

Appendix

A. Training Data Distribution

Table 1 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[12]	209	4.42
	FigureQA[4]	348	7.36
	PlotQA[15]	2	0.04
	TabMWP[11]	483	10.22
	Subtotal	1042	22.04
GeoVQA	GeoQA+[2]	16	0.34
	Geometry3K[8]	852	18.02
	Subtotal	868	18.36
Science VQA	ScienceQA[10]	449	9.50
	Subtotal	449	9.50
Math VQA	CLEVR-Math[6]	549	11.61
	Super-CLEVR[5]	542	11.46
	IconQA[9]	350	7.40
	Subtotal	1441	30.48
OCR VQA	DocVQA[13]	4	0.08
	TextVQA[17]	7	0.15
	InfographicVQA[14]	6	0.13
	Subtotal	17	0.36
General VQA	A-OKVQA[16]	640	13.54
	VizWiz[1]	66	1.40
	MMC-Instruct[7]	49	1.04
	Original	119	2.52
	GPT-4V	32	0.68
	ShareGPT4V[3]	5	0.11
	Subtotal	911	19.27
	Total	4728	

Table 1. 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 1, 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 1. 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 2 and 3, 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 4. 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 2. 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 3. 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 4. 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 5 and 6 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)

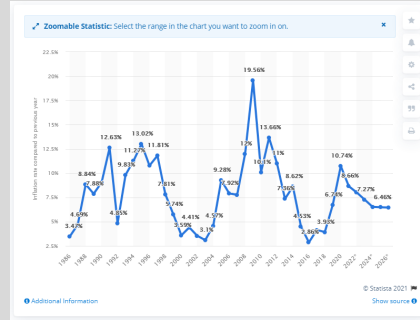
Reference Response:

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:

- 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.
- 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.
- 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.
- 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.
- 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.
- 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 5. Visualization of training set.

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

E. Computational Cost and Latency Analysis

We analyze the computational overhead of our proposed PaLMR framework compared to the standard GRPO base-

line. Table 2 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.

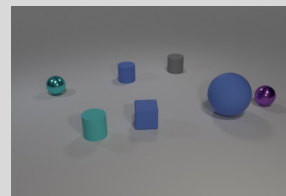
The **light blue cube** is to the right of the teal cylinder and below the light blue cylinder.

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

References

- [1] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 1
- [2] 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. 1
- [3] 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. 1
- [4] 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. 1
- [5] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Ko-

- rtylewski, 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. 1
- [6] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 1
- [7] 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. 1
- [8] 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. 1
- [9] 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. 1
- [10] 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. 1
- [11] 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. 1
- [12] 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. 1
- [13] 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. 1
- [14] 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. 1
- [15] 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. 1
- [16] 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. 1
- [17] 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. 1