Recent Advances in Multimodal Reasoning

Yen-Shuo Su, Karthic Palaniappan, Ivy He

10/23/2025

Georgia Tech

Visionary-R1: Mitigating Shortcuts in Visual Reasoning with Reinforcement Learning

Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, Kaiyang Zhou

Georgia Tech

Outline

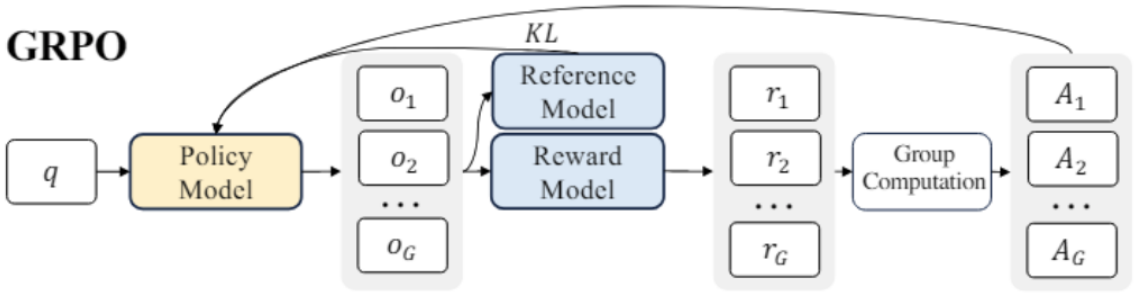
- Problem Statement
- Related Works
- Approach
- Experiments & Results
- Limitations, Societal Implications
- Summary of Strengths, Weaknesses, Relationship to Other Papers

Problem Statement - Challenges

- Training AI models to reason is challenging due to lack of large-scale human-annotated reasoning data
- Recent LLM work DeepSeel-R1 shows promising results using RL with only Q&A pairs
- Key Question: Can we train VLMs to reason using only visual Q&A pairs without explicit CoT supervision?

Background – Group Relative Policy Optimization (GRPO)

- Sample multiple responses for each question
- Score each response with reward function
- Normalize rewards within group to get advantages
- Update policy to favor high-advantage responses
- Key Benefits
  - No critic model needed
  - Group-based normalization stabilizes training
  - Successfully induces reasoning in LLMs



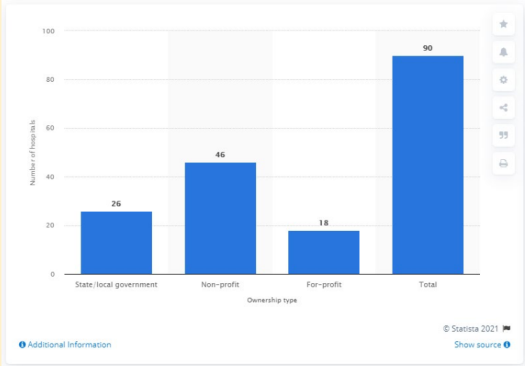
Georgia Tech

Problem Statement - Critical Problems

- Simply applying GRPO to VLMs leads to shortcut learning
- Models generate short, uninformative reasoning for easy questions
- Poor generalization to harder, unseen questions

Easy Sample (Train)

Q: How many hospitals in Colorado were owned by the state or local government in 2019? A: 26



GRPO

<think> The number of hospitals in Colorado owned by the state or local government in 2019 is 26.</think> <answer> 26 </answer>

Visionary-R1

<info> The image is a bar graph that displays the number of hospitals in Colorado by ownership type. Here are the key points:
- The total number of hospitals in Colorado is 90.
- 26 hospitals are owned by the state/local government.
- 46 hospitals are non-profit.
- 18 hospitals are for-profit.
</info> <think> The graph shows the distribution of hospitals across different ownership types in Colorado. **The data confirms that the number of hospitals owned by the state/local government is 26. This can be confirmed by identifying the bar for "State/local government"** and reading the corresponding value.
</think> <answer> 26 </answer>

Hard Sample (Test) **Q:How many queries have a p-value lower than 0.50? A:4**

	shutdown	bullet	objective	superintelligent	rot23	paperclips	paperclips_large	water
p-value	**0.92**	**0.69**	0.01	**0.10**	**0.58**	**0.43**	**0.60**	0.01

Table 1: **Goodness-of-fit results for power law models.** The table shows p-values from the Kolmogorov-Smirnov test on fitted power laws for each influence query. Values above the 0.1 thresholds suggested by Clauset et al. (2009) indicate the power law is a plausible fit.

GRPO

<think> To answer the question, **we need to count the number of p-values that are lower than 0.50.** </think>
<answer> 3 </answer>

Visionary-R1

<info> The image contains a table titled "Table 1: Goodness-of-fit results for power law models." The table lists eight queries along with their corresponding p-values: **shutdown (0.92), bullet (0.69), objective (0.01), superintelligent (0.10), rot23 (0.58), paperclips (0.43), paperclips_large (0.60), and water (0.01).** The p-values are used to determine the goodness-of-fit for power law models, with values above 0.1 suggested by Clauset et al. (2009) indicating a plausible fit. </info> <think> To determine how many queries have a p-value lower than 0.50, **we need to count the number of p-values that are less than 0.50. From the table, the p-values are: 0.92, 0.69, 0.01, 0.10, 0.58, 0.43, 0.60, and 0.01. The p-values that are less than 0.50 are 0.01, 0.01, 0.1, and 0.43.** Therefore, there are four queries with p-values lower than 0.50. </think> <answer> 4 </answer>

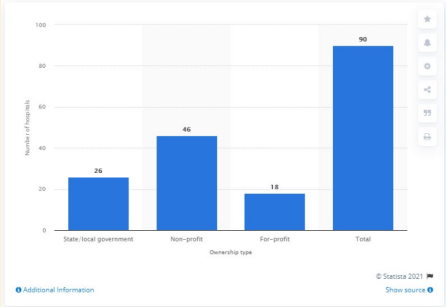
Georgia Tech

Approach – Shortcut Problem

- GRPO encourages models to exploit easy patterns in training data
- Model ignores visual input, and replies on textual patterns from questions
- Resulting in correct answers on simple problems without grounding in images
- Visonary-R1 difference
  - Forces detailed image interpretation first
  - Consistent approach regardless of difficulty
  - Generalizes to hard questions

Easy Sample (Train)

Q: How many hospitals in Colorado were owned by the state or local government in 2019? A: 26



GRPO

<think> The number of hospitals in Colorado owned by the state or local government in 2019 is 26.</think> <answer> 26 </answer>

Visionary-R1

<info> The image is a bar graph that displays the number of hospitals in Colorado by ownership type. Here are the key points:
- The total number of hospitals in Colorado is 90.
- 26 hospitals are owned by the state/local government.
- 46 hospitals are non-profit.
- 18 hospitals are for-profit.
</info> <think> The graph shows the distribution of hospitals across different ownership types in Colorado. **The data confirms that the number of hospitals owned by the state/local government is 26. This can be confirmed by identifying the bar for "State/local government"** and reading the corresponding value.
</think> <answer> 26 </answer>

Hard Sample (Test) **Q:How many queries have a p-value lower than 0.50? A:4**

	shutdown	bullet	objective	superintelligent	rot23	paperclips	paperclips_large	water
p-value	**0.92**	**0.69**	0.01	**0.10**	**0.58**	**0.43**	**0.60**	0.01

Table 1: **Goodness-of-fit results for power law models.** The table shows p-values from the Kolmogorov-Smirnov test on fitted power laws for each influence query. Values above the 0.1 thresholds suggested by Clauset et al. (2009) indicate the power law is a plausible fit.

GRPO

<think> To answer the question, **we need to count the number of p-values that are lower than 0.50.** </think>
<answer> 3 </answer>

Visionary-R1

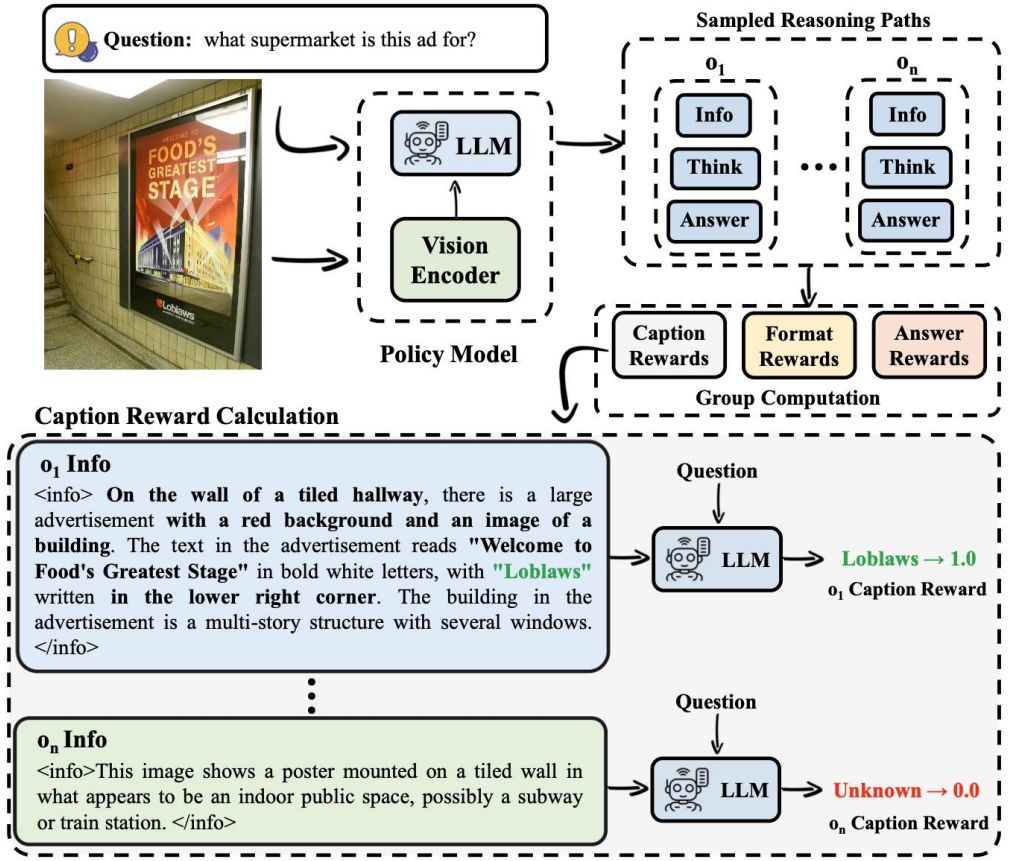
<info> The image contains a table titled "Table 1: Goodness-of-fit results for power law models." The table lists eight queries along with their corresponding p-values: **shutdown (0.92), bullet (0.69), objective (0.01), superintelligent (0.10), rot23 (0.58), paperclips (0.43), paperclips_large (0.60), and water (0.01).** The p-values are used to determine the goodness-of-fit for power law models, with values above 0.1 suggested by Clauset et al. (2009) indicating a plausible fit. </info> <think> To determine how many queries have a p-value lower than 0.50, **we need to count the number of p-values that are less than 0.50. From the table, the p-values are: 0.92, 0.69, 0.01, 0.10, 0.58, 0.43, 0.60, and 0.01. The p-values that are less than 0.50 are 0.01, 0.01, 0.1, and 0.43.** Therefore, there are four queries with p-values lower than 0.50. </think> <answer> 4 </answer>

Georgia Tech

Approach – Visionary-R1 Framework

- Proposed solution: Caption-Reason-Answer output format

- Three-stage output structure
  - Caption (<info>): Generate detailed image description
  - Reason (<think>): Construct reasoning chain based on caption
  - Answer (<answer>): Provide final answer



Approach – Visionary-R1 Rewards

- Format reward
  - Binary reward checking adherence to the format of '<info>…</info> <think>…</think> <answer>…</answer>'
- Caption reward
  - Extracts caption content between <info> tags
  - Feeds caption + question to LLM
  - LLM answers based only on captions
  - Determines if the caption is informative based on the correctness of the answers
  - Prevents reward hacking: Special prompt to filter out reasoning/answers in captions
- Accuracy reward
  - Standard correctness check
- Combined reward

$$R_i = r_a + r_f + \alpha r_c$$

Approach – Training with Cosine Annealing KL

- Stabilizing RL training with dynamic KL penalty
- Standard GRPO objective

$$\mathcal{J}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{n} \sim \pi_{\theta_{\text{old}}}(O|q)]$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta \mathrm{D}_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right)\right)$$

- KL divergence penalty

$$\mathbb{D}_{KL}\left[\pi_\theta || \pi_{ref}\right] = \frac{\pi_{ref}\left(o_i \mid q\right)}{\pi_\theta\left(o_i \mid q\right)} - \log\frac{\pi_{ref}\left(o_i \mid q\right)}{\pi_\theta\left(o_i \mid q\right)} - 1$$

  - Prevents model from deviating too far from reference policy
  - Stabilizes training

Georgia Tech

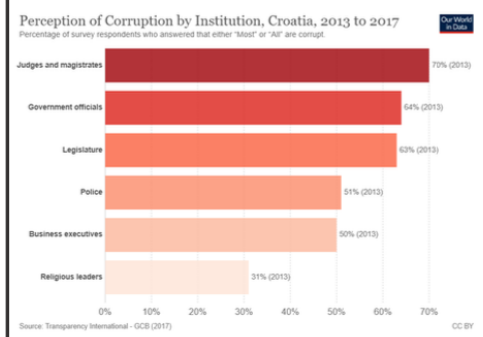
Approach – Training with Cosine Annealing KL

- The KL coefficient Challenge
  - Large KL coefficients causes the performance to stay close to the baseline, falls short in terms of reasoning abilities, causing the model to collapse
  - Small KL coefficients encourages more exploration, but causes the reward hacking problem

- Solution: Cosine Annealing

$$\hat{\beta} = \frac{\beta}{2} \times \left(1 + \cos\left(\pi \times \frac{T_{\text{cur}}}{T_{\max}}\right)\right)$$

Training Data



Q: What is the difference in the value of Police and Business executives in corruption? A: 1

Static KL (0.04)

<info> The image presents a bar chart detailing the perception of corruption by institution in Croatia from 2013 to 2017.
1. Identify the values:
- Percentage values:
- Government officials: 64%
- Judges and magistrates: 70%
- Legislation: 63%
- Police: 51%
- Business executives: 50%
- Religious leaders: 31%
</info> </think>
<info> Judging from above Chain Comparison Logical explicit logical Explicit Chain Calculating conclusion Analysis. </info></think></think></think><!-- Chain Chain Chain ……<Repeating>……

Static KL (0.008)

<info>Police is the percentage of corruption which is 51, Business executives is the percentage of corruption which is 50, the difference is 1 </info> <think> </think> <answer>1 </answer>

Georgia Tech

Experiments and Results - Setup

- A total of 11 aggregated Q&A training datasets with diverse coverage in various domains
- Base model:
  - Qwen2.5-VL-3B
  - Strong visual understanding from pretraining
  - No reasoning post-training
- Training details
  - n = 8 responses per question
  - α = 0.1 (caption reward weight)
  - β = 0.04 (KL coefficient, with cosine annealing)
  - ~1,500 GPU hours on NVIDIA A800-80G

Table 4: Composition of our training data.

Dataset	Size	Answer Type	Visual Format
A-OKVQA [38]	17.1K	Multi-choice	General Scene
ChartQA [40]	28.3K	Open-text+Num	Chart
AI2D [44]	15.5K	Multi-choice	Diagram
ScienceQA [45]	6.2K	Multi-choice	Scene + Chart
GeoQA+ [42]	12.1K	Multi-choice	Math
DocVQA [43]	39.5K	Open-text	Document
CLEVR-Math [46]	32.6K	Num	3D
Icon-QA [47]	29.9K	Multi-choice	Diagram
TabMWP [48]	23.1K	Open-text+Num	Table
RoBUT SQA [41]	34.1K	Open-text+Num	Chart
TextVQA [39]	34.6K	Multi-choice	General Scene
Total	272.6K		

Experiments and Results - Setup

- Evaluation Benchmarks
  - MathVista: Logical, algebraic, scientific reasoning
  - MathVision: Mathematical visual reasoning
  - MMStar: Perception, math understanding, science, technology, logical reasoning
  - MMBench: Comprehensive visual and mathematical reasoning
- Baselines
  - SFT: Direct supervised training on Q&A data
  - GRPO: Standard GRPO without caption-first design
  - SOTA: GPT-4o, Claude3.5, Gemini, open-source reasoning models

	Size	Strategy	Data	MathVista	MathVision	MMStar	MMBench
Close-source models							
GPT-4o* [28]	-	-	-	63.8	31.2	65.1	84.3
GPT-o1* [29]	-	-	-	71.8	63.2	67.5	83.8
Claude3.5-Sonnet [30]	-	-	-	67.7	37.9	65.1	82.6
Claude3.7-Sonnet* [31]	-	-	-	74.5	58.6	68.8	82.0
Gemini-1.5-Pro [32]	-	-	-	63.9	19.2	59.1	73.9
Gemini-2.5-Pro* [33]	-	-	-	82.7	73.3	77.5	90.1
Open-source models							
Qwen2.5-VL [26]	3B	-	-	62.3	21.2	55.9	79.1
InternVL2.5 [34]	4B	-	-	60.5	20.9	58.3	81.1
MiniCPM-V2.6 [35]	8B	-	-	60.6	17.5	57.5	81.5
LLaMA3.2 [36]	11B	-	-	51.5	-	49.8	65.8
Reasoning models							
Ovis [18]	4B	SFT	CoT	66.6	-	59.5	79.3
Mulberry [19]	7B	SFT	CoT	63.1	-	61.3	-
R1-Onevision [24]	7B	SFT+RL	CoT	64.1	29.9	-	-
Insight-V [21]	7B	SFT+RL	CoT	59.9	-	61.5	82.3
R1-VL [22]	7B	SFT+RL	CoT	63.5	24.7	60	-
LLaVA-CoT [17]	11B	SFT	CoT	54.8	-	57.6	75
Our models							
Base Model	3B	-	-	61.5	19.1	52.4	82.1
SFT	3B	SFT	QA	54.6	7.0	61.9	80.7
GRPO	3B	RL	QA	61.8	20.3	54.3	78.6
Visionary-R1	3B	RL	QA	69.4	24.7	66.5	84.1

Georgia Tech

Experiments and Results – Main Results

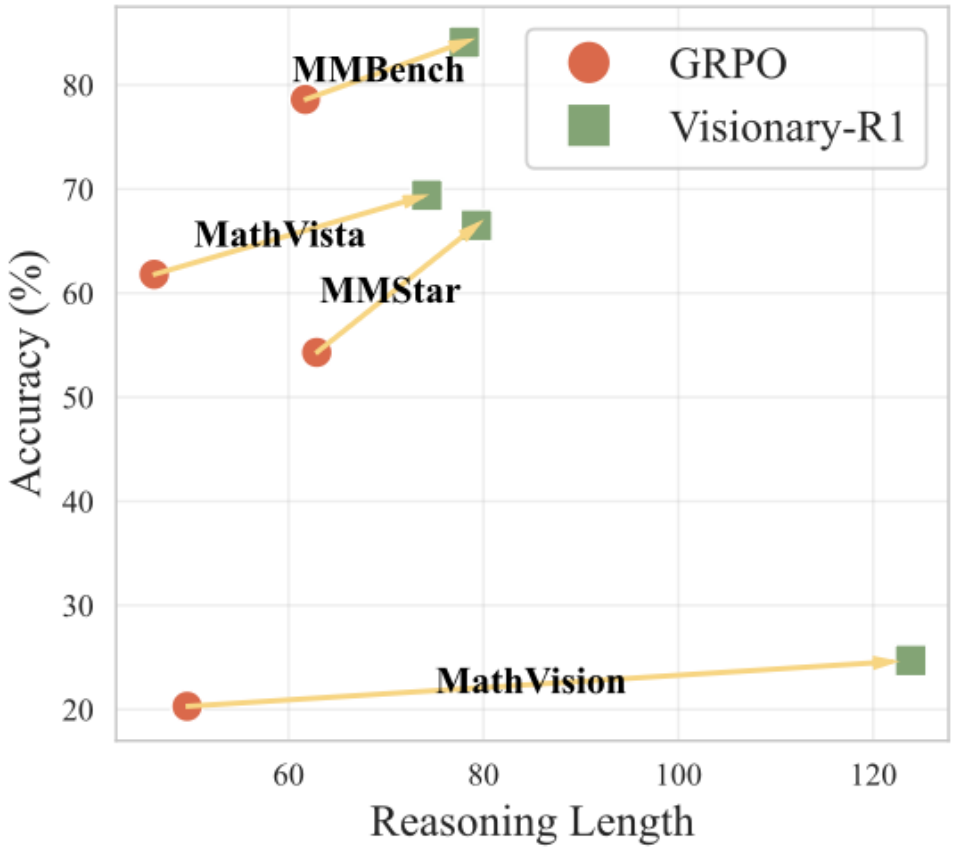
- Visionary-R1 outperforms strong commercial models
- Key Observations
  - SFT Fails: Worse than base model on 3 out of the 4 benchmarks
  - GRPO Marginal: Slight improvements, but unstable
  - Visionary-R1: Substantial improvements across all benchmarks
- Performance highlights
  - Beats GPT-4o on MathVista, MMStar
  - Beats Claude3.5-Sonnet on MathVista, MMStar
  - Beats Gemini-1.5-Pro on all benchmarks
  - Competitive or better performance vs open-source reasoning models despite smaller size and no CoT supervision

	Size	Strategy	Data	MathVista	MathVision	MMStar	MMBench
Close-source models							
GPT-4o* [28]	-	-	-	63.8	31.2	65.1	84.3
GPT-o1* [29]	-	-	-	71.8	63.2	67.5	83.8
Claude3.5-Sonnet [30]	-	-	-	67.7	37.9	65.1	82.6
Claude3.7-Sonnet* [31]	-	-	-	74.5	58.6	68.8	82.0
Gemini-1.5-Pro [32]	-	-	-	63.9	19.2	59.1	73.9
Gemini-2.5-Pro* [33]	-	-	-	82.7	73.3	77.5	90.1
Open-source models							
Qwen2.5-VL [26]	3B	-	-	62.3	21.2	55.9	79.1
InternVL2.5 [34]	4B	-	-	60.5	20.9	58.3	81.1
MiniCPM-V2.6 [35]	8B	-	-	60.6	17.5	57.5	81.5
LLaMA3.2 [36]	11B	-	-	51.5	-	49.8	65.8
Reasoning models							
Ovis [18]	4B	SFT	CoT	66.6	-	59.5	79.3
Mulberry [19]	7B	SFT	CoT	63.1	-	61.3	-
R1-Onevision [24]	7B	SFT+RL	CoT	64.1	29.9	-	-
Insight-V [21]	7B	SFT+RL	CoT	59.9	-	61.5	82.3
R1-VL [22]	7B	SFT+RL	CoT	63.5	24.7	60	-
LLaVA-CoT [17]	11B	SFT	CoT	54.8	-	57.6	75
Our models							
Base Model	3B	-	-	61.5	19.1	52.4	82.1
SFT	3B	SFT	QA	54.6	7.0	61.9	80.7
GRPO	3B	RL	QA	61.8	20.3	54.3	78.6
Visionary-R1	3B	RL	QA	69.4	24.7	66.5	84.1

Georgia Tech

Experiments and Results – Main Results

- Clear positive correlation between the length of reasoning and better accuracy
- Validates the hypothesis that detailed reasoning improves performance



Georgia Tech

Ablation Studies

- Component-wise contribution analysis

- Methodology
  - Train on individual datasets
  - Two experimental setups for diversity

- Key insights:
  - Caption format alone provides substantial improvement
  - Caption reward proves additional gains
  - Both components essential for best performance
  - Consistent across domains

Table 2: Ablation study on different components in Visionary-R1.

	Train: ChartQA		Train: A-OKVQA	
Method	MathVista	MathVision	MMStar	MMBench
Zero-shot	61.5	19.1	52.4	82.1
GRPO	59.0	18.2	54.2	82.6
GRPO+caption	62.6	20.9	60.4	85.5
Visionary-R1	**64.6**	**22.7**	**62.9**	**87.6**

Georgia Tech

Ablation Studies

- Shows qualitative difference in reasoning quality
- GRPO: Short, pattern-based
- GRPO + Caption: Better, but caption may be superficial
- Visionary-R1: Detailed caption + thorough reasoning

Test Data

Boxes per shipment

Stem	Leaf
1	9
2	6
3	4 9
4	1 9
5	0 3 9
6	4 8

Q:A shipping company keeps track of the number of boxes in each shipment they send out. How many shipments had exactly 56 boxes? A:0

GRPO

<think> To find how many shipments had exactly 56 boxes, we need to look for the number 56 in the plot. </think>
<answer> 1 </answer>

GRPO+Caption

<info> The image is a stem-and-leaf plot representing the number of boxes in each shipment. The stem represents the tens digit, and the leaf represents the ones digit. </info> <think> To find the number of shipments with exactly 56 boxes, **we need to look at the row where the stem is 5 and the leaf is 6.** So, there is no entry in the row where the stem is 5 and the leaf is 6, which means there is no shipment with exactly 56 boxes. </think> **<answer> 0 </answer>**

GRPO+Caption+r_c

<info> The image is a stem-and-leaf plot representing the number of boxes in each shipment. The stem represents the tens digit, and the leaf represents the ones digit. **The plot shows the following data: 19, 26, 34, 39, 41, 49, 50, 53, 59, 64, 68.** </info> <think> To find how many shipments had exactly 56 boxes, we need to look for the number 56 in the plot. **In the stem-and-leaf plot, the number 56 would be represented by the stem 5 and the leaf 6. We can see that 56 is not listed in the plot.** Therefore, there are no shipments with exactly 56 boxes. </think> **<answer> 0 </answer>**

Georgia Tech

Additional Analysis – KL Coefficient Deep Dive

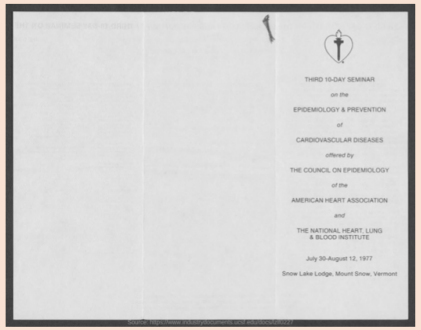
- Static values fail for Visionary-R1
- Dynamic scheduling essential: Both linear and cosine dramatically better
- Cosine slightly better than linear: Smoother transition preferred
- Minimal impact on GRPO: Suggests this specifically helps captioning component
- Early high β prevents collapse and maintains caption quality
- Late low β allows exploration and enables detailed outputs
- Caption generation particularly sensitive to exploration-exploitation tradeoff

Method	Strategy	MathVista	MathVision
Visionary-R1	Static (0.04)	60.9	19.3
	Static (0.008)	60.7	18.7
	Linear	63.4	22.4
	Cosine	64.6	22.7
GRPO	Static (0.04)	59.0	18.2
	Cosine	59.6	18.4

Georgia Tech

Qualitative Analysis

Document



Q: Where the seminar is conducted?
A: Snow Lake Lodge, Mount Snow, Vermont

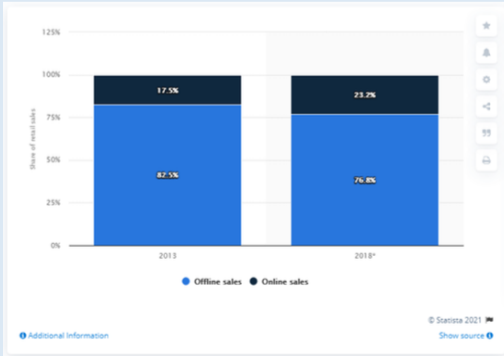
Visionary-R1

<info> The image shows a black and white document that appears to be a invitation or informational flyer. The details of the seminar are as follows:

- Title: Third Day Seminar
- Topic: Epidemiology & Prevention of Cardiovascular Diseases
- Organizer: Council on Epidemiology of the American Heart Association
- Collaborator: The National Heart, Lung & Blood Institute
- Date: July 30 - August 12, 1977
- Location: Snow Lake Lodge, Mount Snow, Vermont

The flyer has a minimalistic design with a logo featuring a heart and a torch at the top-left corner, which could be associated with the American Heart Association, given their logo often resembles a flame or torch within a heart shape. </info> <think> Given the information on the document, the key aspect to focus on is the location provided, which is explicitly mentioned as "Snow Lake Lodge, Mount Snow, Vermont." </think> <answer> Snow Lake Lodge, Mount Snow, Vermont </answer>

Chart

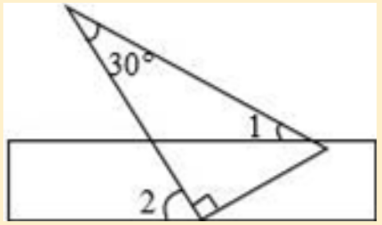


Q: What is the predicted increase in online sales of electricals and electronics by 2018?
A: 23.2

Visionary-R1

<info>
The image is a bar chart with two data series: Offline sales and Online sales. The x-axis represents the years 2013 and 2018, while the y-axis represents the share of retail sales as a percentage.
For the year 2013, the Online sales are shown at 17.5%, and Offline sales are 82.5%.
For 2018, the Online sales increased to 23.2%, and Offline sales decreased to 76.8%.
It's worth noting that the source of the data is Statista 2021.
</info>
<think> The increase in online sales is shown in the percentage, from 17.5 percent in 2013 to 23.2 percent in 2018.
</think>
<answer> 23.2 </answer>

Math

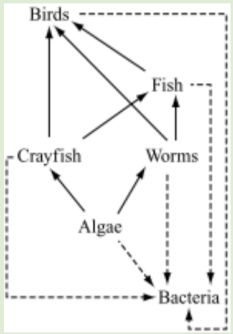


Q: Place a ruler and a right-angled triangle ruler as shown in the figure (∠ACB is a right Angle). Given that ∠1=15°, what is the size of ∠2?
A: 45°

Visionary-R1

<info> A ruler and a right-angled triangle board. The right-angled triangle board has a 30° Angle. There are two marks on the ruler. One is in the middle and the other is in the lower left corner. The Angle indicated by the mark at the lower left corner is ∠2, which is calculated by the Angle between the right Angle of the right-angled triangle and the ruler. The Angle marked in the middle is ∠1. </info>
<think> According to the problem statement, one Angle of a right triangle is 30 degrees and the other Angle is 90 degrees. The problem states that ∠1=15 degrees. According to the formula for the sum of interior angles of a triangle, the other Angle (the Angle adjacent to the right triangle) is 60 degrees. Because the triangle set is a right triangle, the adjacent angles are 30 degrees. A right triangle has a right Angle of 90 degrees. So the Angle that ∠2 is adjacent to 90 degrees of a right triangle is 60 degrees minus 15 degrees, that is, 45 degrees. </think> <answer> 45° </answer>

Diagram



Q: Which is the top predator in the food web? Options: Birds, Crayfish, Fish, None of the above
A: Birds

Visionary-R1

<info> This image depicts a food web where organisms are connected by arrows indicating the direction of energy flow from one organism to another. The food web includes various organisms such as birds, fish, crayfish, worms, algae, and bacteria. The arrows show the feeding relationships among these organisms, with the arrows pointing from the prey to the predator. The food web forms a hierarchical structure where energy flows from the bottom to the top, starting from the bacteria and algae at the base and moving through various levels of consumers to the top predators. </info>
<think> To determine the top of the food chain, we need to identify the organisms that do not have any arrows pointing towards them. In this food web, birds are positioned at the top, with arrows pointing towards them from the organisms below. There are no arrows pointing towards birds from the bottom of the web. </think>
<answer> Birds </answer>

Georgia Tech

Limitations & Societal Implications

- Limitations
  - Scale Constraints
    - Only 3B parameter model tested
    - Larger models (7B, 13B, 70B) unexplored due to budget (~1,500 GPU hours already)
  - Dataset Considerations
    - No preprocessing/filtering applied to training data
    - Potential inclusion of low-quality samples
  - Caption Reward Design
    - Relies on LLM component quality
    - Potential circular dependency (model judges its own captions)
    - No independent caption quality validation
  - Limited Analysis
    - Few failure case discussions
    - Generalization limits unclear

Limitations & Societal Implications

- Society Implications
  - Positive
    - Democratization: Removes dependency on proprietary models (GPT-4o)
    - Accessibility: Lower barrier to entry for researchers
    - Transparency: Open-source approach promotes reproducibility
    - Cost reduction: Pure RL more efficient than distillation pipelines
  - Potential Concerns
    - Misuse potential: Could generate misleading visual analyses
    - Deployment responsibility: Needs monitoring and safeguards
    - Evaluation challenges: How to verify reasoning quality at scale?

- Future Work
  - Scale to larger models (investment in compute)
  - Apply cosine annealing KL to other RL domains
  - Explore additional visual domains (video, 3D, medical imaging)
  - Develop independent caption quality metrics

Summary of Strengths, Weaknesses

- Strengths:
  - First to systematically identify and characterize shortcut learning in visual RL
  - Caption-reason-answer format is intuitive and practically implementable
  - Strong empirical results that beats GPT-4o, Claude3.5-Sonnet on multiple benchmarks
  - No CoT supervision needed—removes expensive distillation step
  - Innovative RLAIF-based caption reward and Cosine annealing KL coefficient
  - Clear methodology, open-source commitment

- Weaknesses:
  - Only 3B parameters—leaves scalability questions unanswered
  - Model judges its own captions via its LLM component
  - Data quality unexplored with no filtering nor curation
  - Comparison fairness concerns of 3B vs. likely much larger commercial models (GPT-4o, Claude)
  - Insufficient failure analysis of when does the approach fail or what reasoning types remain challenging
  - Does not include the latency/cost implications for deployment

Discussion Points & Questions

- Shortcut Learning Universality
  - Is this phenomenon specific to vision-language tasks?
  - Do pure LLMs exhibit similar shortcuts in reasoning?
  - What makes VLMs more susceptible?
- Caption vs. Reasoning Trade-offs
  - Could enforcing detailed captions constrain reasoning flexibility?
  - Is there unnecessary verbosity?
  - Where's the optimal balance?
- Alternative Grounding Mechanisms
  - Beyond captioning: scene graphs, object detection, segmentation masks?
  - Could multimodal embeddings provide implicit grounding?
  - What's the minimal grounding needed?

VL-Rethinker: Incentivizing Self-Reflection of VLMs with Reinforcement Learning

H. Wang et. al., Univ. of Waterloo | Poster @ NeurIPS '25

Georgia Tech

Agenda

- Inspiration
  - DeepSeek-R1-Zero
- Constraints in adapting GRPO to VLMs
- Modifications to GRPO
  - Selective Sample Replay
  - Forced Rethinking
- Evaluation
- Ablation tests
- Discussion

DeepSeek-R1-Zero

We directly apply RL to the base model without relying on supervised fine-tuning (SFT) as a preliminary step. This approach allows the model to explore chain-of-thought (CoT) for solving complex problems, resulting in the development of DeepSeek-R1-Zero. DeepSeek-R1-Zero demonstrates capabilities such as self-verification, reflection, and generating long CoTs, marking a significant milestone for the research community. Notably, it is the first open research to validate that reasoning capabilities of LLMs can be incentivized purely through RL, without the need for SFT. This breakthrough paves the way for future advancements in this area.

A description of methodology and emerging capabilities of DeepSeek-R1-Zero

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

Table 1 | Template for DeepSeek-R1-Zero. prompt will be replaced with the specific reasoning question during training.

Group Relative Policy Optimization In order to save the training costs of RL, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which foregoes the critic model that is typically the same size as the policy model, and estimates the baseline from group scores instead. Specifically, for each question q, GRPO samples a group of outputs $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_θ by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right), \quad (1)$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \quad (2)$$

$$A_i = \frac{r_i - mean(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}.$$

GRPO cost function and advantages formula

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

Table 2 | Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.

DeepSeek V3 671B: 15.6%

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a-\sqrt{a+x}} = x$ is equal to

Response: <think>
To solve the equation $\sqrt{a-\sqrt{a+x}} = x$, let's start by squaring both $\cdots$
$\left(\sqrt{a-\sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.
Rearrange to isolate the inner square root term:
$(a-x^2)^2 = a+x \implies a^2 - 2ax^2 + (x^2)^2 = a+x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
$\cdots$
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$
We started with the equation:

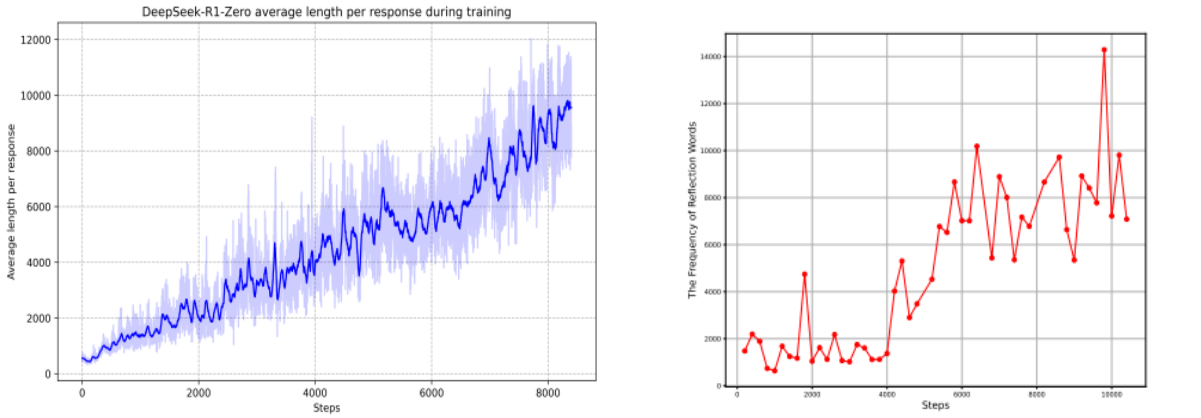


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

Training: 146k prompts (math, code, STEM, logic, general) with 64*8 H800 GPUs for 198 hours

Adapting GRPO to VLMs

Vanishing Advantages

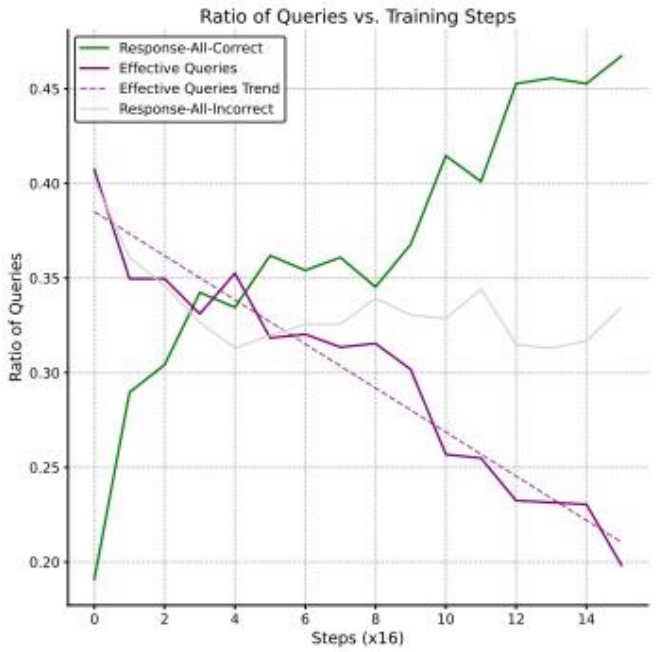


Figure 2: Illustration of the Vanishing Advantages problem. Training of 72B rapidly saturates, leading to a significant decrease of effective queries to only 20% within 256 steps.

- Training: 38k prompts (math, charts, STEM, spatial reasoning) with base model Qwen-VL-2.5-72B
- Uniform rewards within group results in zero advantage
  - Base model training w/o reasoning
  - Insufficient batch size, diversity in prompts, etc.

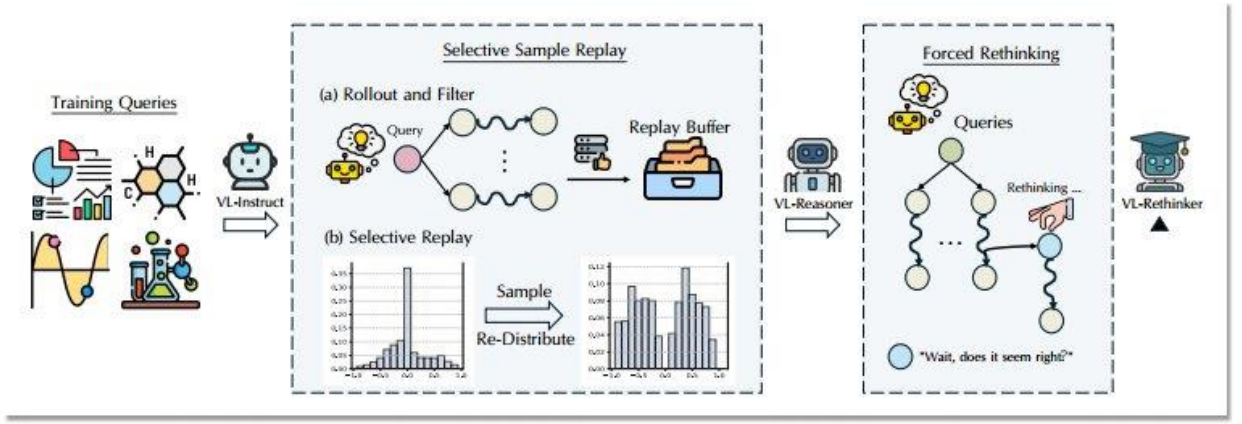


Figure 4: **Method Overview.** We present a two-stage RL method based on Qwen2.5-VL-Instruct. The first stage enhances general reasoning through GRPO with Selective Sample Replay (SSR), which retains explored trajectories with non-zero advantages and selectively replay samples based on their advantages. The second stage promotes deliberate reasoning using forced rethinking, where we append a specific rethinking trigger.

Selective Sample Replay

Algorithm 1 Selective Sample Replay (SSR)

Input: Buffer $\mathcal{B}_{\text{replay}}$, raw training batch $\mathcal{D}_{\text{raw}} = \{(x_i, y_i, \hat{A}_i)\}$, intensity $\alpha \geq 0$.
Output: Training batch $\mathcal{D}_{\text{train}}$, updated buffer $\mathcal{B}_{\text{replay}}$
Let $N_{\text{batch}} = |\mathcal{D}_{\text{raw}}|$
Initialize list for effective current samples $\mathcal{D}_{\text{effective}} \leftarrow \emptyset$
for each sample $(x_i, y_i, \hat{A}_i)$ in $\mathcal{D}_{\text{raw}}$ **do**
  Add $(x_i, y_i, \hat{A}_i)$ to $\mathcal{D}_{\text{effective}}$ when $|\hat{A}_i| > 0$
end for
Update buffer: $\mathcal{B}_{\text{replay}} \leftarrow \mathcal{B}_{\text{replay}} \cup \mathcal{D}_{\text{effective}}$
Let $n_{\text{effective}} = |\mathcal{D}_{\text{effective}}|$
Calculate number of samples needed from buffer: $n_{\text{from\_buffer}} = \max(0, N_{\text{batch}} - n_{\text{effective}})$
Initialize list for samples from buffer $\mathcal{D}_{\text{from\_buffer}} \leftarrow \emptyset$
if $n_{\text{from\_buffer}} > 0$ **then**
  Calculate sampling probabilities $P(\text{select } j)$ for all $j \in \mathcal{B}_{\text{replay}}$ according to Eq. 1
  Form $\mathcal{D}_{\text{from\_buffer}}$ by drawing $n_{\text{from\_buffer}}$ samples from $\mathcal{B}_{\text{replay}}$
end if
$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{effective}} \cup \mathcal{D}_{\text{from\_buffer}}$

$$P(\text{select } j) = \frac{|\hat{A}_j|^\alpha}{\sum_{k \in \mathcal{B}_{\text{replay}}} |\hat{A}_k|^\alpha}$$

- Similar to Prioritized Experience Replay (ICLR '16)
- Improves convergence rate with sparse rewards

Georgia Tech

Forced Rethinking

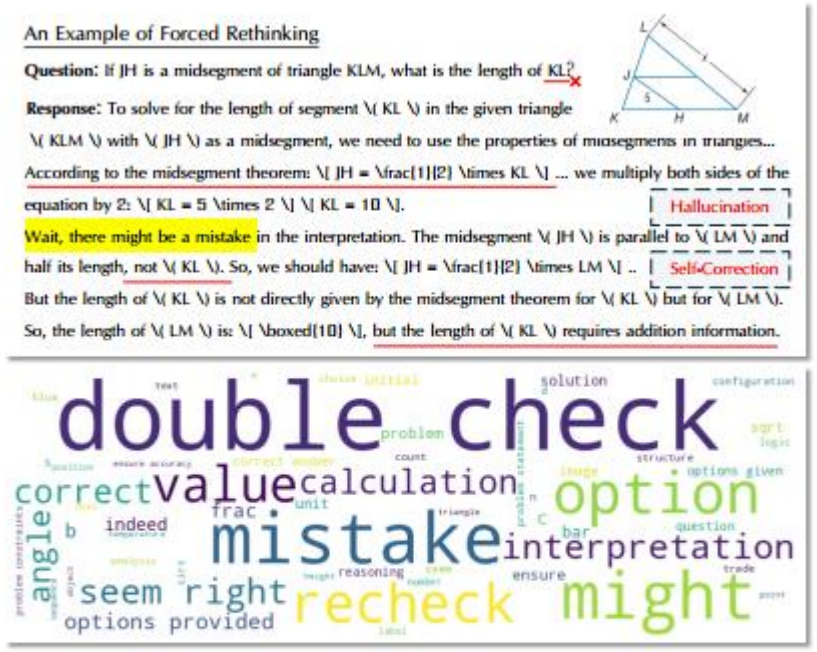


Figure 3: An example of Forced Rethinking (Top). VL-Rethinker discovers a flawed problem via rethinking upon its hallucinations. The word cloud of VL-Rethinker (Bottom) shows the learned rethinking pattern of self-verification, self-correction and self-questioning.

Forced Rethinking employs two means to stimulate the model's deliberate reasoning. The first, a straightforward means, involves a hint within the instruction prompt itself, e.g., "regularly perform self-reflection on your ongoing reasoning". This contextual cue serves to increase the model's propensity for generating rethinking sequences. The core principle of Forced Rethinking, however, lies in a targeted intervention within the RL rollout procedure, as depicted in Fig. 4. Following the VLM's initial generation of a response y_1 to a given input x, we append a specific textual "rethinking trigger" to y_1. This augmented sequence is then fed back into the model, urging it to generate a subsequent response segment y_2. Consequently, the complete generated sequence becomes $y = y_1 \oplus \text{trigger} \oplus y_2$. To elicit a diverse range of reasoning behaviors, we designed three distinct categories of triggers: self-verification, self-correction, and self-questioning. Detailed descriptions of these rethinking triggers are provided in the appendix.

This approach functions as a form of guided exploration [Wang et al., 2025], but it carries the inherent risk of disrupting the policy's native distribution. To mitigate this, we apply this forced rethinking to only a fraction $q < 1$ of the generated responses. Furthermore, we retain only those rethinking trajectories that lead to a correct final answer. Based on these successful forced rethinking trajectories, we incorporate an additional Supervised Fine-tuning (SFT) loss, which directly incentivizes the model to generate the desired deliberate thinking patterns.

Rethinking Triggers

```
self_questioning = "\n\nWait, does it seem right?"
self_correction = "\n\nWait, there might be a mistake"
self_verification = "\n\nWait, let's double check"
```

Georgia Tech

Evaluation

Model	Math-Related			Multi-Discipline			Real-World
	MathVista testmini	MathVerse testmini	MathVision test	MMMU-Pro overall	MMMU val	EMMA full	MEGA core
Proprietary Model							
OpenAI-o1	73.9	57.0	60.3	62.4	78.2	45.7	56.2
OpenAI-GPT-4o	60.0	41.2	30.6	51.9	69.1	32.7	52.7
Claude-3.5-Sonnet	67.7	47.8	33.5	51.5	68.3	35.1	52.3
Gemini-2.0-Flash	73.4	54.6	41.3	51.7	70.7	33.6	54.1
Open-Source Models							
Llama4-Scout-109B	70.7	-	-	52.2	69.4	24.6	31.8
InternVL-2.5-78B	72.3	51.7	34.9	48.6	61.8	27.1	44.1
QvQ-72B	71.4	48.6	35.9	51.5	70.3	32.0	8.8
LLava-OV-72B	67.5	39.1	30.1	31.0	56.8	23.8	29.7
Qwen-2.5-VL-32B	74.7	48.5	38.4	49.5	†59.4	31.1	13.3
Qwen-2.5-VL-72B	74.8	57.2	38.1	51.6	†67.0	34.1	49.0
VL-Rethinker-32B	78.8	56.9	40.5	50.6	65.6	37.9	19.9
VL-Rethinker-72B	**80.4**	**63.5**	**44.9**	**55.9**	**68.8**	**38.5**	**51.3**
Δ (Ours - Open SoTA)	+5.6	+6.3	+6.8	+3.7	-1.4	+4.4	+2.3

Table 1: Comparison between our 72B model and other state-of-the-art models. The notation of † indicates reproduced results using our evaluation protocols.

Model	Math-Related			Multi-Discipline			Real-World
	MathVista testmini	MathVerse testmini	MathVision test	MMMU-Pro overall	MMMU val	EMMA full	MEGA core
General Vision-Language Models							
InternVL2-8B	58.3	-	17.4	29.0	51.2	19.8	26.0
InternVL2.5-8B	64.4	39.5	19.7	34.3	56.0	-	30.4
QwenVL2-7B	58.2	-	16.3	30.5	54.1	20.2	34.8
QwenVL2.5-7B	68.2	46.3	25.1	36.9	†54.3	21.5	35.0
Llava-OV-7B	63.2	26.2	-	24.1	48.8	18.3	22.9
Kimi-VL-16B	68.7	44.9	21.4	-	†55.7	-	-
Vision-Language Reasoning Models							
MM-Eureka-8B (Intern)	67.1	40.4	22.2	27.8	49.2	-	-
MM-Eureka-7B (Qwen)	73.0	50.3	26.9	-	-	-	-
R1-VL-7B	63.5	40.0	24.7	7.8	44.5	8.3	29.9
R1-Onevision-7B	64.1	46.4	29.9	21.6	-	20.8	27.1
OpenVLThinker-7B	70.2	47.9	25.3	37.3	52.5	26.6	12.0
VL-Rethinker-7B	**74.9**	**54.2**	**32.3**	**41.7**	**56.7**	**29.7**	**37.2**
Δ (Ours - Prev SoTA)	+4.7	+6.3	+2.4	+4.4	+0.7	+3.1	+2.2

Table 2: Comparison between our 7B model and other general and reasoning vision-language models. † means that the results are reproduced by us.

- Not dramatic enough

Georgia Tech

Ablation Tests

Same as base model

Yeah, Science!

Model	RL-Algo	Data	MathVision	MathVista	MathVerse	MMMU-Pro	EMMA
VL-Rethinker-7B	SSR	16K	32.3	74.9	54.2	41.7	29.7
w/o 'Forced-Rethinking'	SSR	16K	29.8	72.4	53.2	40.9	29.5
- no SSR	Filter	16K	28.5	72.0	50.0	40.0	26.9
- no SSR& Filter	GRPO	16K	26.0	70.9	51.4	38.8	26.2
- no Text	SSR	13K	29.1	73.5	53.5	41.1	28.7
- no Science&Text	SSR	11K	28.0	71.6	50.3	39.7	28.0

Table 3: Ablation Results to show the impact of SSR and Data Mix.

- Test-time forced rethinking not effective
- Varying rethinking ratio across datasets for VL-Rethinker

(selects rethinking in hindsight)

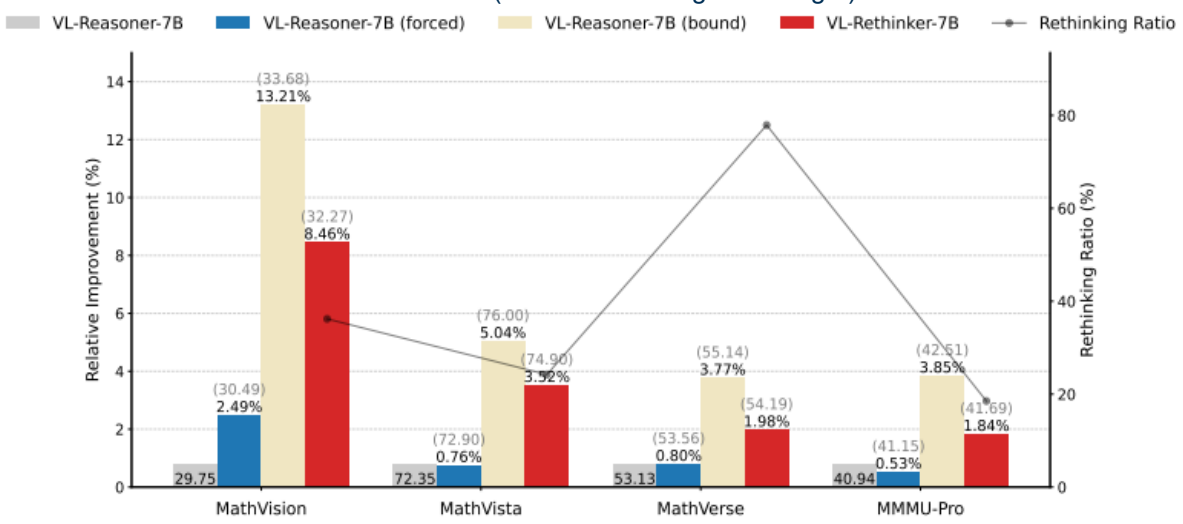


Figure 7: **Relative Improvement with Different Re-thinking Strategies.** We compare: (a) VL-Reasoner (forced), which is forced to rethink at test time; (b) VL-Reasoner (bound), represents the upper bound of test-time forced re-thinking; and (c) VL-Rethinker is trained for self-reflection. The results indicate that forcing VL-Reasoner to rethink at test time yields positive performance gains. Training for self-reflection significantly enhances performance, achieving closer results to the upper bound of forced re-thinking. The overlaid line plot shows the rethinking ratio (right y-axis) of VL-Rethinker across different benchmarks, showing VL-Rethinker adaptively performs re-thinking, unlike the fixed forced re-thinking strategy.

Georgia Tech

Discussion

- Multi-stage pipeline with cold start data, SFT (DeepSeek-R1)
- Retrieve reasoning traces from LLMs using image captioner (R1-Onevision)

Multimodal Chain-of-Thought Reasoning in Language Models

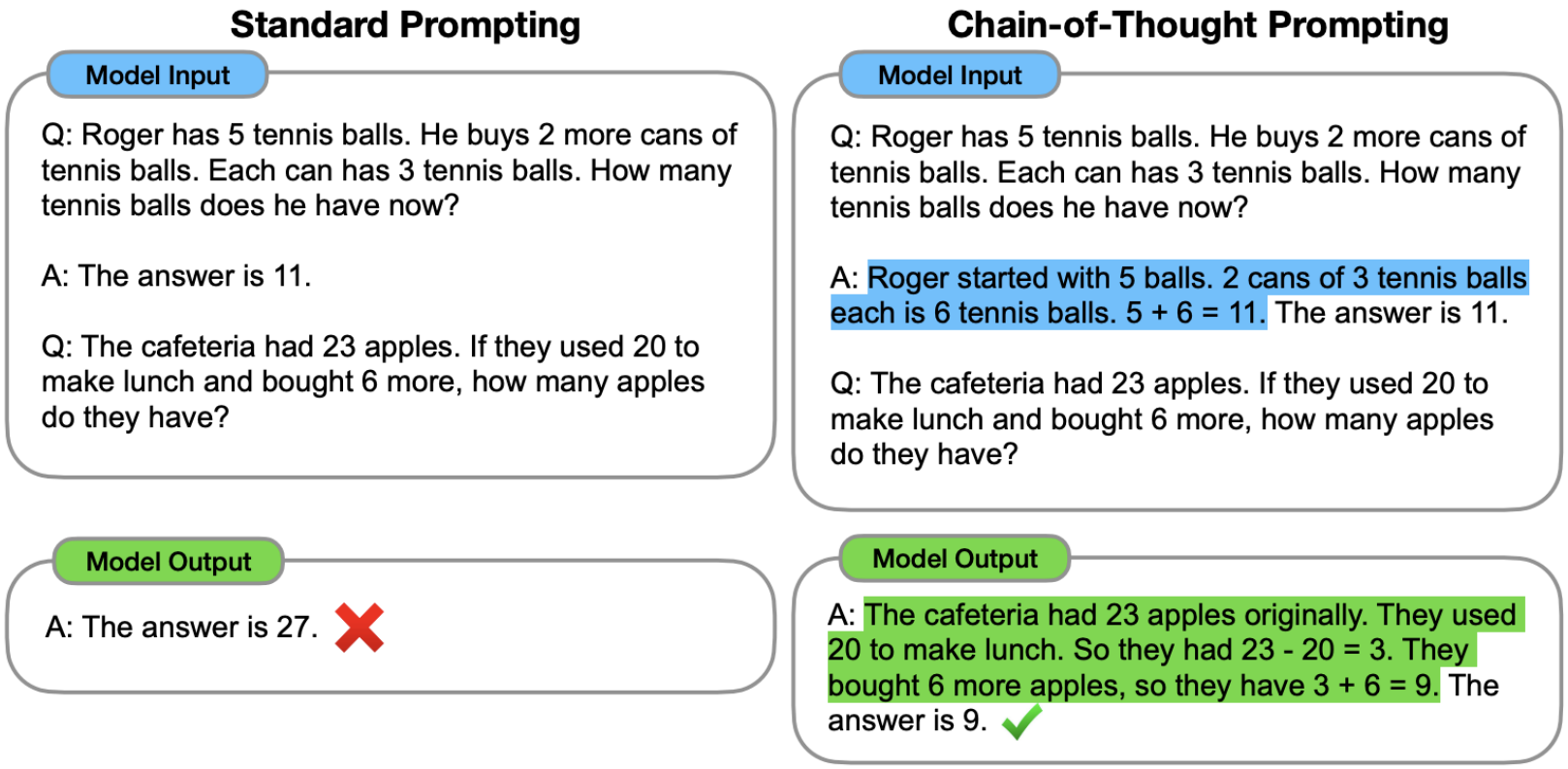
Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, Alex Smola

TMLR

Georgia Tech

Chain-of-Thought (CoT) Prompting

- LLMs can improve reasoning by generating intermediate steps

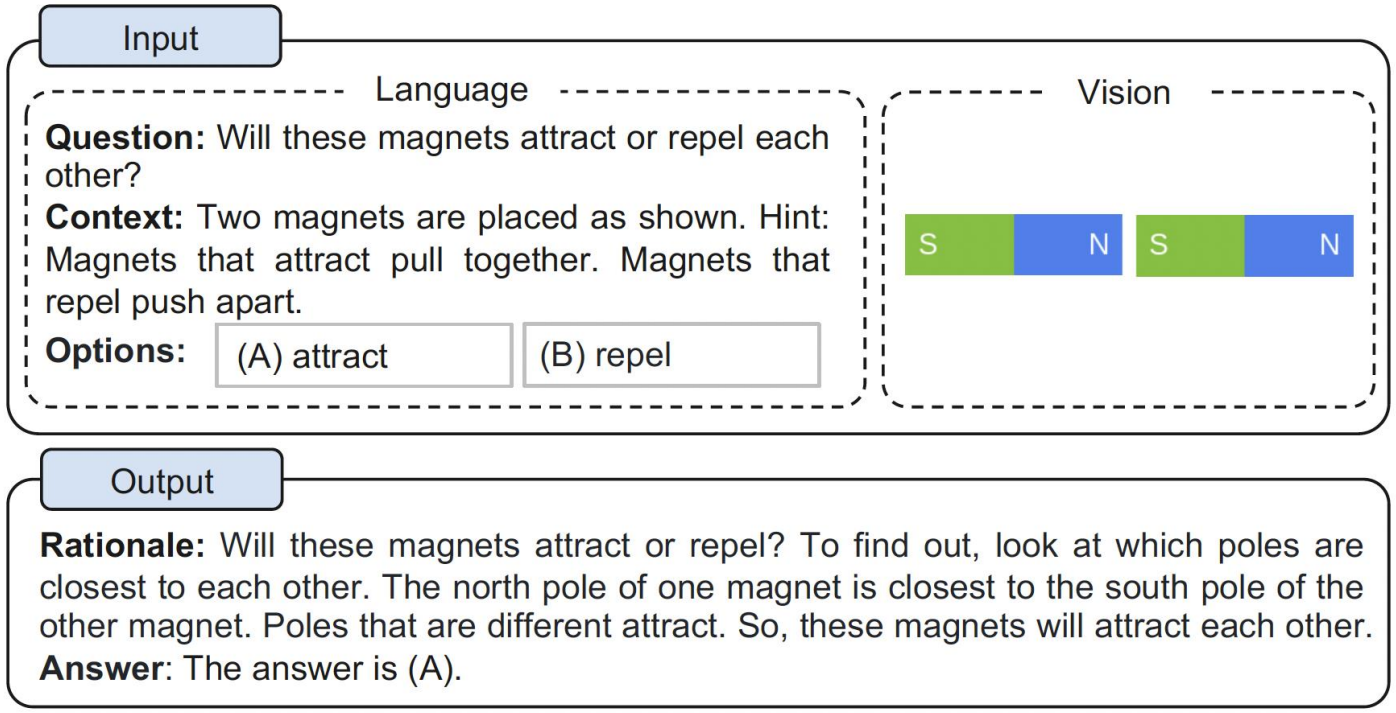


Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

Georgia Tech

Multimodal CoT Prompting

- Traditional CoT focuses on **language modality** alone
- Real-world reasoning requires both text and visual information
  - e.g., Understanding science questions requires reading text AND interpreting diagrams
- Challenge: How to effectively combine vision and language for reasoning?



Georgia Tech

Two Primary Paradigms

- Prompting LLMs
  - Transform the input of different modalities into a unified modality and prompt LLMs to perform CoT
    - e.g., generate a caption for an image by a captioning model, and then concatenate the caption with the original language input to be fed into LLMs
  - Risk of **information loss** when transforming vision signals into textual descriptions.
- Fine-tuning smaller models
  - Fine-tune smaller LMs by fusing multimodal features
  - Challenge: Smaller models (<100B parameters) struggle with hallucination

Challenge of Multimodal-CoT

- Fine-tune a text-only baseline (FLAN-Alpaca$_{Base}$) for CoT reasoning on ScienceQA
- The model takes the concatenation of tokens of
  - Question text (Q)
  - Context text (C)
  - Multiple options (M)
- Three formats
  - **No-CoT:** predicts the answer directly (QCM→A)
  - **Reasoning:** answer inference is conditioned to the rationale (QCM→RA)
  - **Explanation:** rationale is used for explaining the answer inference (QCM→AR)
- **Results:** rationales might not necessarily contribute to predicting the right answer, why?

Method	Format	Accuracy
No-CoT	QCM→A	81.63
Reasoning	QCM→RA	69.32
Explanation	QCM→AR	69.68

Georgia Tech

Challenge of Multimodal-CoT

- Separate the CoT problem into two stages
  - Rationale generation (RougeL: 90.73%)
  - Answer inference (Accuracy: 78.57%)
- **Reason:** model tends to generate hallucinated rationales that mislead the answer inference

Method	(i) QCM→ R	(ii) QCMR→ A
Two-Stage Framework	90.73	78.57

Problem

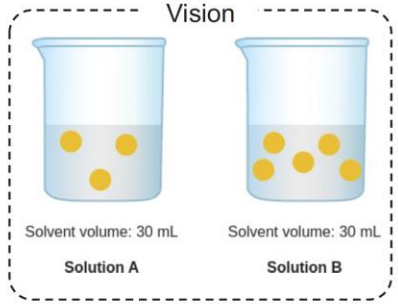
Question: Which solution has a higher concentration of yellow particles?

Context: The diagram below is a model of two solutions. Each yellow ball represents one particle of solute.

Options: (A) Solution B | (B) Solution A | (C) neither; their concentrations are the same

Gold Rationale: In Solution A and Solution B, the yellow particles represent the solute. To figure out which solution has a higher concentration of yellow particles, look at both the number of yellow particles and the volume of the solvent in each container. Use the concentration formula to find the number of yellow particles per milliliter. Solution B has more yellow particles per milliliter. So, Solution B has a higher concentration of yellow particles.

Answer: The answer is (A).



Baseline

Rationale: In Solution A and Solution B, the yellow particles represent the solute. To figure out which solution has a higher concentration of yellow particles, look at both the number of yellow particles and the volume of the solvent in each container. Use the concentration formula to find the number of yellow particles per milliliter. Solution A has more yellow particles per milliliter. So, Solution A has a higher concentration of yellow particles.

Answer: The answer is (B).

+ Vision Features

Rationale: In Solution A and Solution B, the yellow particles represent the solute. To figure out which solution has a higher concentration of yellow particles, look at both the number of yellow particles and the volume of the solvent in each container. Use the concentration formula to find the number of yellow particles per milliliter. Solution B has more yellow particles per milliliter. So, Solution B has a higher concentration of yellow particles.

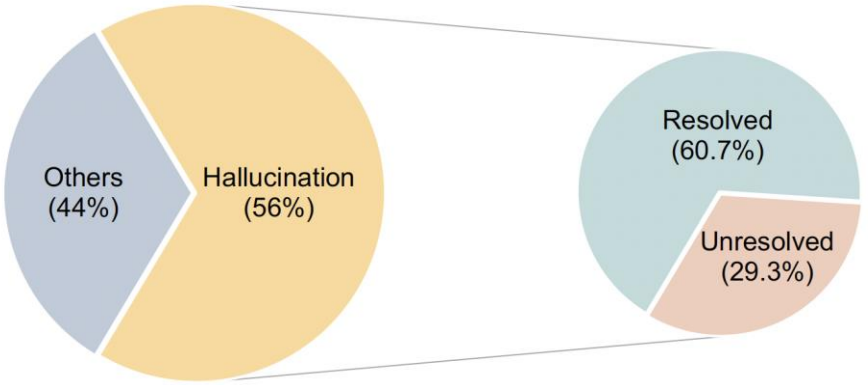
Answer: The answer is (A).

Georgia Tech

Can Multimodality Contributes to Effective Rationales?

- **Hypothesis:** Hallucination is due to a lack of necessary vision contexts
- **Solution 1 (basic)**
  - transform the image into a caption and then append the caption in the input of both stages
  - yields marginal performance gains (↑0.80%)
- **Solution 2 (advanced)**
  - Feed the image to ViT to extract vision features, and then fuse the vision features with the encoded language representations before feeding the decoder
  - RougeL score of the rationale generation: 90.73% -> 93.46%
  - Answer accuracy of 78.57% -> 85.31%
  - 60.7% hallucination mistakes have been corrected

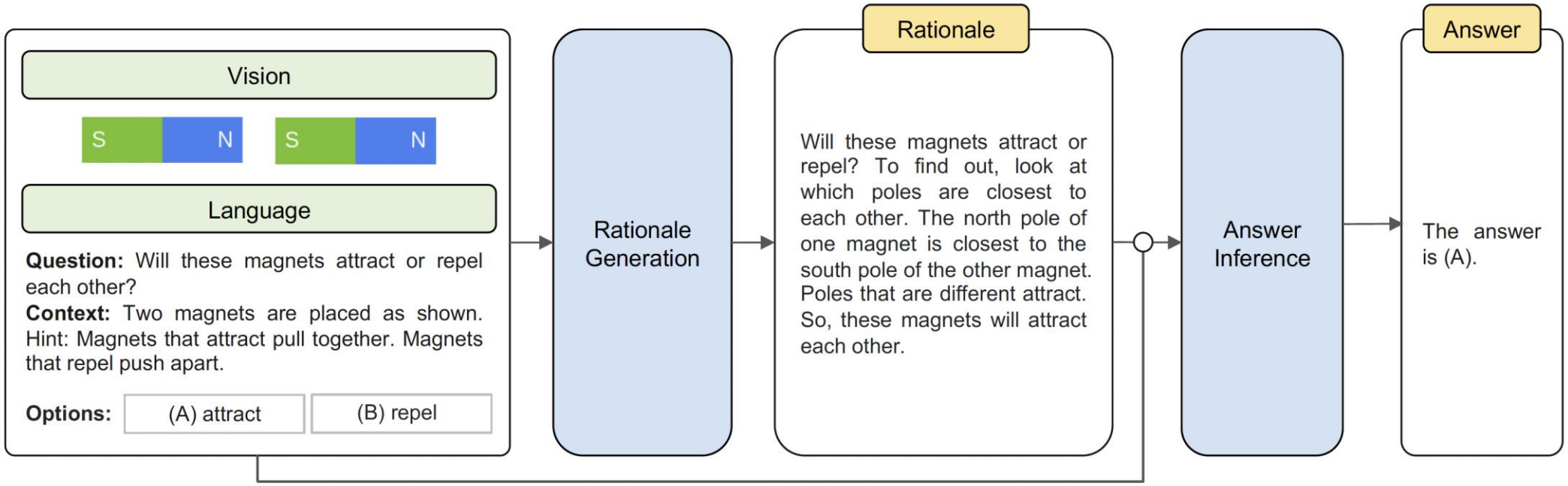
Method	(i) QCM→ R	(ii) QCMR→ A
Two-Stage Framework	90.73	78.57
w/ Captions	90.88	79.37
w/ Vision Features	93.46	85.31



(a) ratio of hallucination mistakes (b) correction rate w/ vision features

Multimodal-CoT: Framework

- **Two stages**
  - **Rationale generation:** language and vision inputs -> rationales
  - **Answer inference:** language and vision inputs + rationales -> answer
- Both stages share the same model structure but differ in the input and output



Georgia Tech

Multimodal-CoT: Model Architecture

- **Encoding**
  - Language: Transformer encoder for text
  - Vision: Frozen ViT-large for patch-level image features
  - Output: Aligned representations (H_language, H_vision)

$$
\begin{aligned}
H_{\text{language}} &= \text{LanguageEncoder}(X_{\text{language}}), \\
H_{\text{vision}} &= W_h \cdot \text{VisionExtractor}(X_{\text{vision}}),
\end{aligned}
$$

- **Interaction**
  - Single-head attention between text tokens and image patches
  - Gated fusion mechanism combining language and vision representations

- **Decoding**
  - Transformer decoder generates target text

$$
\begin{aligned}
\lambda &= \text{Sigmoid}(W_l H_{\text{language}} + W_v H^{\text{attn}}_{\text{vision}}), \\
H_{\text{fuse}} &= (1-\lambda) \cdot H_{\text{language}} + \lambda \cdot H^{\text{attn}}_{\text{vision}},
\end{aligned}
$$

Georgia Tech

Experiments: Setup

- **Datasets**
  - **ScienceQA**: 21k multimodal multiple-choice questions (science domain)
  - **A-OKVQA:** 25k knowledge-based VQA questions
- **Implementation Details**
  - **Backbone:** T5 (Base 200M, Large 700M)
  - **Vision features:** Frozen ViT-large

Experiments: Setup

Example of Science QA

- **Question:** Which animal's mouth is also adapted for bottom feeding?
- **Options:** ["discus", "armored catfish"]

- **Answer:** 1 ("armored catfish")
- **Subject:** natural science

Example of Science QA

- **Question:** What is the occupation of the person driving?
- **Options:** ["waiter", "farmer", "cashier", "musician"]

- **Answer:** 1 ("farmer")

Georgia Tech

Experiments: Main Results - ScienceQA

- **Key Results**
  - Multimodal-CoT Large (738M): 90.45%
  - Previous best published: 86.54%
  - Human performance: 88.40%

Model	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Human	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
MCAN (Yu et al., 2019)	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down (Anderson et al., 2018)	70M	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN (Kim et al., 2018)	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF (Gao et al., 2019)	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT (Kim et al., 2021)	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM (Lu et al., 2021)	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT (Li et al., 2019)	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA (Lu et al., 2022a)	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
GPT-3.5 (text-davinci-002) (Lu et al., 2022a)	173B	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
GPT-3.5 (text-davinci-003)	173B	77.71	68.73	80.18	75.12	67.92	81.81	80.58	69.08	76.47
ChatGPT (Lu et al., 2023)	-	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03	78.31
GPT-4 (Lu et al., 2023)	-	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04	83.99
Chameleon (ChatGPT) (Lu et al., 2023)†	-	81.62	70.64	84.00	79.77	70.80	86.62	81.86	76.53	79.93
Chameleon (GPT-4) (Lu et al., 2023)†	-	89.83	74.13	89.82	88.27	77.64	92.13	88.03	83.72	86.54
LLaMA-Adapter (Zhang et al., 2023a)†	6B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaVA (Liu et al., 2023)†	13B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
InstructBLIP (Dai et al., 2023)†	11B	-	-	-	-	90.70	-	-	-	
Mutimodal-CoT$_{\text{Base}}$	223M	84.06	92.35	82.18	82.75	82.75	84.74	85.79	84.44	85.31
Mutimodal-CoT$_{\text{Large}}$	738M	91.03	93.70	86.64	90.13	88.25	89.48	91.12	89.26	90.45

Georgia Tech

Experiments: Main Results - A-OKVQA

- **Multimodal-CoT Base:** 50.57%
- Baseline comparisons
  - Language-only: 47.86%
  - Improvement: +2.71% over language-only baseline
- Shows the approach generalizes beyond ScienceQA

Model	Accuracy
BERT	32.93
GPT-3 (Curie)	35.07
IPVR (OPT-66B)	48.6
ViLBERT	49.1
Language-only Baseline	47.86
Multimodal-CoT$_{\text{Base}}$	50.57

Georgia Tech

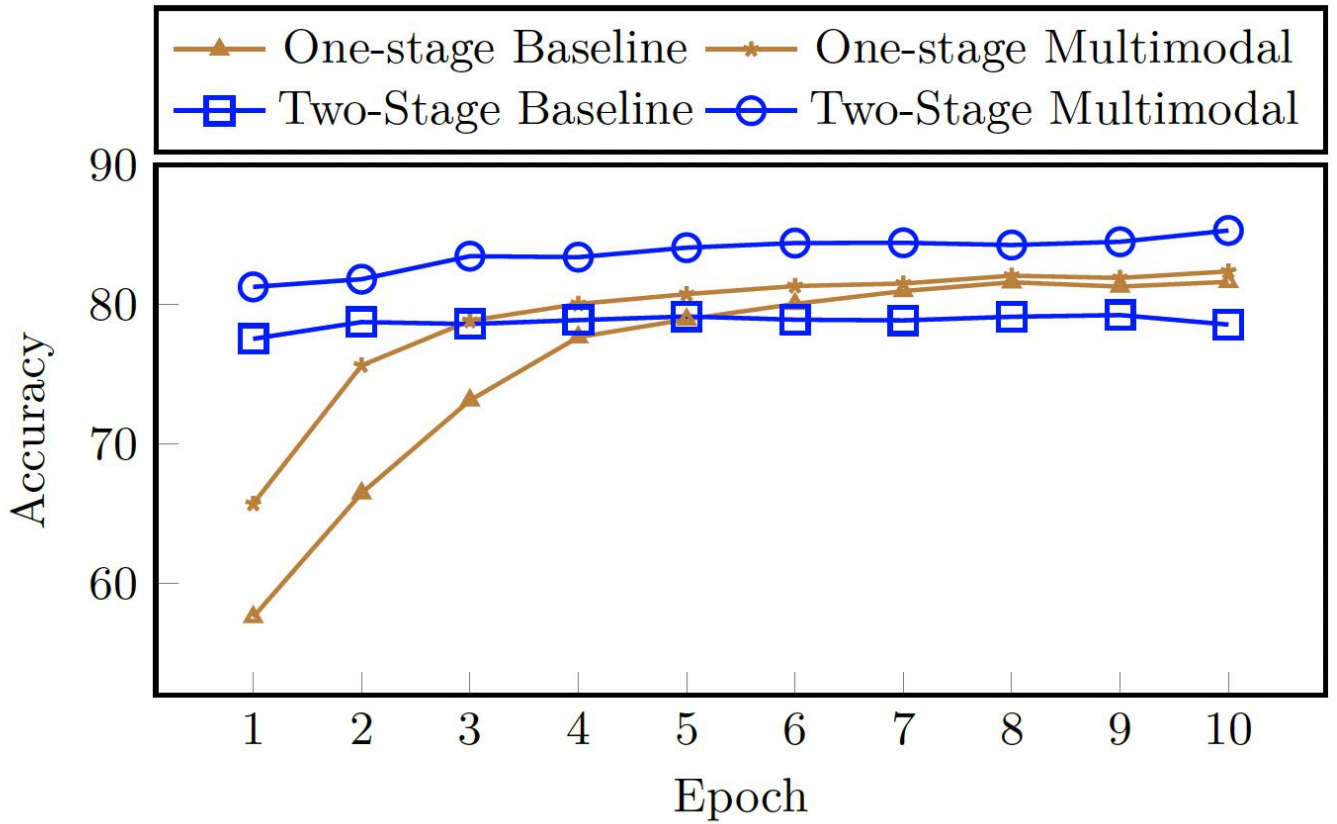
Ablation Study

- **Full model:** 85.31%
  - Without two-stage framework: 82.62% (-2.69%)
  - Without vision features: 78.57% (-6.74%)
- Both components essential for strong performance

Model	Base	Large
Multimodal-CoT	85.31	90.45
w/o Two-Stage Framework	82.62	84.56
w/o Vision Features	78.57	83.97

Analysis - Convergence Boost

- Two-stage methods achieve higher accuracy earlier in training
- Vision features enable faster convergence to better performance
- **Practical benefit:** Reduced training time needed



Georgia Tech

Analysis - Scaling to Large Models

- Leverage InstructBLIP and ChatGPT to generate rationales
- Trains with generated rather than human-annotated rationales
- **Results with generation:** 87.76% (vs. 90.45% w/ annotation)
- **Implication:** Approach scales to domains without existing annotated reasoning chains

Model	IMG	TXT	AVG
InstructBLIP	60.50	-	-
ChatGPT	56.52	67.16	65.95
Multimodal-CoT w/ Annotation	88.25	90.13	90.45
Multimodal-CoT w/ Generation	83.54	85.73	87.76

Georgia Tech

Analysis - Robustness

- **Backbone Model Generalization:** Approach benefits multiple backbone architectures
- **Vision Feature Comparison:** ViT superior but other features also effective

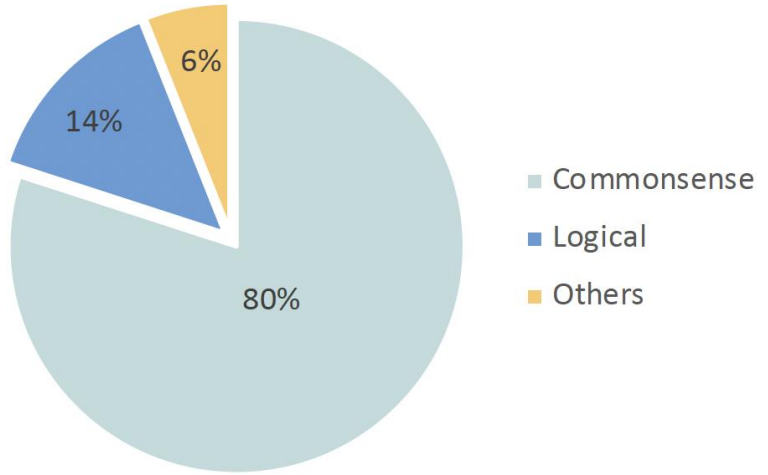
Method	Accuracy
Prior Best (Lu et al., 2022a)	75.17
MM-CoT on UnifiedQA	82.55
MM-CoT on FLAN-T5	83.19
MM-CoT on FLAN-Alpaca	85.31

Feature	Feature Shape	Accuracy
ViT	(145, 1024)	85.31
CLIP	(49, 2048)	84.27
DETR	(100, 256)	83.16
ResNet	(512, 2048)	82.86

Georgia Tech

Error Analysis (50 error cases)

- **Commonsense mistakes:** 80%
  - e.g., map interpretation, object counting
- **Logical mistakes:** 14%
  - e.g., comparison errors, contradictions in reasoning
- **Other:** 6%
  - e.g., CoT empty
- **Implications:** Future work should focus on visual feature quality and commonsense knowledge integration



Problem

Question: Which of these cities is marked on the map?

Context: N/A

Options: (A) Las Vegas | (B) San Francisco | (C) Phoenix | (D) Salt Lake City

Rationale : The city is Las Vegas, Nevada. Salt Lake City, Phoenix, and San Francisco are marked with gray circles on the map below.

Answer: The answer is (A).

Vision

Prediction

Rationale: The city is San Francisco, California. Salt Lake City, Las Vegas, and Phoenix are marked with gray circles on the map below.

Answer: The answer is (B).

Georgia Tech

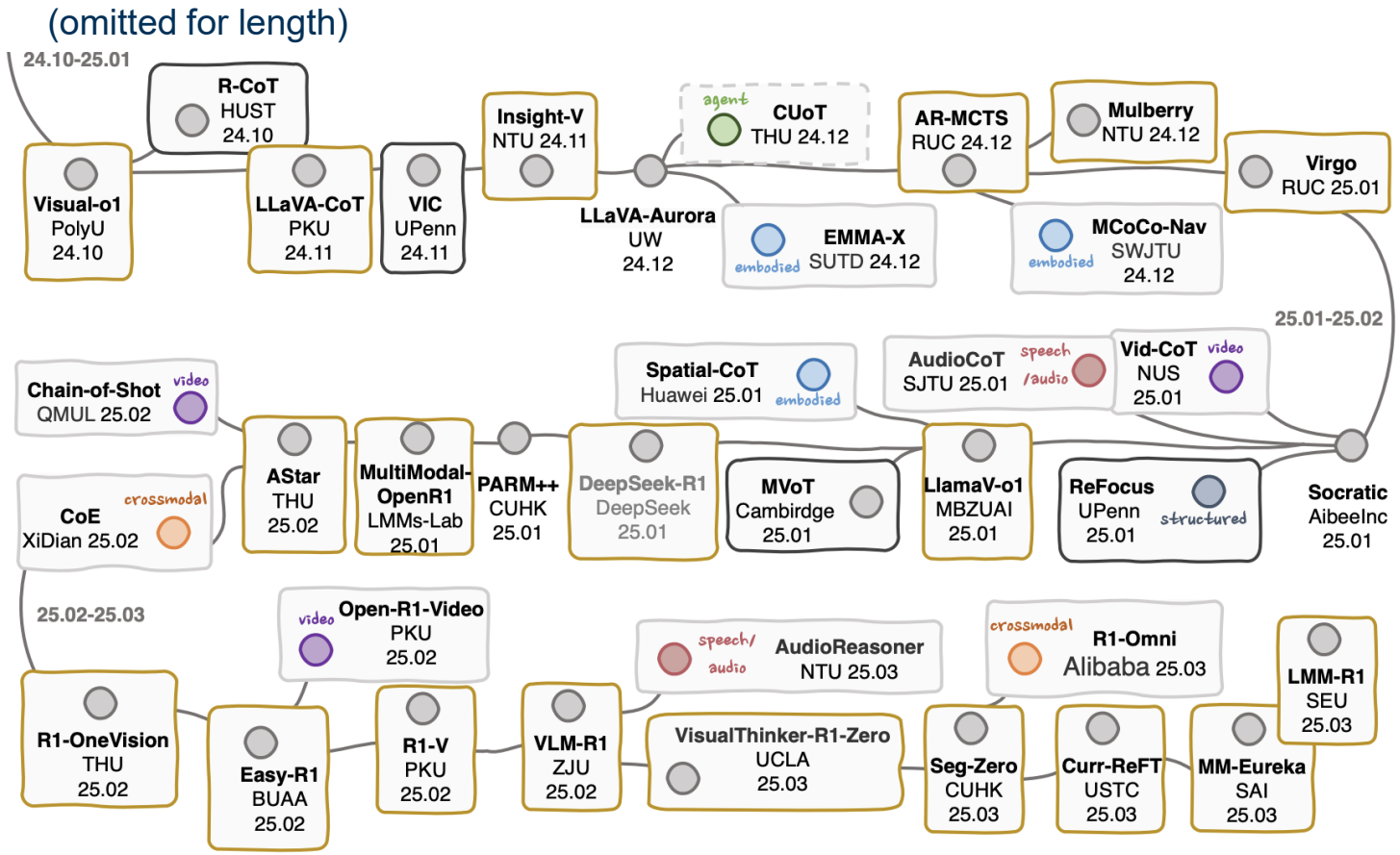
Discussion

- **Strengths**
  - Two-stage framework:
    - Rationale generation (vision + language)
    - Answer inference (uses generated rationale)
  - Analysis of why naive CoT fails and how vision features solve it
- **Weaknesses**
  - Error analysis only on 50 samples
  - Frozen ViT features, no end-to-end vision-language optimization
- **Limitations**
  - **Commonsense gap**: 80% errors need external knowledge (e.g., maps, counting)
  - **Domain-specific**: Evaluated primarily on educational/scientific QA

Discussion

- **Implications**
  - Multimodal information is crucial for grounding reasoning chains
  - Two-stage design is useful for multimodal reasoning tasks
- **Future Works**
  - Integrate commonsense knowledge bases
  - Extend to other modality pairs
  - Apply to more reasoning-heavy tasks

Recent Development in Multimodal CoT Reasoning



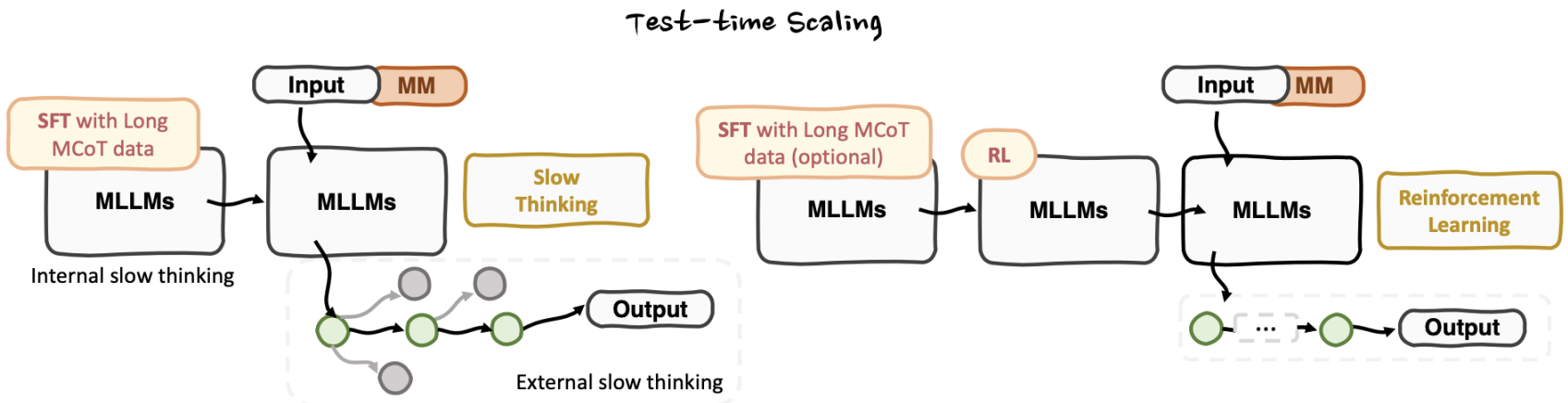
Wang, Yaoting, et al. "Multimodal chain-of-thought reasoning: A comprehensive survey." *arXiv preprint arXiv:2503.12605* (2025).

Georgia Tech

Related Works - MCoT Methodologies

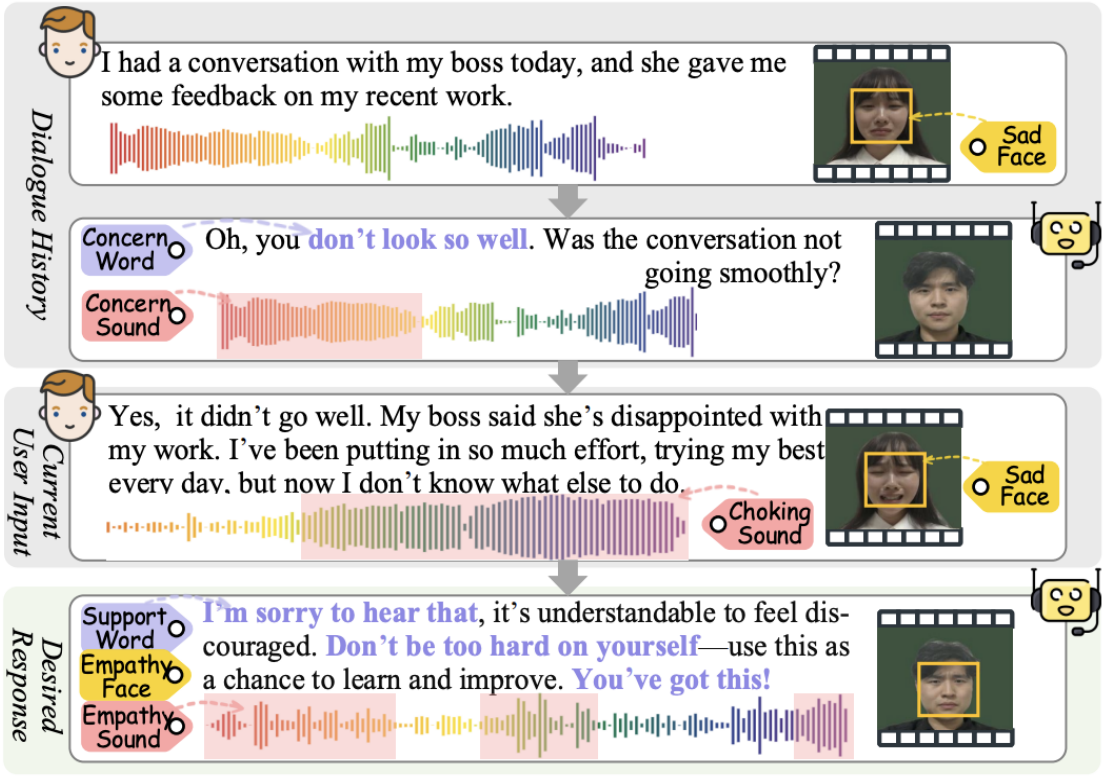
- **Prompt-based**
  - Zero- / Few-shot
- **Plan-based**
  - Dynamic exploration during reasoning
  - Debate-based reasoning
- **Learning-based**
  - Fine-tuning with reasoning data
  - Trend: Shift toward test-time scaling post-OpenAI o1

Model	Foundational LLMs	Modality	Learning	Cold Start	Algorithm	Aha-moment
Deepseek-R1-Zero [137]	Deepseek-V3	T	RL	✗	GRPO	✔
Deepseek-R1 [137]	Deepseek-V3	T	SFT+RL	✔	GRPO	-
LLaVA-Reasoner [138]	LLaMA3-LLaVA-NEXT-8B	T,I	SFT+RL	✔	DPO	-
Insight-V [91]	LLaMA3-LLaVA-NEXT-8B	T,I	SFT+RL	✔	DPO	-
Multimodal-Open-R1 [99]	Qwen2-VL-7B-Instruct	T,I	RL	✗	GRPO	✗
R1-OneVision [101]	Qwen2.5-VL-7B-Instruct	T,I	SFT	-	-	-
R1-V [237]	Qwen2.5-VL	T,I	RL	✗	GPRO	✗
VLM-R1 [238]	Qwen2.5-VL	T,I	RL	✗	GPRO	✗
LMM-R1 [239]	Qwen2.5-VL-Instruct-3B	T,I	RL	✗	PPO	✗
Curr-ReFT [244]	Qwen2.5-VL-3B	T,I	RL+SFT	✗	GPRO	-
Seg-Zero [245]	Qwen2.5-VL-3B + SAM2	T,I	RL	✗	GPRO	✗
MM-Eureka [246]	InternVL2.5-Instruct-8B	T,I	SFT+RL	✔	RLOO	-
MM-Eureka-Zero [246]	InternVL2.5-Pretrained-38B	T,I	RL	✗	RLOO	✔
VisualThinker-R1-Zero [247]	Qwen2-VL-2B	T,I	RL	✗	GPRO	✔
Easy-R1 [240]	Qwen2.5-VL	T,I	RL	✗	GRPO	-
Open-R1-Video [243]	Qwen2-VL-7B	T,I,V	RL	✗	GRPO	✗
R1-Omni [130]	HumanOmni-0.5B	T,I,V,A	SFT+RL	✔	GRPO	-
VisRL [248]	Qwen2.5-VL-7B	T,I	SFT+RL	✔	DPO	-
R1-VL [249]	Qwen2-VL-7B	T,I	RL	✗	StepGRPO	-



Related Works - Applications with MCoT Reasoning

- **Embodied AI**
  - Multi-robot coordination
  - Spatial Reasoning
- **Healthcare**
  - Surgical triplet recognition
  - Medical VQA
- **Human-centric**
  - Empathetic dialogue
  - Multimodal sentiment analysis
- **Agentic Systems**
  - Long video understanding
  - Multi-scene video generation
- And more



Zhang, Han, et al. "Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark." *Proceedings of the ACM on Web Conference 2025*. 2025.

Georgia Tech

Related Works - Key Challenges & Future Directions

- **Computational Sustainability**
  - Exponential growth in resources for long-MCoT
  - Balancing reasoning depth vs. cost
- **Limited Generalization**
  - Lack of robust reasoning in general scenarios (vs. math/science)
- **Error Propagation**
  - Small inaccuracies compound
- **Modality Imbalance**
  - Uneven progress across modalities (text/image > audio)
- **Other Open Problems**
  - Ethical and safety considerations

Thank you!

Queries? ToTs?

Georgia Tech