


Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

Current Training Process of Multimodal Chain-of-Thought Reward Models:


- Relies on **labor-intensive** annotations
- Provides only **single-step** rewards
- **Lacks** sufficient evaluation **dimensions**

→ As a result, the reward model cannot offer step-level, multi-dimensional evaluation


Query:
What is the color of birds in sky?



There are many birds in the picture.
A bird is *on the top-left of* the image.
And several *white* doves on a *pine*.
So the answer is *white*.



Existing Reward Model:




According to the given picture and answer,
the bird in the sky is *white*, so the answer is *correct*.

One-step Single-dimensional

SVIP-Reward:

Multi-step Multi-dimensional

	Relevance	Logic	Attribute
Step1:	✓	✓	✓
Step2:	✗	✓	✓
Step3:	✓	✓	✗
Step4:	✓	✗	✓

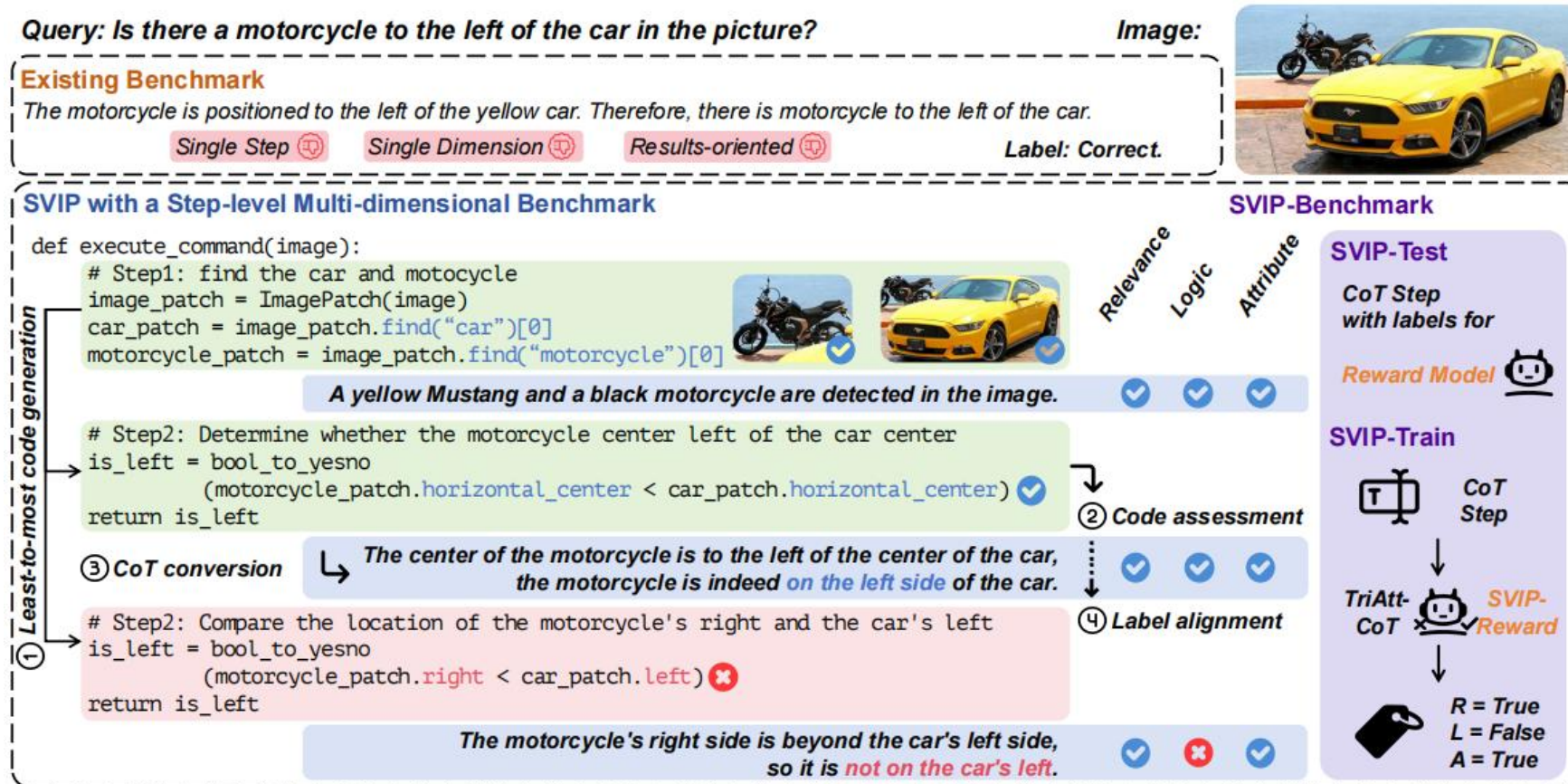


Although the answer is correct, after analysis,
the second step of is *irrelevant* to the problem,
and there is an *attribute* error in the third step,
so the answer is *logically unreliable*.

Existing reward models: Evaluation is **single and holistic**.
While SVIP, evaluates **each CoT step** across **three dimensions**.

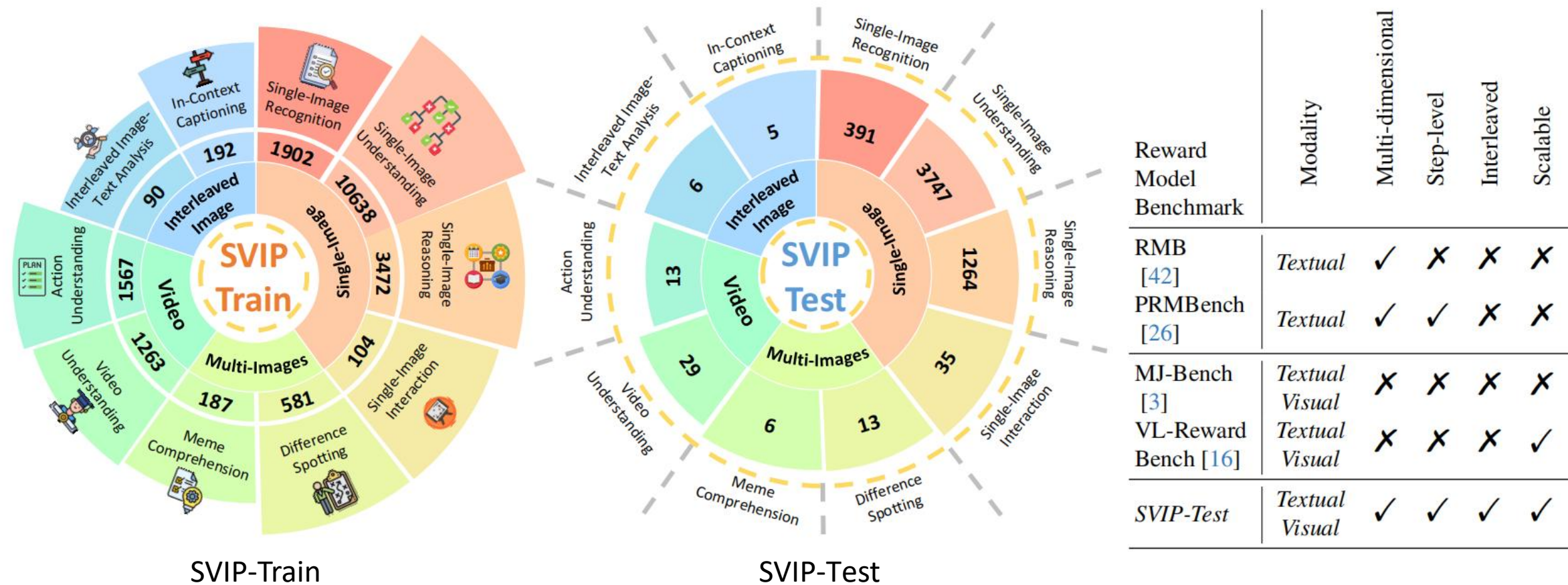
Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

- We propose SVIP: an automatic approach to training stepwise, multi-dimensional CoT reward models.
- SVIP Process: 1) Generates code to solve visual tasks. 2) Transforms code block analysis into CoT step evaluations as training samples. 3) Trains the SVIP-Reward model with a multi-head attention mechanism, TriAtt-CoT.



Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

- SVIP introduces two datasets for training and evaluation: SVIP-Train and SVIP-Test.
- Compared with existing benchmarks, it provides step-level data with scores for each step.




Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

- SVIP defines three types of CoT step labels and performs corresponding evaluations through program analysis.

Relevance

What is the gender ratio of the people in the image?



Code Block

```
...
return gender_ratio =
    len(male) / len(female)
```

Compilation results

ZeroDivisionError:
division by zero!

Convert to CoT Step

The number of female is zero, which prevents the division operation and results in an error. ❌

Code Block

```
...
if len(female) == 0:
    return "No female at all."
```

Compilation results


Successful Compilation!

Convert to CoT Step

There are no females in the image. The gender ratio consists entirely of males. ✅

Logic

Is the charger in the picture yellow and round?



Code Block

```
...
is_yellow_and_round
= is_yellow or is_round
...
```

ProgTest results*

```
assert
(is_yellow and is_round)
== is_yellow_and_round
```

Convert to CoT Step

The charger in the picture is yellow but not round, so the charger is yellow and round. ❌

Code Block

```
...
is_yellow_and_round
= is_yellow and is_round
...
```

ProgTest results*

```
assert
(is_yellow and is_round)
== is_yellow_and_round
```


Convert to CoT Step

The charger in the picture is yellow but not round, so the charger is not yellow and round. ✅

*ProgTest only tests the necessary logic according to the requirements of the problem

Attribute

Who is the man in yellow?



Code Block

```
...
return man_in_yellow.
    simple_query("Who is he?")
```

De-fine results

He is Tadej Pogacar, 2024 Tour de France Champion

Convert to CoT Step

Therefore, the man in yellow is Tadej Pogacar ❌

Code Block

```
...
return man_in_yellow.
    simple_query("Who is he?")
```

De-fine results

He is Jonas Vingegaard, 2023 Tour de France Champion

Convert to CoT Step

Therefore, the man in yellow is Jonas Vingegaard ✅

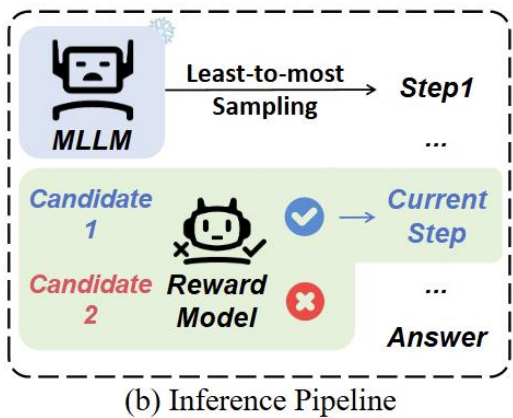
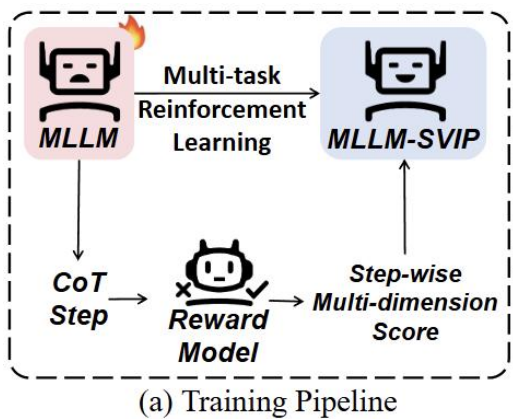
Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

- The results on SVIP-Test show that, among existing reward models, SVIP-Reward achieves the highest performance.
- Notably, with Qwen2-VL-7B as the base model, SVIP outperforms existing tuning methods by 7.3%.

Reward Model	MJ-Bench	VLRewardBench	SVIP-Test				
			Overall	Step			
				Relevance	Logic	Attribute	Avg
GPT-4o	65.8	65.8	58.4	82.2	54.7	56.8	64.6
IXC-2.5-Reward-7B	69.2	65.8	63.7	89.4	60.1	53.6	67.7
Qwen2-VL-7B (<i>Zero-shot</i>)	72.3	28.3	55.1	73.8	50.2	47.3	57.1
Qwen2-VL-7B (<i>Tuning</i>)	73.1	55.4	63.9	81.5	55.9	52.7	63.4
Qwen2-VL-7B (<i>SVIP-Reward</i>)	73.7	68.0	67.3	92.3	61.0	58.9	70.7
InternVL-2.5-2B (<i>Zero-shot</i>)	48.0	35.1	46.2	75.6	52.1	51.1	59.6
InternVL-2.5-2B (<i>Tuning</i>)	50.4	49.6	54.5	78.6	53.5	53.5	61.9
InternVL-2.5-2B (<i>SVIP-Reward</i>)	51.6	56.3	60.8	87.0	56.4	56.2	66.5

Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

- SVIP-Reward efficiently improves performance in both training and testing phases.



MLLM	Reward Model	MME _{sum}	MMMU _{val}	MMMU-Pro _{overall}	MathVista _{mini}	MMT _{val}	POPE _{avg}
DeepSeek-VL-7B	-	1835	36.9	18.2	36.7	53.6	87.9
Qwen2-VL-7B	-	2374	58.5	30.9	59.3	64.6	89.3
Qwen2-VL-7B	Qwen2-VL-7B (<i>Zero-shot</i>)	2390	59.9	31.3	59.4	64.8	89.5
Qwen2-VL-7B	Qwen2-VL-7B (<i>Tuning</i>)	2387	60.7	32.4	59.6	67.4	89.6
Qwen2-VL-7B	Qwen2-VL-7B (<i>SVIP-Reward</i>)	2466	61.8	33.0	60.6	70.9	90.1
Openflamingo-3B	-	912	37.2	22.0	33.2	51.4	73.4
InternVL-2.5-2B	-	2144	59.4	24.3	51.7	55.8	90.6
InternVL-2.5-2B	InternVL-2.5-2B (<i>Zero-shot</i>)	2202	60.8	24.7	52.5	56.2	90.6
InternVL-2.5-2B	InternVL-2.5-2B (<i>Tuning</i>)	2261	61.2	26.1	53.6	58.6	90.8
InternVL-2.5-2B	InternVL-2.5-2B (<i>SVIP-Reward</i>)	2332	62.0	27.4	54.9	63.5	90.8

MLLM	Reward Model	MME _{sum}	MMMU _{val}	MMMU-Pro _{overall}	MathVista _{mini}	MMT _{val}	POPE _{avg}
GPT-4o	-	2328	69.1	51.9	63.8	65.4	86.9
Deepseek-VL-7B	-	1847	36.6	18.1	36.1	53.2	88.1
Openflamingo-3B	-	668	21.8	11.6	29.5	47.4	58.6
Qwen2-VL-7B	-	2327	54.1	30.5	58.2	64.0	88.1
Qwen2-VL-7B	Qwen2-VL-7B (<i>Zero-shot</i>)	2301	59.1	30.7	59.0	64.8	89.0
Qwen2-VL-7B	Qwen2-VL-7B (<i>Tuning</i>)	2383	60.4	31.4	59.2	67.7	89.5
Qwen2-VL-7B	Qwen2-VL-7B (<i>SVIP-Reward</i>)	2472	61.3	32.3	61.6	70.5	90.2
InternVL-2.5-2B	-	2138	40.9	23.7	51.3	54.5	90.6
InternVL-2.5-2B	InternVL-2.5-2B (<i>Zero-shot</i>)	2185	41.2	24.4	53.1	56.1	90.4
InternVL-2.5-2B	InternVL-2.5-2B (<i>Tuning</i>)	2246	43.0	25.9	53.7	58.6	90.8
InternVL-2.5-2B	InternVL-2.5-2B (<i>SVIP-Reward</i>)	2319	43.7	26.8	55.4	63.3	90.9

Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

- An Illustrative Example of SVIP

1) **Query:** What is the name of the landmark in the picture?

Golden answer: Trondheimsfjorden

CoT step: *We attempt to analyze information about a landmark in an image, but I don't know what island it is.*

Relevance: True; Logic: True; Attribute: False.



Figure 7. Attribute errors due to external knowledge errors.

In this scenario, if GPT4V is utilized as the visual module for visual programming, it would accurately identify the landmark in the image as Munkholmen, a small islet located in the Trondheim Fjord in Norway. However, less capable models, such as blip2, would fail to provide the correct answer, subsequently leading to their detection by the SVIP system, which would mark the Attribute as false.