truth]
PSEUDO VISUAL GROUND TRUTH
[The End of Caption Ground Truth]

[The Start of Assistant A's Answer]
MODEL RESPONSE 1
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
MODEL RESPONSE 2
[The End of Assistant B's Answer]

Figure 10. User prompt template for Visual-Aware Scoring.

Prompt for Policy Model:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed.

Figure 11. Prompt template for Policy Model.

inforcement learning. To solve this problem, we remove all `<think>` and `<\think>` tokens during the judge's generation phase.

Furthermore, despite prompt instructions requesting a structured format, large language models occasionally produce free-form text containing multiple potential verdict tokens. To address this, we adopt a robust extraction method similar to those employed in multimodal benchmarks like MMMU. Specifically, we utilize regular expressions to identify all valid verdict candidates within the response. We select the *last* matching token as the final decision, exploiting the tendency of instruction-tuned models to output their definitive conclusion at the end of the reasoning trace.

Mitigating Positional Bias via Input Shuffling. LLM-as-a-Judge metrics are susceptible to positional bias, where the model exhibits a preference for options presented in specific locations (e.g., "Option A" vs. "Option B") regardless of content quality. To ensure accuracy, we implement a stochastic rolling strategy for input construction. For each evaluation pair consisting of the model's current rollout and the reference response, we randomize their assignment to the first or second position in the prompt context. The judge's output is then dynamically mapped back to the corresponding source (model or reference) to derive the final visual fidelity score.

D. Visualized Examples for Training Dataset

Figures 12 and 13 present qualitative examples from our training dataset alongside the inference outputs of the trained PaLMR model. In these visualizations, red text denotes incorrect reasoning or hallucinations, whereas green text highlights visually accurate observations relevant to the user's query.

First, we observe that the pseudo-visual ground truth synthesized by Gemini-2.5-Flash consistently provides precise descriptions of visual content, capturing details essential for the reasoning process. This reliability validates the use of pseudo visual GTs within our visual-aware scoring framework as a robust metric for assessing visual faithfulness. Furthermore, while the base model suffers from visual inconsistencies—such as inaccurate data extraction in ChartQA and object recognition failures in ClevrMath—PaLMR effectively rectifies these hallucina-

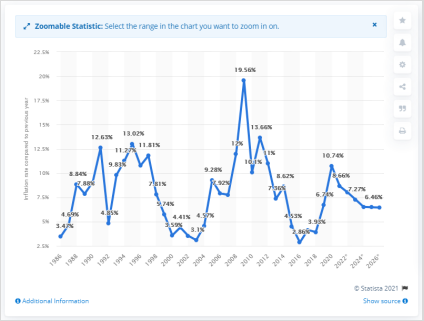
ChartQA

Question: What is Pakistan's inflation rate forecast to level off at over the next few years?
Answer: 6.46
Pseudo GT:
- Here is a detailed caption for the image:
- * **Overall Image:** The image displays a line graph titled "Zoomable Statistic: Select the range in the chart you want to zoom in on," presenting historical and projected inflation rates compared to the previous year, from 1986 to 2026. The graph is overlaid on a white background, with social media icons visible on the right sidebar.
- * **Graph Title:** A blue bar at the top of the graph contains the text "Zoomable Statistic: Select the range in the chart you want to zoom in on."
- * **Icon:** An icon resembling an open box with arrows pointing outwards is present next to the text "Zoomable Statistic."
- * **Close Icon:** A white 'x' symbol is located on the far right of this blue bar, likely to close the statistic.
- * **Y-Axis:** Represents the "Inflation rate compared to previous year."
- * **Labels:** Percentage values are aligned vertically along the left side, starting from 2.5% and increasing in 2.5% increments to 22.5%. The specific labeled values are 2.5%, 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, and 22.5%.
- * **X-Axis:** Represents the "Year."
- * **Labels:** Years are aligned horizontally along the bottom, starting from 1986 and increasing in two-year increments to 2026. The years 2022, 2024, and 2026 are marked with an asterisk (*), indicating they are projected values.
- * **Gridlines:** Faint gray horizontal and vertical gridlines create a structured background for the data points, aiding in reading specific values.
- * **Line Graph Data:** A prominent blue line with circular markers indicates the inflation rate for each year. Each marker is accompanied by its corresponding percentage value.
- * **Data Points (Year, Inflation Rate):**
- * 1986: 3.47%

…….

- * 2009: 13.66%
- * 2010: 11%
- * 2011: 8.62%
- * 2012: 7.36%
- * 2013: 4.53%
- * 2014: 2.86% (Lowest Point)
- * 2015: 3.93%
- * 2018: 6.74%
- * 2019: 10.74%
- * 2020: 8.66%
- * 2021: 7.27%
- * 2022*: 6.74% (Projected)
- * 2023*: 6.46% (Projected)

……



Reference Response:
To provide a forecast for Pakistan's inflation rate, I need to consider recent trends, economic policies, and potential future developments. Here's an internal monologue leading to a hypothetical answer:

1. **Current Inflation Trends**: Pakistan's inflation rate has fluctuated significantly in recent years, reaching highs around the 10% mark in 2020 due to global supply chain disruptions and local economic measures to combat the COVID-19 pandemic.

2. **Economic Policies and Measures**: The Pakistani government has implemented a number of economic reforms and measures aimed at controlling inflation. This includes interest rate adjustments, fiscal discipline, and efforts to stabilize the exchange rate.

3. **Global Economic Conditions**: The global economy is still recovering from the pandemic, with uncertainties surrounding geopolitical tensions and trade tensions. These factors could influence global commodity prices and, consequently, Pakistan's inflation rate.

4. **Domestic Economic Conditions**: Domestic economic conditions, such as agricultural production, industrial growth, and domestic demand, are also crucial in determining inflation levels. If there's a significant improvement in these areas, it could help moderate inflation.

5. **Inflation Forecasts from Institutions**: According to recent economic reports and forecasts by organizations like the International Monetary Fund (IMF), the World Bank, and local financial institutions, Pakistan's inflation rate is expected to gradually reduce. These reports suggest a steady decline over the next few years, although the pace and specifics may vary.

6. **Final Forecast Adjustment**: Considering the above factors, a reasonable forecast for Pakistan's inflation rate to level off over the next few years is likely to be in the range of 6-8%. This forecast accounts for gradual improvements in economic conditions and global factors.

Given this reasoning, here is the final answer:

\boxed{6-8%}

PaLMR Response:
<think>
To answer the question about Pakistan's inflation rate forecast to level off at over the next few years:
Identify the trend near the end of the provided data:
- The line graph shows recent years (2018-2026) with specific annualized inflation rates marked.
- Key years near the graph's end-in-focus area (2024*, 2026*) are shown:
  In 2024*, the inflation rate is forecast to be **7.27%**.
  In 2026*, it's forecast slightly lower at **6.46%**.

Look for a flattening or stabilization:
- The values in the later years exhibit a declining trend but are close or showing a slow rate of further decline: 2024: 7.27%, 2026: 6.46%.
- This suggests the inflation rate is trending downward but relatively stable around approx. 6-7% in the forecast period.

Consider the trend line (visual or logical extrapolation):
Graph suggests:
- Rates around 2026 (last focused data point):
- Slightly below 2024 and consistent/no sharp pronounced rise after 7.27%.
- This implies leveling off and leveling slightly downward toward 6.46%.
- The provided specific projected/highest estimated rate toward the near future on this trend, implying a likely near-term outlook (forecast usually targets around observable or predicted trends):
- 6.46% appears stable/finalized (lowest, and without any marked sharp rise in trend):
- The figures give us the forecast stable range, and the closest most centrally stable (median-high) short-term inflation rate forecast toward the given years is **6.46%**, closest aligned across projected 2026 and near future given visible gradual taper.
- Thus, the inflation rate is forecast to level off in the mid-6-7% range (~6.46%) over the next few years (calibrated to the visible trend and given the lowest late-point projection focused on lower years 2024* and 2026*).

</think>
\boxed{6.46}

Figure 12. Visualization of training set.

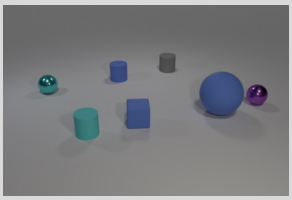
tions through visual guidance. It ultimately yields responses that are both visually perceptive and logically correct.

E. Computational Cost and Latency Analysis

We analyze the computational overhead of our proposed PaLMR framework compared to the standard GRPO baseline. Table 5 details the per-step training latency for Qwen2.5-VL-7B on a node with 8 GPUs, which are allocated symmetrically (4 GPUs for the policy model and 4 for the reward model).

By serving the judge model (Qwen3-30A3B) via SGLang, PaLMR introduces a 37.5% latency overhead relative to the baseline. To investigate this cost, we conducted

ClevrMath



Question: How many gray matte things are left after adding 1 gray matte thing?
Answer: 2
Pseudo GT:
Here is a detailed caption for the image:
* **Scene Description:** The image depicts an assortment of geometric primitive shapes rendered on a flat, light gray surface under ambient lighting conditions, casting subtle shadows to the left and slightly behind each object. The background is a uniform light gray, blending seamlessly with the surface.
* **Objects and Attributes:**
  * **Large Blue Sphere:** A prominent matte blue sphere, positioned in the middle-right of the frame, appearing largest among all objects.
  * **Small Purple Metallic Sphere:** A smaller, highly reflective metallic purple sphere located to the right and slightly behind the large blue sphere.
  * **Small Teal Metallic Sphere:** A small, highly reflective metallic teal sphere situated in the far left of the frame.
  * **Teal Cylinder:** A matte teal cylinder placed in the bottom-left quadrant of the image.
  * **Light Blue Cylinder:** A matte light blue cylinder positioned in the top-middle of the frame.
  * **Gray Cylinder:** A matte gray cylinder located in the top-right quadrant of the image.
  * **Light Blue Cube:** A matte light blue cube situated in the central-bottom portion of the frame, between the teal cylinder and the large blue sphere.
* **Spatial Relationships:**
  * The **teal metallic sphere** is positioned to the left of all other objects.
  * The **teal cylinder** is below and slightly to the right of the teal metallic sphere.
  * The **light blue cylinder** is above and slightly to the right of the teal cylinder.
  * The **light blue cube** is to the right of the teal cylinder and below the light blue cylinder.
  * The **gray cylinder** is above and to the right of the light blue cube, and to the right of the light blue cylinder.
  * The **large blue sphere** is to the right of the light blue cube and below the gray cylinder.
  * The **purple metallic sphere** is to the right and slightly behind the large blue sphere, appearing closer to the right edge of the frame.

* **Overall Composition:** The objects are disbursed across the surface, creating an impression of scattered, distinct shapes rather than a deliberate arrangement. The varying colors and finishes (matte vs. metallic) add visual interest to the minimalist composition.

Reference Response:
First, let's think through the question step-by-step:

1. The question asks about gray matte objects.
2. It mentions adding 1 gray matte object.
3. We need to determine how many gray matte objects remain after this addition.

If we start with zero gray matte objects (as there is no initial count given, and the question asks specifically about the state "after adding 1 gray matte thing"), then adding 1 gray matte object means we now have 1 gray matte object in total.

Therefore, the final answer is:

\boxed{1}

PaLMR Response:
<think>
To answer the question, let's break it down step by step:
Count the current number of gray matte objects:
- In the image, there is 1 gray matte cylinder already present.
2. **Consider the addition:**
- The question asks how many gray matte objects there will be after adding 1 more gray matte object.
3. **Calculate the total after adding:**
- Currently, there is 1 gray matte object.
- Adding 1 more gray matte object results in: (1 + 1 = 2).
Reasoning Breakdown:
- Initially, there is 1 gray matte object in the image.
- Adding 1 more gray matte object results in a total of 2 gray matte objects.
</think>
Final Answer:\boxed{2}"

Figure 13. Visualization of training set.

an ablation study by disabling the Chain-of-Thought (CoT) generation in the judge model (PaLMR w/o CoT). While this nearly eliminates the computational overhead (+0.7%), it substantially degrades downstream performance, reducing the MMMU score from 59.3 to 57.3 and dropping human alignment to approximately 60%. These findings demonstrate that the additional computational cost incurred by CoT is a necessary trade-off. The CoT process is essential for providing stable, low-noise reward signals that effectively mitigate judge model bias, thereby enabling superior reasoning capabilities.

Method	Latency (s)	Overhead	MMMU
Baseline (GRPO)	397	+0%	57.8
PaLMR (w/o CoT)	400	+0.7%	57.3
PaLMR (Ours)	**546**	**+37.5%**	**59.3**

Table 5. Comparison of per-step training latency, computational overhead, and downstream performance (MMMU). Evaluated on Qwen2.5-VL-7B using 8 GPUs.