A Training Details

Datasets

Text-Only Dataset: We use DeepScaleR-Preview [1] as the text-only math reasoning dataset to train FRE-Text. This dataset is composed of AIME(American Invitational Mathematics Examination) problems (1984-2023), AMC (American Mathematics Competition) problems (prior to 2023), Omni-MATH [2] and STILL [3]. It is filtered to keep those that have verifiable answers and to remove redundant problems, leaving 40.3k high-quality data points.

VerMulti-65K: We use the MathV360K [4] dataset, which contains 360k multimodal problems with diverse domains, including arithmetic, geometry, science, and more. The dataset is further filtered, saving prompts that have verifiable numeric answers (*e.g.*, 123, 4.11) or option answer (*e.g.*, A, B, C, D). Finally, we get 130k multimodal problems. To reduce the cost of training, we randomly sample 65k data from the filtered dataset for rule-based RL training. The detailed statistics of VerMulti are shown in Tab. A1.

We exclude samples that are not objectively verifiable for rule-based reward. Typical rejected cases include: (i) open-ended or descriptive answers (*e.g.*, "Describe the trend in the figure," "What can you infer from the chart?"); (ii) problems with multiple valid solutions where a single canonical answer cannot be determined; (iii) free-form text or paragraph answers that do not reduce to a numeric value or option label. Retaining only numeric or option-style answers ensures that the accuracy reward can be computed deterministically.

Split	Count	Split	Count
IconQA [5]	7166	TQA [6]	2130
PMC-VQA [7]	6760	DocVQA [8]	1974
TabMWP [9]	6732	TextVQA [10]	1462
A-OKVQA [11]	6185	VQA2.0 [12]	1316
FigureQA [13]	4995	ChartQA [14]	1115
ScienceQA [15]	4243	PlotQA [16]	1020
GeoQA+ [17]	4062	Super-CLEVR [18]	1016
DVQA [19]	3317	VQA-AS [20]	404
Geometry3K [21]	2845	MapQA [22]	278
UniGeo [23]	2767	VizWiz [24]	204
AI2D [11]	2603	GEOS [25]	129
CLEVR-Math [26]	2393	VQA-RAD [27]	2

Table A1: Statistics of VerMulti.

VerMulti-Geo15K: We use 15k geometry problems from MathV360K [4] to train MGT-Geo, which consists with GEOS [25], Geometry3K [21], GeoQA+ [17], UniGeo [23], and TQA [6]. The distribution of data is shown in Tab. A2.

Sokoban Training Datasets: For the Sokoban domain, we generated a diverse set of 11,500 environments by using gym-sokoban [28] with varying difficulty levels to train MGT-Sokoban. Specifically, we created 5,000 environments using the sokoban-small-v0 difficulty setting and

another 5,000 environments with sokoban-small-v1 difficulty. To ensure the model encounters a wide range of scenarios, we further supplemented the dataset with 500 environments each from sokoban-v0, sokoban-v1, and sokoban-large-v0 difficulty settings. This distribution allows the model to learn planning strategies across different board sizes and complexity levels, from simpler small-scale puzzles to more challenging large configurations.

Split	Count
GEOS	271
GeoQA+	8155
Geometry3K	2776
TQA	25
UniGeo	5583

Table A2: Statistics of VerMulti-Geo15K.

Training Hyper-Parameters

Our training infrastructure is built upon OpenRLHF [29] and we use the same set of hyper-parameters for RL, as shown in Tab. A5. All training tasks use 8 A100 GPUs and 1 A100 GPU for evaluation.

We train SFT-Text 2 epochs with batch-size=256, learning-rate=5e-6, warmup-ratio=0.03.

Training was conducted on 8 NVIDIA A100 GPUs, taking approximately 36 hours for the complete pipeline (FRE plus one MGT variant).

B Detailed Performance on MM-Star

To validate the improvement of our model on general visual tasks, we also provide the MM-Star different split results in Tab. A3. The results show that MGT can improve the perception ability of FRE-Text.

Model	**Coarse**	**Fine-grained**	**AVG**
QwenVL-2.5 CoT	68.56	47.52	58.04
FRE-Multi	69.32	52.05	60.68
FRE-Text	65.20	46.30	55.75
MGT-PerceReason	67.99	48.19	58.09

Table A3: Performance comparison (%) on different perception categories in MM-Star benchmark. Coarse Perception includes tasks requiring basic object recognition, while Fine-grained Perception involves more detailed visual discrimination tasks.

On the Coarse Perception subset (basic object recognition), MGT-PerceReason (67.99%) remains close to the baseline (68.56%), indicating that reasoning enhancement does not come

Table A4: Performance (%) of different model sizes on representative benchmarks. "MM-AVG" is the mean of MathVerse, MathVision, and MathVista. Reported scores for MM-Eureka-Qwen-7B are taken from its original paper.

Model	MathVerse	MathVision	MathVista	MM-AVG	MATH500
			7B-size		
Qwen2.5-VL-7B (Baseline)	45.41	26.51	68.60	46.84	62.20
FRE-Qwen2.5-VL-7B (Ours)	47.08	**30.10**	69.50	48.89	**73.00**
MGT-Qwen2.5-VL-7B (Ours)	**50.91**	28.55	71.80	**50.42**	68.00
			2B-size		
Qwen2-VL-2B (Baseline)	21.04	11.64	49.60	27.43	10.00
FRE-Qwen2-VL-2B (Ours)	21.88	13.45	50.60	28.64	**19.80**
MGT-Qwen2-VL-2B (Ours)	**22.96**	**14.64**	**52.50**	**30.03**	18.00

at the cost of simple visual perception and the model does not exhibit "over-thinking" on trivial queries.

C Vision-Only Subset and the Role of FRE

On the MathVerse "Vision Only" subset, FRE-Text (25.00%) drops relative to the baseline (28.43%), yet it provides the reasoning foundation on which MGT builds. MGT-Geo recovers and surpasses both, reaching 36.68% (+11.68% over FRE-Text, +8.25% over baseline). This suggests that the enhanced reasoning from FRE may help the model attend to and integrate visual features more effectively during MGT, rather than the gain being due to domain data exposure alone.

D Extended Results on Larger and Smaller LMMs

To examine the scalability of LMM-R1, we reproduced the complete training pipeline on models that are both larger and smaller than the 3B backbone used in the main paper. For the larger setting we fine-tuned Qwen2.5-VL-7B-Instruct, and for the compact setting we used Qwen2-VL-2B-Instruct. We further compare the 7B variant with the concurrent SOTA method MM-Eureka-Qwen-7B. The scores in table A4 demonstrate that our two-stage procedure consistently improves performance across modalities and scales.

LMM-R1 delivers notable gains on Qwen2.5-VL-7B, improving the MathVerse/MM-AVG/MATH500 metrics while matching or surpassing MM-Eureka-Qwen-7B on two of the three multimodal benchmarks. The performance lift on the lightweight 2B model verifies that the framework remains effective even when model capacity is severely constrained. These results support the claim that LMM-R1 generalizes beyond a single backbone and can serve as a scalable recipe for multimodal reasoning enhancement.

E Code and dataset license

Our codebase is built upon OpenRLHF [29] and follows its Apache License 2.0 (Apache-2.0). Our dataset is derived from MathV360K [4] and inherits its license terms. We maintain proper attribution to both original works and comply with their respective licensing requirements.

- For code (based on OpenRLHF):
  - Licensed under Apache-2.0
  - Requires maintaining copyright notices and license
  - Requires proper attribution to OpenRLHF
  - Any modifications are clearly documented
- For dataset (based on MathV360K):
  - Follows MathV360K's original license terms
  - Maintains proper attribution to the original dataset
  - Any modifications or derivatives are clearly documented

The complete license information and detailed terms can be found in our repository.

F Evaluation Details

Text-Only Benchmarks

We use LightEval [30] to evaluate our models on text-only mathematical reasoning benchmarks. LightEval is a lightweight evaluation framework that provides standardized evaluation protocols for language models. For our evaluation, we focus on two challenging mathematical reasoning benchmarks: MATH-500 [31] and GPQA-Diamond [32].

Evaluation Protocol. We implement custom evaluation tasks within the LightEval [30] framework to ensure consistent and reproducible evaluation. For each benchmark, we define specific prompt templates that encourage step-by-step reasoning and clear answer formatting, as shown in Appendix H. The evaluation is conducted with a maximum generation length of 32,768 tokens to accommodate extensive reasoning chains.

Metrics. For MATH-500, we employ a specialized metric that extracts answers from LaTeX expressions, with priority given to boxed answers. This metric uses multilingual extractive matching with a precision of 5 decimal places. For GPQA-Diamond, we use a letter-based extraction metric that identifies the selected multiple-choice option (A, B, C, or D) from the model response.

Both metrics are implemented using LightEval's multilingual extractive match framework, which provides robust answer extraction capabilities even when models deviate slightly from the requested output format. This approach ensures fair comparison across different model architectures and response styles.

	FRE-Text	FRE-Multi	MGT-Geo	MGT-PerceReason	MGT-Sokoban
train_batch_size	128	256	256	256	128
rollout_batch_size	128	256	256	256	128
temperature	1	1	1	1	1
n_samples_per_prompt	16	16	16	16	16
max_epochs	1	1	1	1	1
num_episodes	1	2	2	2	4
generate_max_len	8192	8192	8192	8192	8192
init_kl_coef	1e-3	1e-3	1e-3	1e-3	0.0
lambd	1	1	1	/	1
gamma	1	1	1	/	1
actor_learning_rate	1e-6	1e-6	4e-7	4e-7	1e-6
critic_learning_rate	9e-6	9e-6	9e-6	/	5e-6
warm-up ratio	0.03	0.03	0.03	0.03	0.03

Table A5: Training Hyper-parameters.

Multimodal Benchmark

We use LMMs-Eval [33] for multimodal benchmark evaluation, which is a comprehensive and lightweight evaluation toolkit for LMMs. We evaluate our models in representative challenging multimodal reasoning benchmarks, including MathVision [34], MathVerse [35], MathVista [36], OlympiadBench [37] and MM-Star [38].

Evaluation Protocol. We integrate vLLM [39] into LMMs-Eval and deploy our models on vLLM for efficient evaluation. To accurately extract the answer of the model for each test case, we deploy Qwen2.5-14B-Instruct [40] as a judge, using the prompt shown in Appendix H.

The system prompts of our models used in training and evaluation are the same, which are also listed in Appendix H. Furthermore, we find that using the same system prompt for the baseline model degrades its performance. Thus, for fairness, we do not use our system prompt for the baseline model.

Agent-Related Benchmark

We follow the MageBench [41] to evaluate our model in Sokoban and football tasks.

Agent Formulation. We use two agent designs from MageBench [41] for our evaluations: **Global Planner Agent:** This agent observes the initial environment once and plans all subsequent actions without further observations. Its decision process is:

$$\pi_\theta(p_{sys}, p_{task}, p_{cot}, p_{io}) \rightarrow a_1, a_2, ..., a_T \tag{A1}$$

Online Planner Agent: This agent analyzes each step and acts based on previous actions

(AM) and observations (OM). We set AM = 5 and OM = 1. Its decision process is:

$$\pi_\theta(p_{sys}, a_{t-AM:t}, o_{t-OM:t}, p_{cot}, p_{io}) \rightarrow a_{t+1} \tag{A2}$$

Task Descriptions.

- **Sokoban-Global:** The model sees the initial puzzle once and plans all moves in advance. This tests the model's ability to mentally simulate the environment.
- **Football-Online:** The model controls a player in a dynamic environment, making decisions based on changing game states. This tests real-time decision-making in complex scenarios.

G Training Curves

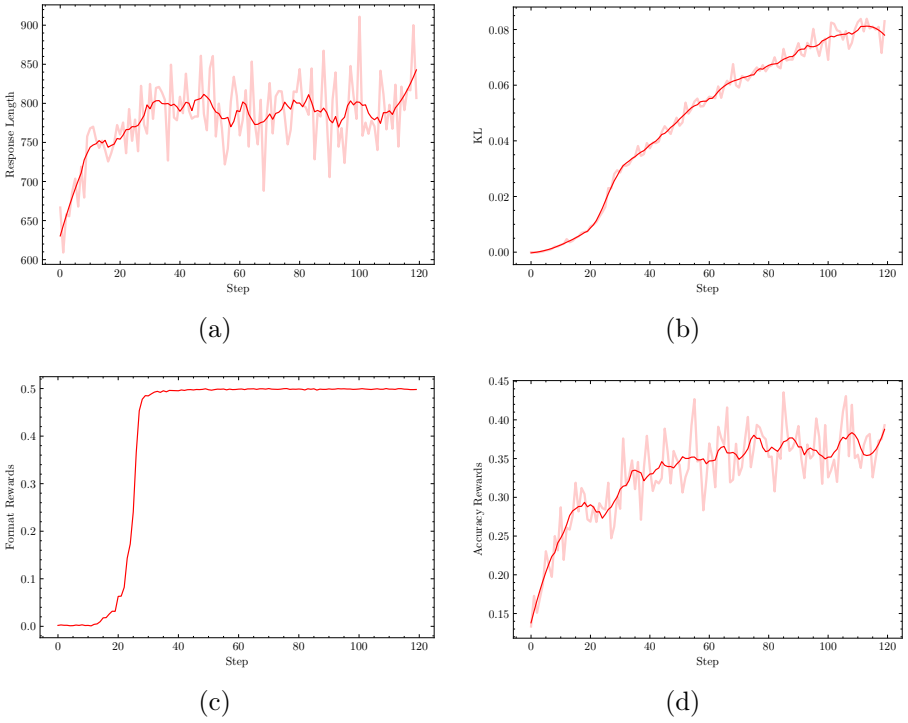
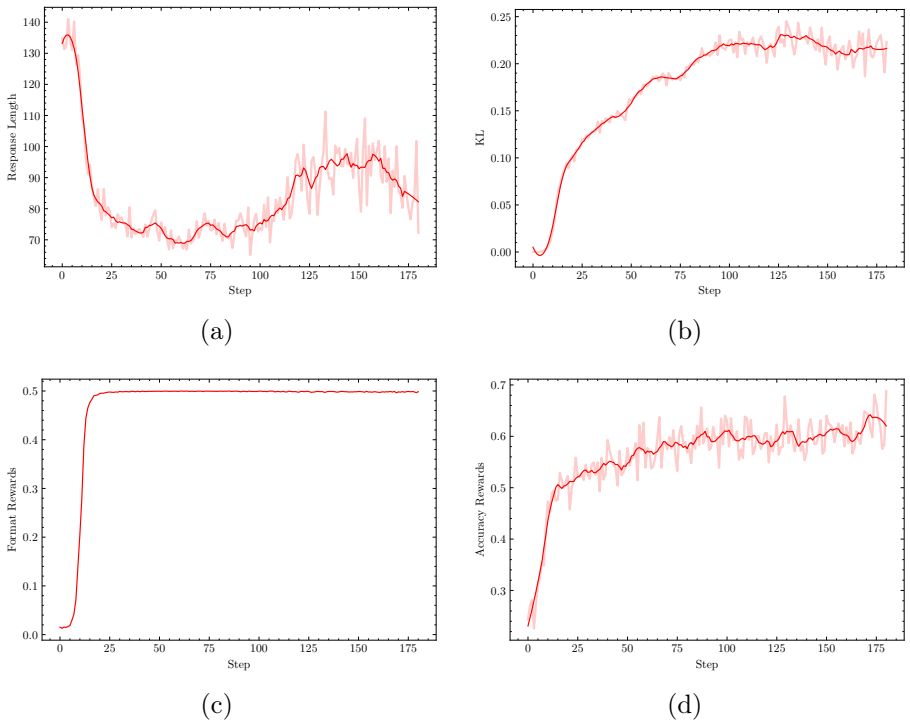
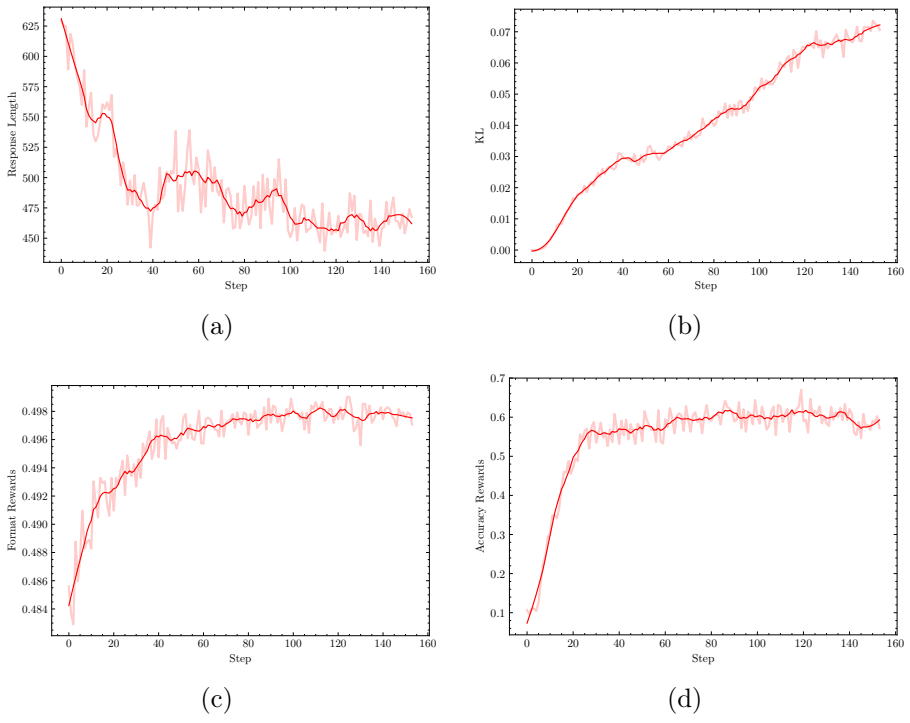


Figure A1: Training Curve of FRE-Text.



(a) (b) (c) (d)

Figure A2: Training Curve of FRE-Multi.



(a) (b) (c) (d)

Figure A3: Training Curve of MGT-Geo.

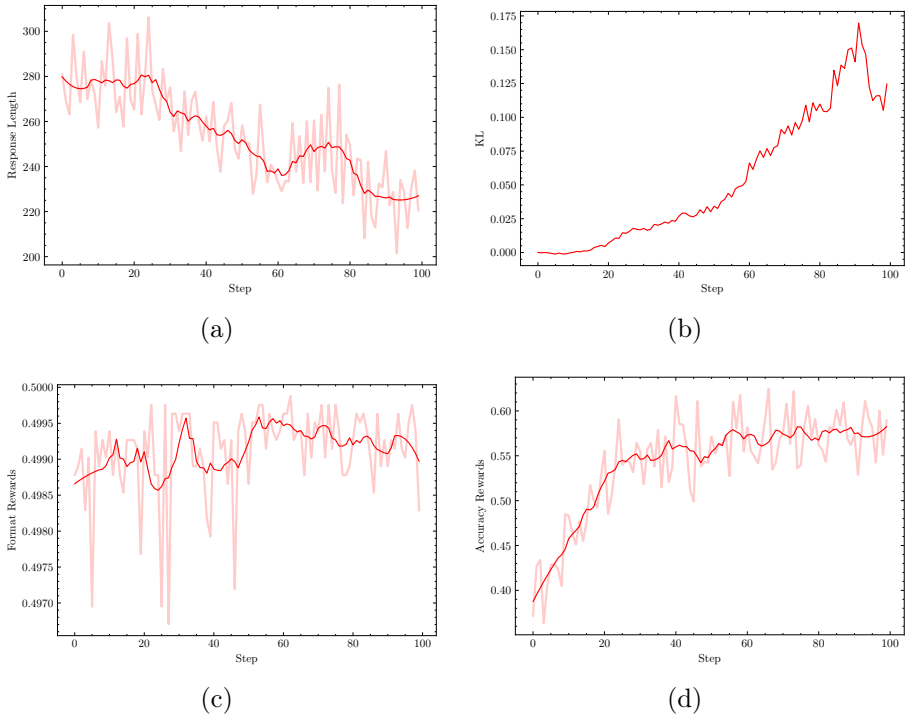


Figure A4: Training Curve of MGT-PerceReason.

H Prompts for Training and Evaluation

System Prompts for Training and Evaluation

```
You are a helpful assistant good at solving math problems with step-by-step
↪ reasoning. You should first thinks about the reasoning process in the mind
↪ and then provides the user with the answer. Your answer must be in latex
↪ format and wrapped in $...$.The reasoning process and answer are enclosed
↪ within <think> </think> and <answer> </answer> tags, respectively, i.e.,
↪ <think> Since $1+1=2$, so the answer is $2$. </think><answer> $2$
↪ </answer>, which means your output should start with <think> and end with
↪ </answer>.
```

Prompt Templates for Text-Only Benchmarks

MATH-500 Prompt Template:

```
Solve the following math problem efficiently and clearly. The last line of your
↪ response should be of the following format: 'Therefore, the final answer is:
↪ $\boxed{ANSWER}$. I hope it is correct' (without quotes) where ANSWER is just
↪ the final number or expression that solves the problem. Think step by step
↪ before answering.

{Question}
```

GPQA-Diamond Prompt Template:

```
Answer the following multiple choice question. The last line of your response
↪  should be of the following format: 'Answer: $LETTER' (without quotes) where
↪  LETTER is one of ABCD. Think step by step before answering.

{Question}

A) {A}
B) {B}
C) {C}
D) {D}
```

Prompt of Answer Extraction for Qwen2.5-14B-Instruct

```
Please read the following example. Then extract the answer from the model
↪  response and type it at the end of the prompt.

Hint: Please answer the question requiring an integer answer and provide the
↪  final value, e.g., 1, 2, 3, at the end.
Question: Which number is missing?

Model response: The number missing in the sequence is 14.

Extracted answer: 14

Hint: Please answer the question requiring a floating-point number with one
↪  decimal place and provide the final value, e.g., 1.2, 1.3, 1.4, at the end.
Question: What is the fraction of females facing the camera?

Model response: The fraction of females facing the camera is 0.6, which means
↪  that six out of ten females in the group are facing the camera.

Extracted answer: 0.6

Hint: Please answer the question requiring a floating-point number with two
↪  decimal places and provide the final value, e.g., 1.23, 1.34, 1.45, at the
↪  end.
Question: How much money does Luca need to buy a sour apple candy and a
↪  butterscotch candy? (Unit: $)

Model response: Luca needs $1.45 to buy a sour apple candy and a butterscotch
↪  candy.

Extracted answer: 1.45

Hint: Please answer the question requiring a Python list as an answer and
↪  provide the final list, e.g., [1, 2, 3], [1.2, 1.3, 1.4], at the end.
Question: Between which two years does the line  graph saw its maximum peak?

Model response: The line graph saw its maximum peak between 2007 and 2008.
```

```
Extracted answer: [2007, 2008]

Hint: Please answer the question and provide the correct option letter, e.g.,
↪  A, B, C, D, at the end.
Question: What fraction of the shape is blue?\nChoices:\n(A) 3/11\n(B)
↪  8/11\n(C) 6/11\n(D) 3/5

Model response: The correct answer is (B) 8/11.

Extracted answer: B

Hint: Please answer the question and provide the correct option letter, e.g.,
↪  A, B, C, D, at the end.
Question: {question}

Model response: {response}

Extracted answer:
```

I Case Study

Case Study of Qwen2.5-VL CoT and FRE-Text

Question: Find the sum of all real solutions to the equation $\frac{x-2}{x^2+4x+1} = \frac{x-5}{x^2-10x}$.

Qwen2.5-VL CoT:

```
Rearrange equation: divide by x^2 and let's use partial fractions to simplify
↪  the equation.
To simplify, write each fraction as a difference: $1 +
↪  \frac{x-4}{(x+2)(x+2)(x-5)} = 1 + \frac{(x-2) - (x-5)}{x^2-10x}$.
3. 2x^2 - 4x - 16 = 0.
4. Add 16 to both sides: 2x^2 - 4x = 0
5. Divide by 2: x^2 - 2x = 0.
6. Take out x: x(x-2) = 0. Elie: x = 0 or x = 2.
Therefore, the sum of all real solutions to the equation is x = 0 + 2 =
↪  $\boxed{2}$.
```

FRE-Text:

```
<think>
To solve the equation \(\frac{x-2}{x^2+4x+1} = \frac{x-5}{x^2-10x}\), we first
↪  eliminate the denominators by cross-multiplying:

\[
(x-2)(x^2-10x) = (x-5)(x^2+4x+1)
\]
```

```
Expanding both sides, we get:

\[
x^3 - 10x^2 - 2x^2 + 20x = x^3 + 4x^2 + x - 5x^2 - 20x - 5
\]

Simplifying both sides, we have:

\[
x^3 - 12x^2 + 20x = x^3 - x^2 - 19x - 5
\]

Subtracting \(x^3\) from both sides, we get:

\[
-12x^2 + 20x = -x^2 - 19x - 5
\]

Rearranging all terms to one side, we have:

\[
-12x^2 + x^2 + 20x + 19x + 5 = 0
\]

Simplifying further, we get:

\[
-11x^2 + 39x + 5 = 0
\]

This is a quadratic equation in the form \(ax^2 + bx + c = 0\), where \(a =
↪  -11\), \(b = 39\), and \(c = 5\). The sum of the solutions of a quadratic
↪  equation \(ax^2 + bx + c = 0\) is given by \(-\frac{b}{a}\). Therefore, the
↪  sum of the solutions is:

\[
-\frac{39}{-11} = \frac{39}{11}
\]

We need to check if the solutions satisfy the original equation and the
↪  denominators. The solutions to the quadratic equation are:

\[
x = \frac{-39 \pm \sqrt{39^2 - 4(-11)(5)}}{2(-11)} = \frac{-39 \pm \sqrt{1521 +
↪  220}}{-22} = \frac{-39 \pm \sqrt{1741}}{-22}
\]

The solutions are:

\[
x = \frac{39 \mp \sqrt{1741}}{22}
```

```
\]

We need to ensure that these solutions do not make the denominators zero. The
↪  denominators are \(x^2 + 4x + 1\) and \(x^2 - 10x\). Substituting \(x =
↪  \frac{39 \mp \sqrt{1741}}{22}\) into these denominators, we check:

For \(x^2 + 4x + 1\):

\[
\left(\frac{39 \mp \sqrt{1741}}{22}\right)^2 + 4\left(\frac{39 \mp
↪  \sqrt{1741}}{22}\right) + 1 \neq 0
\]

For \(x^2 - 10x\):

\[
\left(\frac{39 \mp \sqrt{1741}}{22}\right)^2 - 10\left(\frac{39 \mp
↪  \sqrt{1741}}{22}\right) \neq 0
\]

Since the solutions do not make the denominators zero, they are valid.
↪  Therefore, the sum of all real solutions is:

\[
\boxed{\frac{39}{11}}
\]
</think><answer> $\frac{39}{11}$ </answer>
```

Case Study of Qwen2.5-VL CoT and FRE-Text

Question: How many positive integers b have the property that $\log_b 729$ is a positive integer?

Qwen2.5-VL CoT:

```
We need to find positive integers $b$ such that $\log_{b} 729$ is a positive
↪  integer. This means that $b^{\log_{b} 729} = 729$.

Let's find the prime factorization of 729. We have $729 = 3^6$.

We know that if $\log_{b} 3^6$ is a positive integer, then $b$ must be a power
↪  of 3. Let's write $b = 3^x$. Then we have $\log_{3^x} 3^6 = 2$ since $b^q =
↪  3^6$ implies that $3^{x \cdot \log_{3} 3^6} = 3^6$ and $\log_{3} 3^6 = 6$.
↪  Thus, we have $x \cdot \log_{3} 3^6 = 6$. This simplifies to $x \cdot 6 =
↪  6$, so $x = 1$.

This means $b = 3^1 = 3$ is a solution. We need to check other powers of 3, but
↪  since $\log_{3} 3^6 = 6$, any other power of 3 greater than 3 would result
↪  in a non-positive integer for $\log_{b} 729$. Therefore, there is only one
↪  positive integer $b$ for which $\log_{b} 729$ is a positive integer.
```

```
The answer is: 1
```

FRE-Text:

```
<think>
To determine how many positive integers \( b \) have the property that
↪  \(\log_{b} 729\) is a positive integer, we start by letting \(\log_{b} 729
↪  = k\), where \( k \) is a positive integer. This equation can be rewritten
↪  in its exponential form as:

\[ b^k = 729 \]

We need to find all possible values of \( b \) such that \( 729 \) is a perfect
↪  power of \( b \). First, we factorize \( 729 \):

\[ 729 = 3^6 \]

For \( b^k = 3^6 \) to hold, \( b \) must be a power of \( 3 \). Let \( b = 3^m
↪  \), where \( m \) is a positive integer. Then we have:

\[ (3^m)^k = 3^{6k} \]

This implies that \( m \cdot k = 6 \). We need to find all pairs \((m, k)\)
↪  such that \( m \cdot k = 6 \). The possible pairs \((m, k)\) are:

1. \( (1, 6) \)
2. \( (2, 3) \)
3. \( (3, 2) \)
4. \( (6, 1) \)

For each pair \((m, k)\), we can find \( b \):

1. If \( m = 1 \) and \( k = 6 \), then \( b = 3^1 = 3 \).
2. If \( m = 2 \) and \( k = 3 \), then \( b = 3^2 = 9 \).
3. If \( m = 3 \) and \( k = 2 \), then \( b = 3^3 = 27 \).
4. If \( m = 6 \) and \( k = 1 \), then \( b = 3^6 = 729 \).

Thus, the possible values of \( b \) are \( 3, 9, 27, \) and \( 729 \). There
↪  are 4 such values.

Therefore, the number of positive integers \( b \) that have the property that
↪  \(\log_{b} 729\) is a positive integer is:\[\boxed{4}\]</think><answer> $4$
↪  </answer>
```

Case Study of Qwen2.5-VL and FRE-Multi

Question: What is the median number of points scored by the team per game?

Points per game

Stem	Leaf
2	0 4
3	4 5
4	9
5	6 7 8
6	0 6 6
7	5 6
8	0 4 5 5
9	0

Qwen2.5-VL:

```
<think>
Step 1: Identify all the data points in the stem and leaf plot.
From the stem and leaf plot, the points per game are as follows:
- 21
- 24
- 34, 35
- 49
- 56, 57, 58
- 66, 66
- 75, 76
- 84, 85, 85, 85
- 90

Step 2: Arrange the data points in ascending, descending or random order.
The points in ascending order are:
21, 24, 34, 35, 49, 56, 57, 58, 66, 66, 75, 76, 84, 85, 85, 85, 90

Step 3: Find the median.
Since there are 17 data points (an odd number), the median is the middle number
↪  when ordered.
In this case, the 9th number in the ordered list is the median.
The 9th number is 56.

Therefore, the median number of points scored per game is 56.

</think><answer> $\\boxed{56}$ </answer>
```

FRE-Multi:

```
<think>
To find the median, I need to follow these steps:
1. List all the values in ascending order.\n2. Identify the middle value(s).

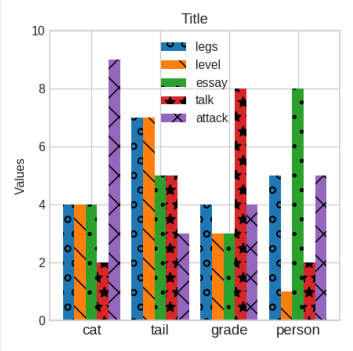
Let's list the values:\n20, 24, 20, 30, 34, 35, 40, 49, 50, 56, 57, 58, 60, 60,
↪  66, 66, 70, 75, 76, 80, 85, 85, 85, 86, 90
```

```
There are 25 values, so the median will be the 13th value in this ordered list.
The 13th value is 60.

Therefore, the median number of points scored per game is 60.
</think><answer> $\\boxed{60}$ </answer>
```

Case Study of Qwen2.5-VL v.s. FRE-Multi

Question: What is the value of 'legs' for 'grade'?



Qwen2.5-VL:

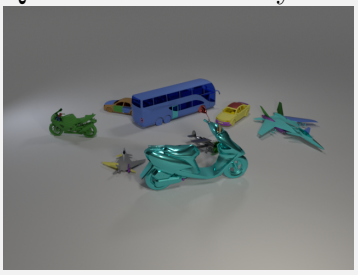
```
<think> To find the value of 'legs' for 'grade', we need to look at the legend
↪  of the graph, which clearly indicates that 'legs' are represented by the
↪  blue circles. After identifying the data for \"grade\" in the graph, we see
↪  that the corresponding value for 'legs' is 4. Therefore, the value of
↪  'legs' for 'grade' is 4.\n</think><answer> $\\boxed{4}$ </answer>
```

FRE-Multi:

```
<think> The 'legs' value for 'grade' is shown on the bar chart as 4.
↪  </think><answer> $\\boxed{4}$ </answer>
```

Case Study of Qwen2.5-VL and FRE-Multi

Question: How many vehicles in the image have wheels?



Qwen2.5-VL:

```
<think>
First, I'll identify each vehicle in the image. There is a double bus (blue), a
↪  car (tan), a car (yellow), a motorcycle (green), a scooter (cyan), and an
↪  airplane (gray). Some of these are likely to have wheels.

Now, examining the vehicles one by one:
- The blue double bus typically has wheels.
- The tan car and the yellow car have wheels.
- The green motorcycle has two wheels.
- The scooter likely has two wheels as well.
- The gray airplane typically has small wheels beneath.

When considering this categorically, all of these vehicles seem to have wheels.
↪  However, classifying vehicles as always having wheels or as having
↪  potential wheels introduces subjectivity. Typically most vehicles have
↪  wheels unless otherwise specified, so I'll make conservative assumptions
↪  regarding certainty.

I'll perform a count suggesting the vehicles have wheels:
1. The double bus has wheels.
2. The cars each usually have wheels.
3. The motorcycle presumably has wheels.
4. The scooter likely has typical two wheels.
5. The airplane typically has wheels.\n\nThus, I will count the potentially
↪  wheel-related vehicles (assumed to have wheels before explicit
↪  confirmation):
1. (Wheel-like others might exist)\n\nFinal number of vehicles likely to have
↪  wheels:
\\boxed{4}
</think><answer> $\\boxed{4}$ </answer>
```

FRE-Multi:

```
<think>
We need to count the number of vehicles with wheels in the image. Let's
↪  identify them:

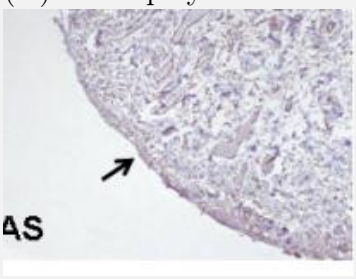
1. The blue bus has wheels.
2. The green motorcycle has wheels.
3. The brown sedan has wheels.
4. The yellow sedan has wheels.
5. The cyan scooter has wheels.

There are no other vehicles with wheels in the image.

Therefore, the total number of vehicles with wheels is 5.
</think><answer> $\\boxed{5}$ </answer>
```

Case Study of Qwen2.5-VL and FRE-Multi

Question: What is the purpose of the left lane in the picture?
Choices:
(A) To show the results of immunofluorescent labeling
(B) To indicate the upper layer of synovial membranes
(C) To show the magnification of the image
(D) To display the results of immunohistochemistry



Qwen2.5-VL:

```
<think> The image appears to be a microscope slide of tissue sections. The
↪ left lane consistently showing up in the micrograph suggests that it is
↪ used to display the results of immunohistochemistry, where antibodies are
↪ applied to stained tissue sections to detect specific molecules. The other
↪ options do not align with the visual presentation and technical names
↪ associated with immunohistochemistry. </think><answer> $\\boxed{D}$
↪ </answer>
```

FRE-Multi:

```
<think> The presence of the arrow and the context suggest that the left lane is
↪ used to display the results of immunohistochemistry. </think><answer>
↪ \\boxed{D} </answer>
```

J Detailed Description of Reward Function

Our reward function consists of two main components that work together to guide the model's behavior:

Format Reward

The format reward $r_f(y)$ evaluates whether the responses follow the required structured format. Specifically, we require:

- The reasoning process must be enclosed within `<think></think>` tags
- The final answer must be provided within `<answer></answer>` tags

This structured format is crucial as it:

- Encourages explicit demonstration of the reasoning process
- Enhances readability of the model's responses
- Facilitates evaluation of reasoning quality
- Helps separate the thought process from the final answer

Accuracy Reward

The accuracy reward $r_a(y)$ evaluates the correctness of the solution using a symbolic verification approach [42]. The verification process involves:

- Parsing both the model's answer and ground-truth solution into comparable representations
- Checking for mathematical equivalence rather than exact string matching
- Allowing recognition of different but equivalent expressions as correct

Combined Reward Function

The final reward function combines both components:

$$r(y) = \alpha \cdot r_f(y) + r_a(y)$$

where:

- $r_f(y)$ is the format reward
- $r_a(y)$ is the accuracy reward
- α is a hyperparameter that controls the relative importance of format versus accuracy

This combined reward function effectively guides the model to:

- Produce correct solutions
- Articulate clear reasoning processes
- Maintain consistent response formatting
- Balance between format adherence and solution accuracy

References

[1] Luo M, Tan S, Wong J, Shi X, Tang W Y, Roongta M, Cai C, Luo J, Zhang T, Li L E, Popa R A, Stoica I. DeepScaleR: Surpassing o1-preview with a 1.5b model by scaling rl, 2025. Technical Report

[2] Gao B, Song F, Yang Z, Cai Z, Miao Y, Dong Q, Li L, Ma C, Chen L, Xu R, Tang Z, Wang B, Zan D, Quan S, Zhang G, Sha L, Zhang Y, Ren X, Liu T, Chang B. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. arXiv preprint arXiv:2410.07985, 2024

[3] Team R S. STILL-3-1.5B-preview: Enhancing slow thinking abilities of small models through reinforcement learning, 2025

[4] Shi W, Hu Z, Bin Y, Liu J, Yang Y, Ng S K, Bing L, Lee w R K. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2024

[5] Lu P, Qiu L, Chen J, Xia T, Zhao Y, Zhang W, Yu Z, Liang X, Zhu S C. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. arXiv preprint arXiv:2110.13214, 2021

[6] Kembhavi A, Seo M, Schwenk D, Choi J, Farhadi A, Hajishirzi H. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017

[7] Zhang X, Wu C, Zhao Z, Lin W, Zhang Y, Wang Y, Xie W. PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415, 2023

[8] Mathew M, Bagal V, Tito R P, Karatzas D, Valveny E, Jawahar C. Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021

[9] Lu P, Qiu L, Chang K W, Wu Y N, Zhu S C, Rajpurohit T, Clark P, Kalyan A. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. arXiv preprint arXiv:2209.14610, 2022

[10] Singh A, Natarajan V, Shah M, Jiang Y, Chen X, Batra D, Parikh D, Rohrbach M. Towards vqa models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019

[11] Kembhavi A, Salvato M, Kolve E, Seo M, Hajishirzi H, Farhadi A. A diagram is worth a dozen images. arXiv preprint arXiv:1603.07396, 2016

[12] Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. IJCV, 2016

[13] Kahou S E, Atkinson A, Michalski V, Kádár Á, Trischler A, Bengio Y. Figureqa: An annotated figure dataset for visual reasoning. arXiv preprint arXiv:1710.07300, 2017

[14] Masry A, Long D X, Tan J Q, Joty S R, Hoque E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244, 2022

[15] Lu P, Mishra S, Xia T, Qiu L, Chang K W, Zhu S C, Tafjord O, Clark P, Kalyan A. Learn to explain: Multimodal reasoning via thought chains for science question answering. arXiv preprint arXiv:2209.09513, 2022

[16] Methani N, Ganguly P, Khapra M M, Kumar P. Plotqa: Reasoning over scientific plots. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2019

[17] Cao J, Xiao J. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In: Proceedings of the International Conference on Computational Linguistics. 2022

[18] Li Z, Wang X, Stengel-Eskin E, Kortylewski A, Ma W, Durme B V, University A Y J H, Southern California o U, Informatics f M P I, Freiburg o U. 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. 2022

[19] Kafle K, Cohen S D, Price B L, Kanan C. DVQA: Understanding data visualizations via question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018

[20] Agrawal A, Lu J, Antol S, Mitchell M, Zitnick C L, Parikh D, Batra D. Vqa: Visual question answering. IJCV, 2015

[21] Lu P, Gong R, Jiang S, Qiu L, Huang S, Liang X, Zhu S C. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2021

[22] Chang S, Palzer D, Li J, Fosler-Lussier E, Xiao N. Mapqa: A dataset for question answering on choropleth maps. arXiv preprint arXiv:2211.08545, 2022

[23] Chen J, Li T, Qin J, Lu P, Lin L, Chen C, Liang X. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. arXiv preprint arXiv:2212.02746, 2022

[24] Gurari D, Li Q, Stangl A, Guo A, Lin C, Grauman K, Luo J, Bigham J P. Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018

[25] Seo M, Hajishirzi H, Farhadi A, Etzioni O, Malcolm C. Solving geometry problems: Combining text and diagram interpretation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2015

[26] Lindström A D, Abraham S S. CLEVR-Math: A dataset for compositional language, visual and mathematical reasoning. arXiv preprint arXiv:2208.05358, 2022

[27] Lau J J, Gayen S, Abacha A B, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data, 2018

[28] Racanière S, Weber T, Reichert D, Buesing L, Guez A, Jimenez Rezende D, Puigdomènech Badia A, Vinyals O, Heess N, Li Y, others . Imagination-augmented agents for deep reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems. 2017

[29] Hu J, Wu X, Zhu Z, Xianyu , Wang W, Zhang D, Cao Y. OpenRLHF: An easy-to-use, scalable and high-performance RLHF framework. arXiv preprint arXiv:2405.11143, 2024

[30] Fourrier C, Habib N, Kydlíček H, Wolf T, Tunstall L. LightEval: A lightweight framework for llm evaluation, 2023

[31] Lightman H, Kosaraju V, Burda Y, Edwards H, Baker B, Lee T, Leike J, Schulman J, Sutskever I, Cobbe K. Let's verify step by step. In: Proceedings of the International Conference on Learning Representations. 2023

[32] Rein D, Hou B L, Stickland A C, Petty J, Pang R Y, Dirani J, Michael J, Bowman S R. GPQA: A graduate-level google-proof q&a benchmark. In: Proceedings of the Conference on Language Modeling. 2024

[33] Zhang K, Li B, Zhang P, Pu F, Cahyono J A, Hu K, Liu S, Zhang Y, Yang J, Li C, Liu Z. Lmms-eval: Reality check on the evaluation of large multimodal models. arXiv preprint arXiv:2407.12772, 2024

[34] Wang K, Pan J, Shi W, Lu Z, Ren H, Zhou A, Zhan M, Li H. Measuring multimodal mathematical reasoning with math-vision dataset. In: Proceedings of Advances in Neural Information Processing Systems. 2025

[35] Zhang R, Jiang D, Zhang Y, Lin H, Guo Z, Qiu P, Zhou A, Lu P, Chang K W, Qiao Y, Gao P, Li H. MathVerse: Does your multi-modal llm truly see the diagrams in visual math problems? In: Proceedings of the European Conference on Computer Vision. 2024

[36] Lu P, Bansal H, Xia T, Liu J, Li C, Hajishirzi H, Cheng H, Chang K W, Galley M, Gao J. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In: Proceedings of the International Conference on Learning Representations. 2024

[37] He C, Luo R, Bai Y, Hu S, Thai Z L, Shen J, Hu J, Han X, Huang Y, Zhang Y, Liu J, Qi L, Liu Z, Sun M. OlympiadBench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2024

[38] Chen L, Li J, Dong X, Zhang P, Zang Y, Chen Z, Duan H, Wang J, Qiao Y, Lin D, Zhao F. Are we on the right way for evaluating large vision-language models? In: Proceedings of Advances in Neural Information Processing Systems. 2024

[39] Kwon W, Li Z, Zhuang S, Sheng Y, Zheng L, Yu C H, Gonzalez J E, Zhang H, Stoica I. Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles. 2023

[40] Team Q. Qwen2.5: A party of foundation models, September 2024

[41] Zhang M, Dai Q, Yang Y, Bao J, Chen D, Qiu K, Luo C, Geng X, Guo B. MageBench: Bridging large multimodal models to agents. arXiv preprint arXiv:2412.04531, 2024

[42] Wang P, Li L, Shao Z, Xu R, Dai D, Li Y, Chen D, Wu Y, Sui Z. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. arXiv preprint arXiv:2312.08935, 2023