



# Virtual Poster Presentation



## See, Think, Learn: A Self-Taught Multimodal Reasoner



Codebase on GitHub



**Sourabh Sharma**  
Undergrad Researcher  
MNIT Jaipur



**Dr. Sonam Gupta**  
Research Scientist  
IBM Research



**Dr. Sadbhawna Thakur**  
Assistant Professor  
MNIT Jaipur

### Contact

Sourabh Sharma  
Malaviya National Institute of Technology Jaipur  
Email: [sourabh125ss@gmail.com](mailto:sourabh125ss@gmail.com)

Dr. Sonam Gupta  
IBM Research  
Email: [sonam.gupta7@ibm.com](mailto:sonam.gupta7@ibm.com)

Dr. Sadbhawna Thakur  
Malaviya National Institute of Technology Jaipur  
Email: [sadbhawna.cse@mnit.ac.in](mailto:sadbhawna.cse@mnit.ac.in)  
Website: <https://sadbhawna.thakur.github.io/>



# See, Think, Learn: A Self-Taught Multimodal Reasoner

Sourabh Sharma<sup>1</sup>, Sonam Gupta<sup>2</sup>, Sadbhawna Thakur<sup>1</sup>

<sup>1</sup>Malaviya National Institute of Technology Jaipur, <sup>2</sup>IBM Research



## Motivation

### Problem Statement:

- Reasoning enhancement in VLMs require **high quality CoT dataset**, obtained via **labor-intensive human annotations** or **costly proprietary models**. Self training algorithms presents a promising alternative for the same.
- Current self-training algorithms overlook the perception** component of multimodal reasoning. Further, the popular augmentation strategy **“positive rationalisation”** causes **shortcut learning**.



### Question:

Where is the location of the bench? Select the correct answer from the following choices:  
 (a) Backyard  
 (b) Beach  
 (c) Public park  
 (d) Courtyard

Correct Answer: (c)

STaR	<b>Rationale: The bench is located in a courtyard.</b> <b>Generated Choice: (d)</b>
R <sup>3</sup> V	<b>Rationale: The wooden bench is located in a courtyard.</b> <b>Generated Choice: (d)</b>
STL (Ours)	<b>Caption:</b> The image shows a wooden bench located in a garden setting. The bench is situated in a dirt area, and it is surrounded by plants. <b>Rationale:</b> The bench is located in a garden setting, which is typically found in public parks or backyards. The dirt area and the presence of plants suggest that it is in a garden, and the bench is likely meant for people to sit and enjoy the surroundings. <b>Generated Choice: (c)</b>

Figure illustrating how STL addresses the limitations of existing self-training frameworks.

### Contributions:

- We propose a **novel self-training framework “See-Think-Learn” (STL)** for VLMs that enforces **explicit visual grounding before reasoning** via a **structured rationale prompt**.
- The model **iteratively generates and learns** from its own structured rationales, **jointly improving perception and reasoning**.
- We introduce **negative rationales** that explain incorrect answer choices, **enhancing discriminative ability and robustness**.

## Contact

Sourabh Sharma  
 Malaviya National Institute of Technology Jaipur  
 Email: sourabh125ss@gmail.com

Dr. Sonam Gupta  
 IBM Research  
 Email: sonam.gupta7@ibm.com

Dr. Sadbhawna Thakur  
 Malaviya National Institute of Technology Jaipur  
 Email: sadbhawna.cse@mmit.ac.in  
 Website: <https://sadbhawnathakur.github.io/>

## Method

### Structured Rationale Prompt:

The prompt has three components:

- Caption**, which gives a **detailed description of the image** based on the question;
- Reasoning**, which involves a **detailed thought process grounded in visual details**; and
- Conclusion**, which gives the **final answer** based on reasoning.

### See-Think-Learn (STL) Self-Training Framework:

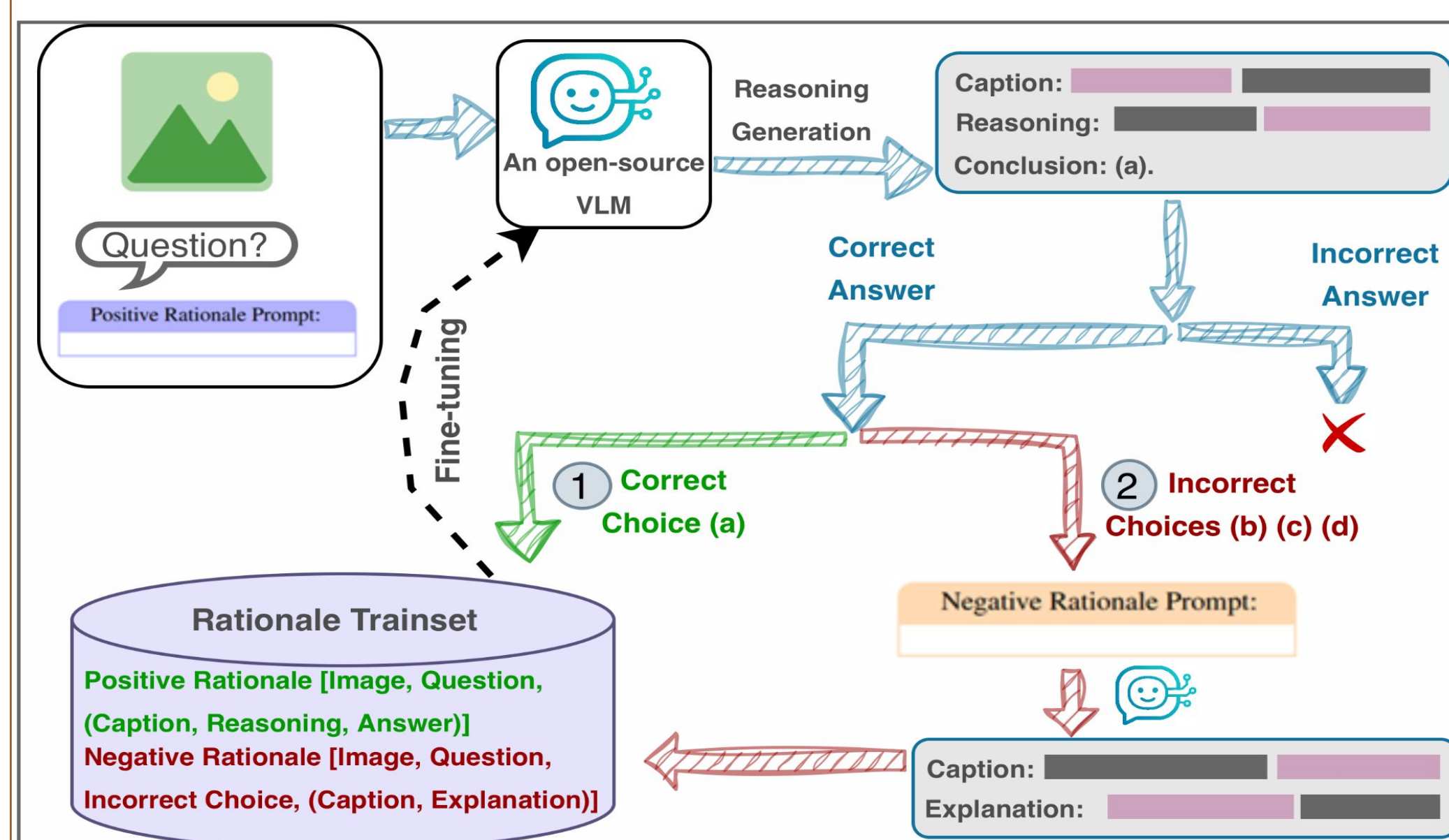


Figure illustrating the architecture of the proposed See-Think-Learn (STL) framework.

- We start with a multiple choice VQA dataset. An **open-source VLM** is prompted with the **Structured Rationale Prompt** to generate CoT rationale for each sample in the dataset.
- If the **conclusion of the generated rationale matches with the ground truth answer**, we generate **Positive Rationale** and **Negative Rationale** for the sample and include them in the **Rationale Trainset**.
- Base VLM is finetuned** on the generated Rationale Trainset.
- The fine-tuned model** is subsequently used to regenerate rationales for the dataset, enabling an **iterative self-improvement process**.
- Through structured rationale supervision and iterative refinement, the **model progressively enhances both its perceptual grounding and reasoning capabilities**.

## Method

### Negative Rationalisation:

- To induce **human strategy of reflective learning**, we introduce **negative rationales**.
- The core idea is that **correct prediction** means that the model has a **good understanding** of the image and the question.
- For samples where the model predicts the correct answer, we prompt it to **explain why each alternative option is incorrect**.
- Training with these counterfactual explanations **improves the model's ability to discriminate the correct answer from distractors**.

## STaR vs. STL

Here's the context that ties everything together: Thorny devils are lizards that live in the deserts of Australia. The is adapted to be camouflaged in a sandy desert. Figure: thorny devil.

(A) Figure A (B) Figure B

**Question:** Which animal is also adapted to be camouflaged in a sandy desert? Select the correct answer from the following choices:  
 (a) Figure A (b) Figure B

**CoT Prompting**  
 You are an image based question-answering expert. Given an image along with a multiple choice question, your task is to select the correct choice based on the image. Provide a detailed thought process to answer the question. Conclude your answer with 'Therefore, the correct answer is (correct choice).'  
**Question:** Which animal is also adapted to be camouflaged in a sandy desert?  
 Select the correct answer from the following choices:  
 (a) Figure A (b) Figure B  
**Response:**  
 The image shows a polar bear, a lizard, and a sandy desert. The polar bear is adapted to be camouflaged in a snowy environment, while the lizard is adapted to be camouflaged in a sandy desert. **Therefore, the correct answer is (b).**

**Positive Rationalization**  
 You are an image based question-answering expert. Given an image along with a multiple choice question, your task is to select the correct choice based on the image. Provide a detailed thought process to answer the question. Conclude your answer with 'Therefore, the correct answer is (correct choice).'  
**Question:** Which animal is also adapted to be camouflaged in a sandy desert?  
 Select the correct answer from the following choices:  
 (a) Figure A (CORRECT) (b) Figure B  
**Response:**  
 The image shows a polar bear, a lizard, and a sandy desert. **The polar bear is an animal adapted to be camouflaged in a sandy desert. Therefore, the correct answer is (a).**

**Structured Rationale Prompt (Positive)**  
 You are an image based question-answering expert. Given an image along with a multiple choice question, your task is to select the correct choice based on the image. Your response should strictly follow the format with three specific sections: CAPTION, REASONING and CONCLUSION. Response:  
 ##CAPTION:[Provide a detailed description of the image, particularly emphasizing the aspects related to the question.]  
 ##REASONING:[Provide a detailed thought process to answer the question.]  
 ##CONCLUSION:[Provide the correct choice based on the reasoning.]  
**Question:** Which animal is also adapted to be camouflaged in a sandy desert?  
 Select the correct answer from the following choices:  
 (a) Figure A (b) Figure B  
**Response:**  
 Caption: The image shows a lizard and a polar bear. The lizard is camouflaged in the sand, while the polar bear is not.  
**Reasoning:** The lizard is adapted to be camouflaged in a sandy desert, as it is able to blend in with the sandy environment. The polar bear, on the other hand, is not adapted to be camouflaged in a sandy desert, as it is not a natural habitat for it.  
**Conclusion:** (a)  
**Negative Rationalization**  
 You are an image based question-answering expert. Given an image along with a multiple choice question and an answer, your task is to explain why the answer is wrong. Your response should strictly follow the format with two specific sections: CAPTION and EXPLANATION. Response:  
 ##CAPTION:[Provide a detailed description of the image, particularly emphasizing the aspects related to the question.]  
 ##EXPLANATION:[Provide a detailed explanation for why the answer is wrong.]  
**Question:** Which animal is also adapted to be camouflaged in a sandy desert?  
 Select the correct answer from the following choices:  
 (a) Figure A (b) Figure B  
**Response:**  
 The correct choice is (a).  
 Explain why this answer is wrong: (b)  
**Response:**  
 Caption: The image shows a lizard, a polar bear, and a sandy desert. The lizard is camouflaged in the sandy desert, while the polar bear is not.  
**Explanation:** The correct answer is (a) because the lizard is adapted to be camouflaged in a sandy desert. It has a colouration that blends in with the sandy environment, making it difficult to spot. The polar bear, on the other hand, is not adapted to be camouflaged in a sandy desert. It is a large, white animal that stands out in the environment, making it easy to spot. Therefore, the answer (b) is incorrect.

Figure illustrating how negative rationalisation prevents shortcut learning.

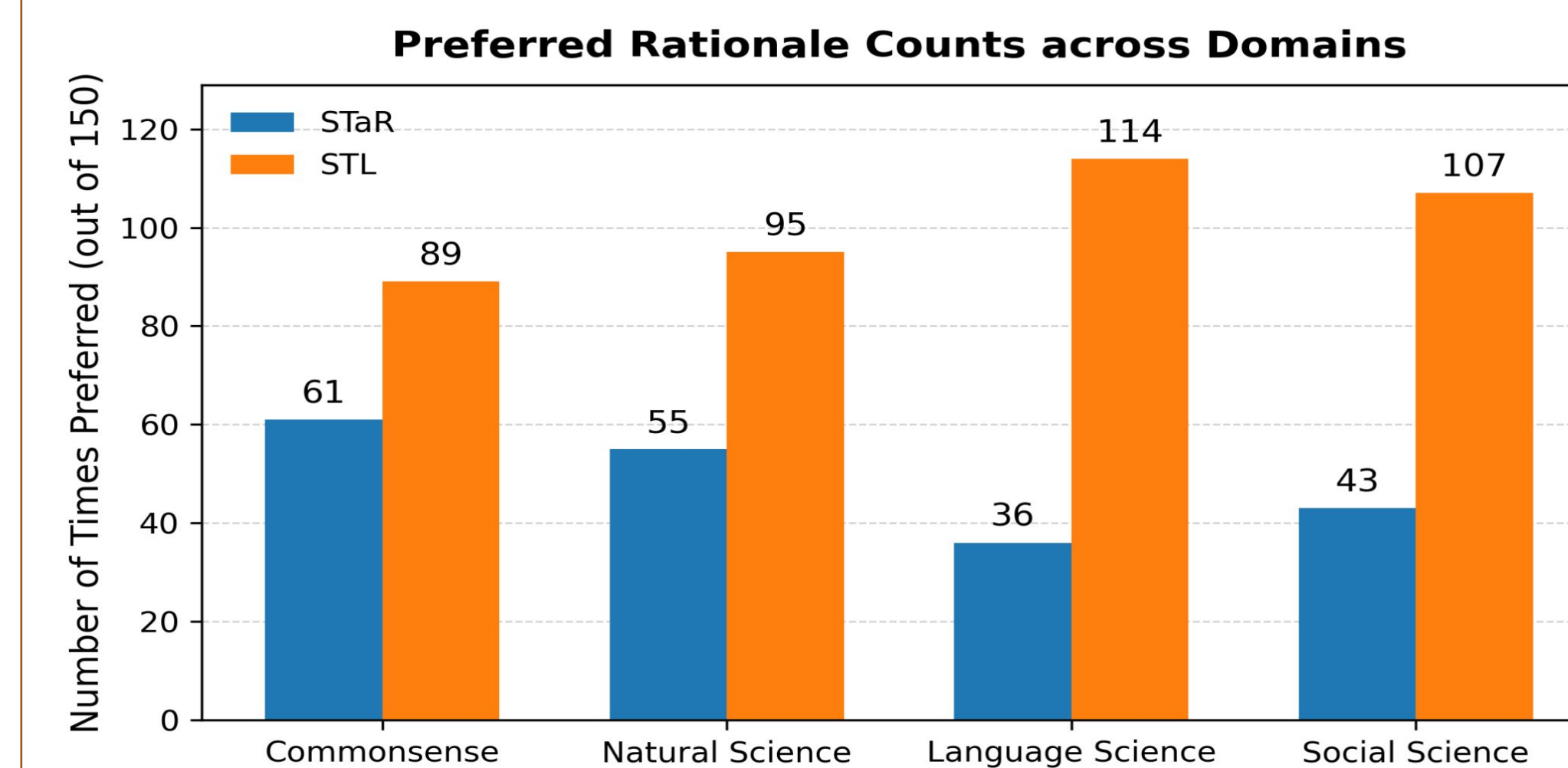
## Quantitative Results

Method	Commonsense	Language Science
<b>Zero-Shot Methods</b>		
Direct VQA	57.58	45.02
CoT	54.94	35.54
Structured Rationale Prompt	53.40	40.28
<b>Direct SFT</b>		
Direct SFT	60.22	46.92
<b>Self-Training Methods</b>		
STaR	64.98	48.82
R <sup>3</sup> V	62.64	45.97
Ours	<b>67.19</b>	<b>55.92</b>

Performance on M3CoT Domain Splits (LLaVA): STL outperforms baselines.

Method	Commonsense	Language Science
W/O (Cap+Neg)	59.56	46.44
W/O Neg	64.62	48.34
Ours (W/ (Cap+Neg))	<b>67.19</b>	<b>55.92</b>

Ablation Study. Neg: Negative Rationalization, Cap: Structured Rationale Prompt



Subjective Analysis: STL rationales are preferred over STaR.

## References

- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In Advances in Neural Information Processing Systems, 2022.
- Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. The North American Chapter of the Association for Computational Linguistics, 2025.