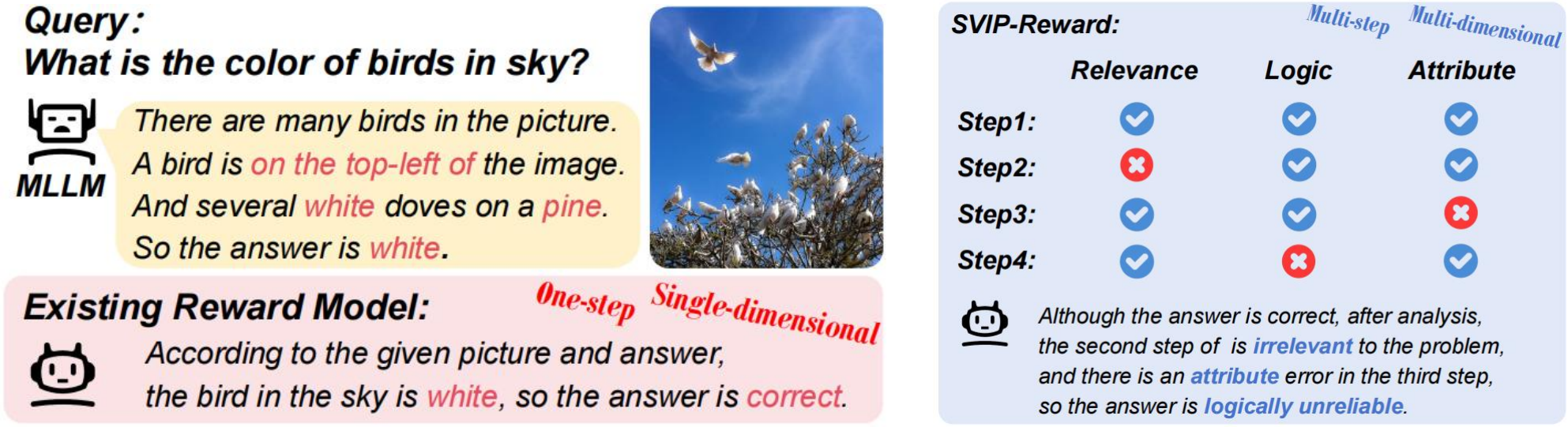
Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

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

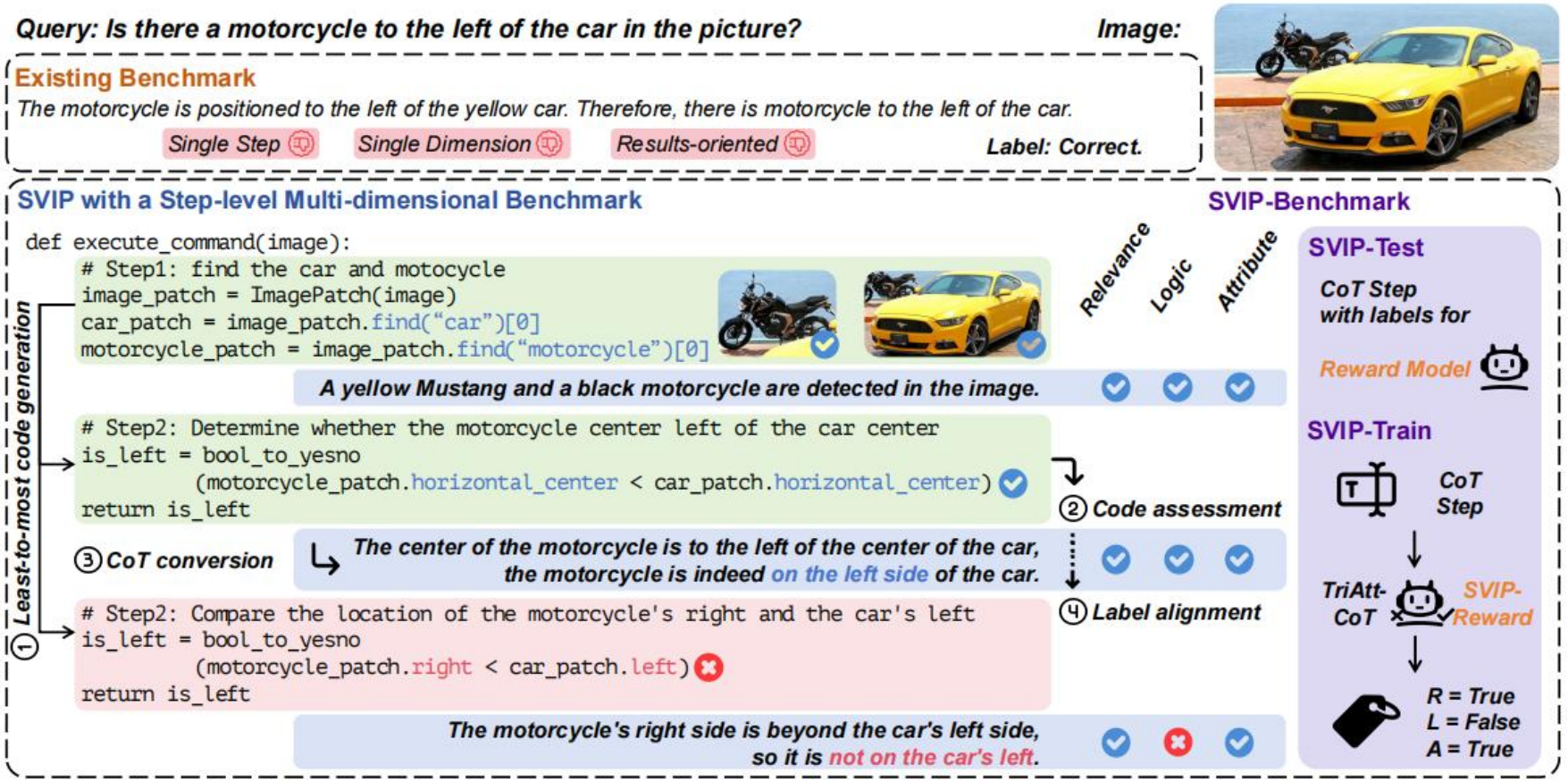


Existing reward models: Evaluation is **single and holistic.**

While SVIP, evaluates **each CoT step** across **three dimensions.**

Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

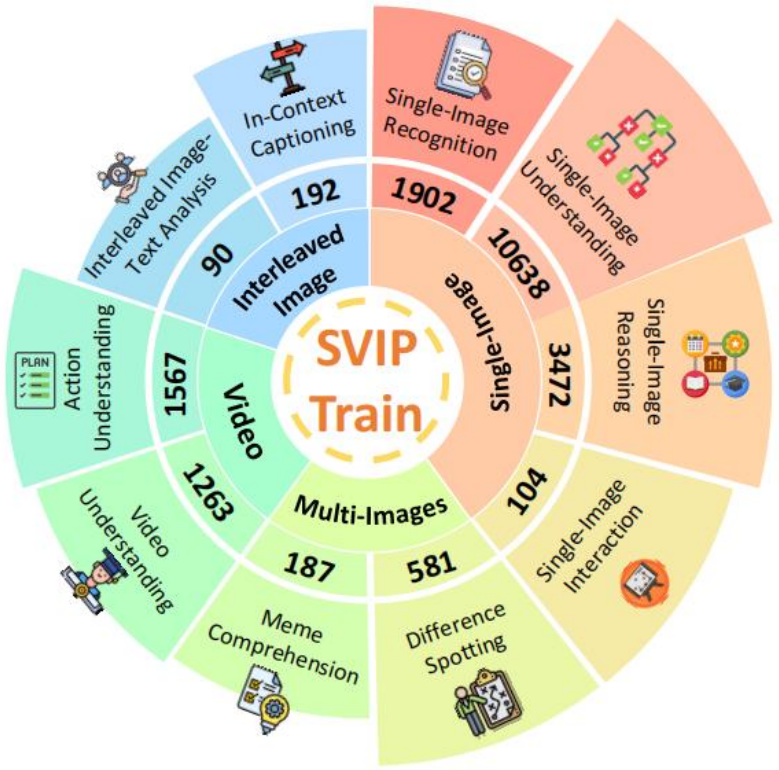
- We propose SVIP: an automatic approach to training stepwise, multi-dimensional CoT reward models.
- SVIP Process: 1) Generates code to solve visual tasks. 2) Transforms code block analysis into CoT step evaluations as training samples. 3) Trains the SVIP-Reward model with a multi-head attention mechanism,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



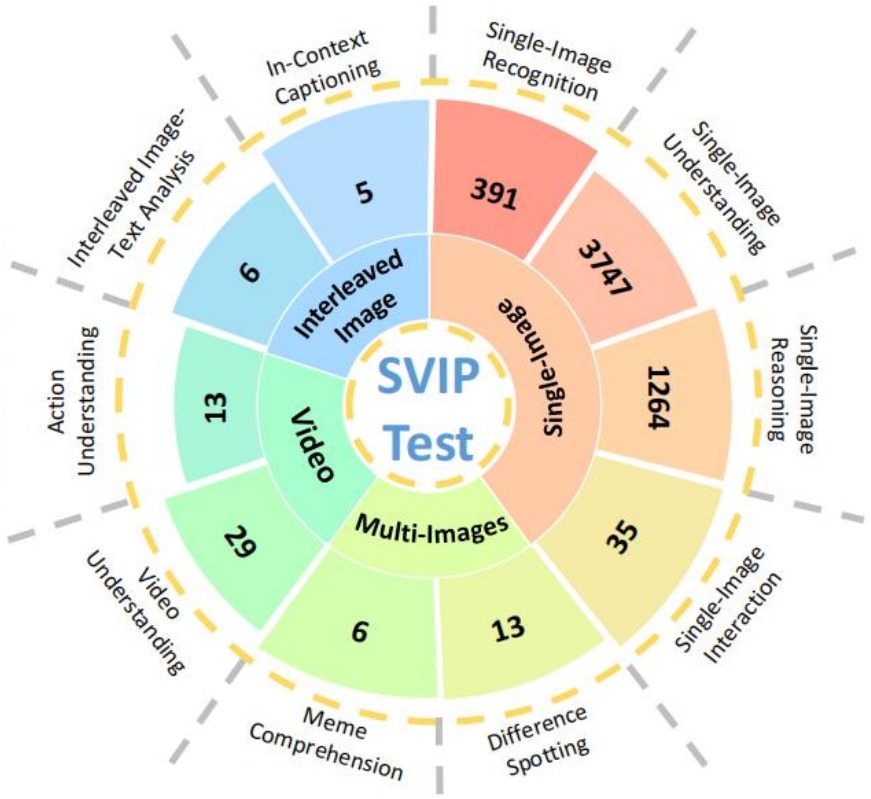
Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program. **ICCV2025.**

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.



SVIP-Train



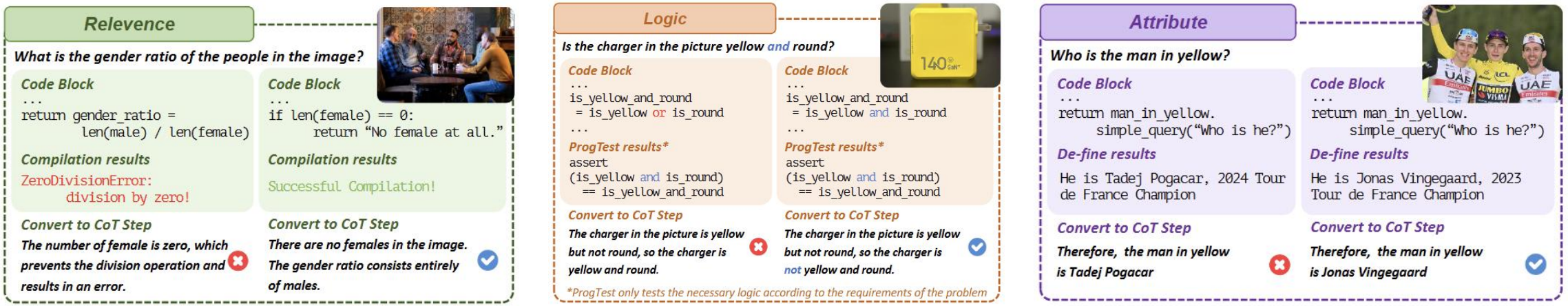
SVIP-Test

Reward Model Benchmark	Modality	Multi-dimensional	Step-level	Interleaved	Scalable
RMB [42]	*Textual*	✓	✗	✗	✗
PRMBench [26]	*Textual*	✓	✓	✗	✗
MJ-Bench [3]	*Textual* *Visual*	✗	✗	✗	✗
VL-Reward Bench [16]	*Textual* *Visual*	✗	✗	✗	✓
SVIP-Test	*Textual* *Visual*	✓	✓	✓	✓

Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program. **ICCV2025.**

Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

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



Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program. **ICCV2025.**

Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

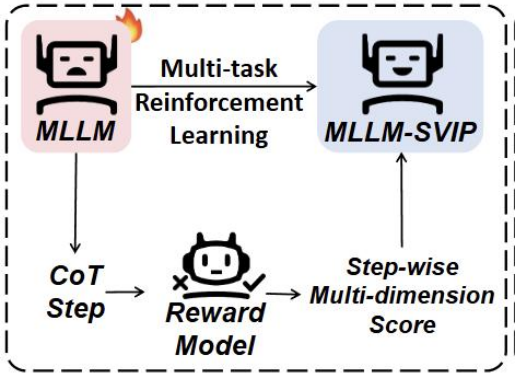
- The results on SVIP-Test show that, among existing reward models, SVIP-Reward achieves the highest performance.
- Notably, with Qwen2-VL-7B as the base model, SVIP outperforms existing tuning methods by 7.3%.

Reward Model	MJ-Bench	VLRewardBench	*SVIP-Test*				
			Overall	Step			
			Correctness	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 (*Zero-shot*)	72.3	28.3	55.1	73.8	50.2	47.3	57.1
Qwen2-VL-7B (*Tuning*)	73.1	55.4	63.9	81.5	55.9	52.7	63.4
Qwen2-VL-7B (`SVIP-Reward`)	**73.7**	**68.0**	**67.3**	**92.3**	**61.0**	**58.9**	**70.7**
InternVL-2.5-2B (*Zero-shot*)	48.0	35.1	46.2	75.6	52.1	51.1	59.6
InternVL-2.5-2B (*Tuning*)	50.4	49.6	54.5	78.6	53.5	53.5	61.9
InternVL-2.5-2B (`SVIP-Reward`)	**51.6**	**56.3**	**60.8**	**87.0**	**56.4**	**56.2**	**66.5**

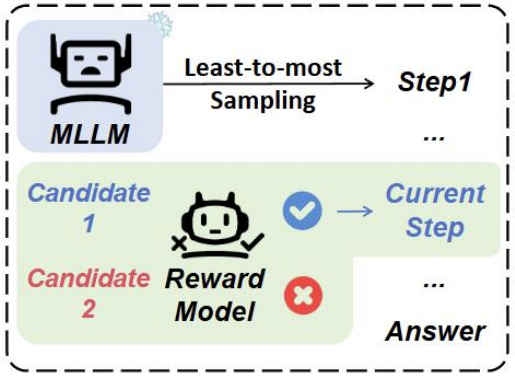
Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program. **ICCV2025.**

Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program.

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



(a) Training Pipeline



(b) Inference Pipeline

MLLM	Reward Model	MME_{sum}	$MMMU_{val}$	$MMMU\text{-}Pro_{overall}$	$MathVista_{mini}$	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 (*Zero-shot*)	2390	59.9	31.3	59.4	64.8	89.5
Qwen2-VL-7B	Qwen2-VL-7B (*Tuning*)	2387	60.7	32.4	59.6	67.4	89.6
Qwen2-VL-7B	Qwen2-VL-7B (`SVIP-Reward`)	**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 (*Zero-shot*)	2202	60.8	24.7	52.5	56.2	90.6
InternVL-2.5-2B	InternVL-2.5-2B (*Tuning*)	2261	61.2	26.1	53.6	58.6	90.8
InternVL-2.5-2B	InternVL-2.5-2B (`SVIP-Reward`)	**2332**	**62.0**	**27.4**	**54.9**	**63.5**	**90.8**

MLLM	Reward Model	MME_{sum}	$MMMU_{val}$	$MMMU\text{-}Pro_{overall}$	$MathVista_{mini}$	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 (*Zero-shot*)	2301	59.1	30.7	59.0	64.8	89.0
Qwen2-VL-7B	Qwen2-VL-7B (*Tuning*)	2383	60.4	31.4	59.2	67.7	89.5
Qwen2-VL-7B	Qwen2-VL-7B (`SVIP-Reward`)	**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 (*Zero-shot*)	2185	41.2	24.4	53.1	56.1	90.4
InternVL-2.5-2B	InternVL-2.5-2B (*Tuning*)	2246	43.0	25.9	53.7	58.6	90.8
InternVL-2.5-2B	InternVL-2.5-2B (`SVIP-Reward`)	**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.

Benchmarking Multimodal CoT Reward Model Stepwise by Visual Program. **ICCV2025.**