

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

See to Believe: Aligning Multimodal Chain-of-Thought Reasoning with Visual Evidence

Visual Reasoning



Vision & Language Prompt

User: Who is most likely to be the ***player*** in this picture?

Reasoning-Chain Guided

Seg-Zero

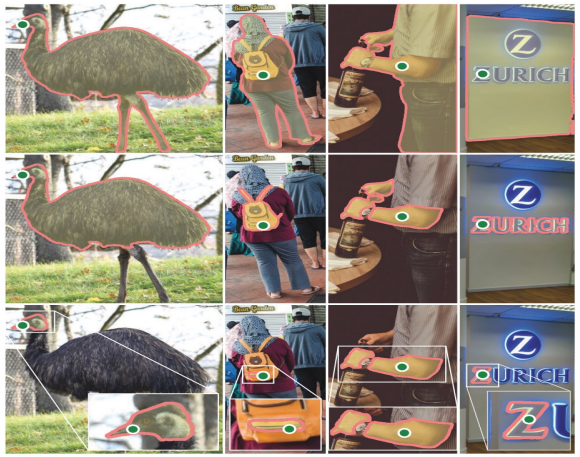
Assistant: The player is most likely the one ***wearing the baseball uniform,*** including the helmet, belt, and baseball glove. The person kneeling down is likely the coach or another adult, as ***they are not wearing a uniform***. So, the answer is [SEG-REF] In this image, **[Captioning Part]**



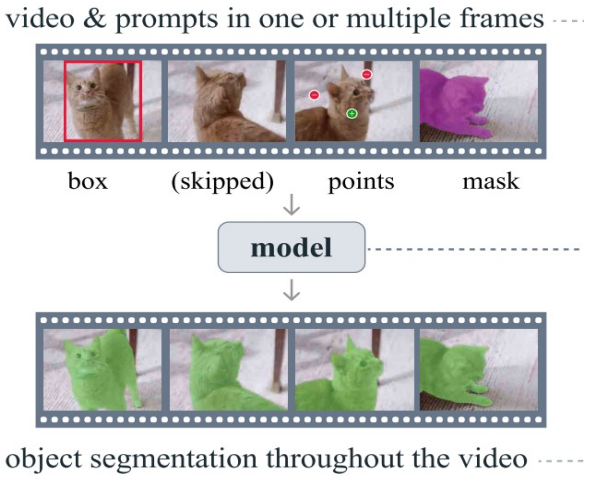
Visual reasoning: bridging high-level textual semantics to pixel-level visual contents.

Visual Reasoning

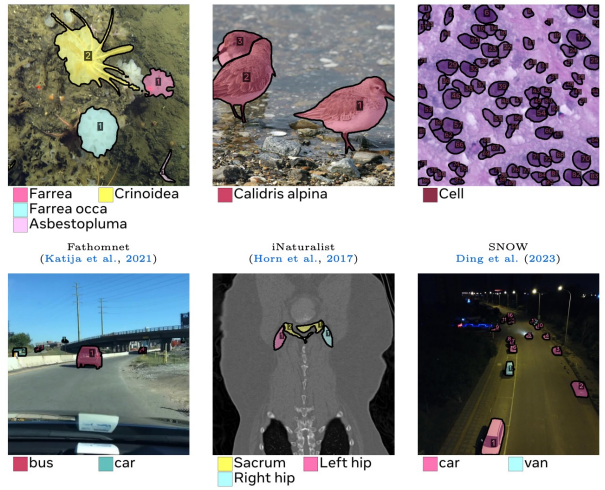
The development of SAM-series visual foundation models



SAM1: box, point…



SAM2: video



SAM3: concept

Visual Reasoning



the object used for stirring milk or coffee

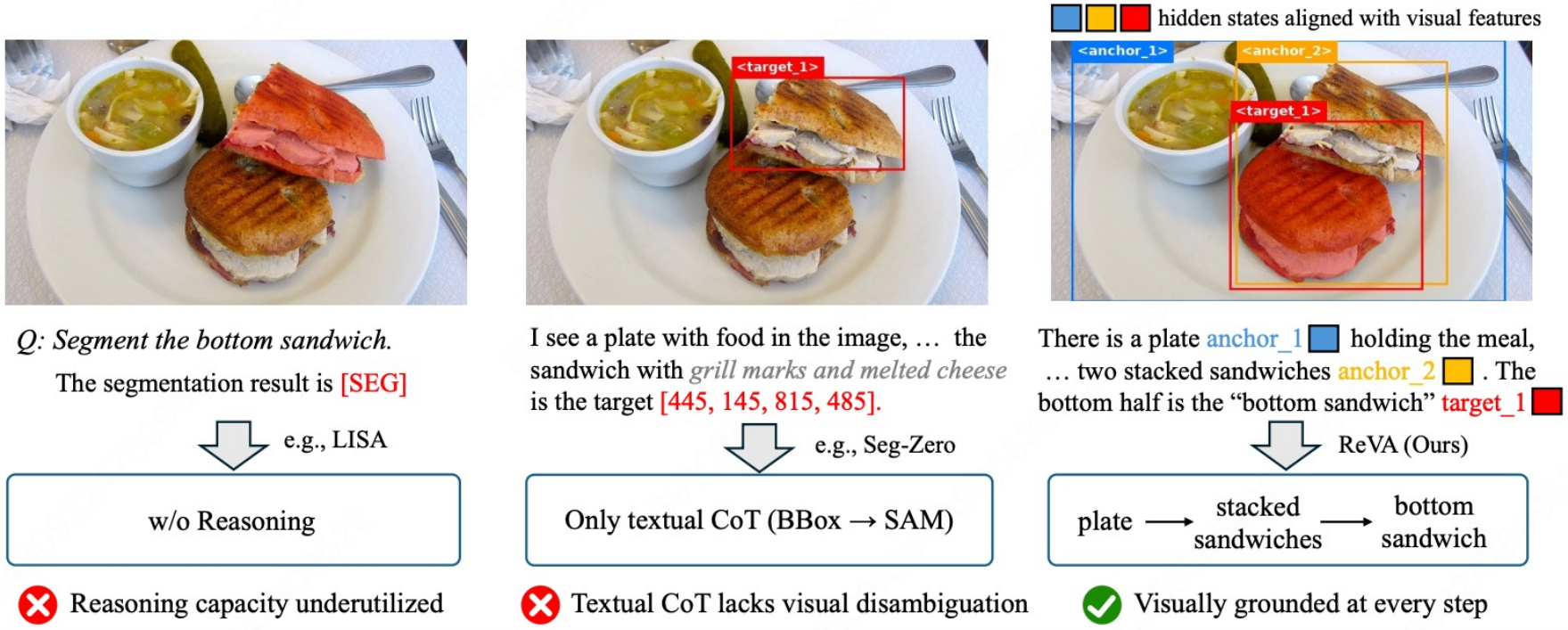
Model		Training	RefCOCO			RefCOCO+			RefCOCOg		
Name	Version	RES	val	testA	testB	val	testA	testB	val (U)	test (U)	val (G)
LISA	LLaVA 7B	✓	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	–
GSVA	13B	✓	79.2	81.7	77.1	70.3	73.8	63.6	75.7	77.0	–
GLaMM	Vicuna 7B	✓	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	–
SAM4MLLM	LLaVA1.6 7B	✓	79.6	82.8	76.1	73.5	77.8	65.8	74.5	75.6	–
SAM4MLLM	LLaVA1.6 8B	✓	79.8	82.7	74.7	74.6	80.0	67.2	75.5	76.4	–
GLEE	Plus	✓	79.5	–	–	68.3	–	–	70.6	–	–
GLEE	Pro	✓	80.0	–	–	69.6	–	–	72.9	–	–
DETRIS	DETRIS-L	✓	81.0	81.9	79.0	75.2	78.6	70.2	74.6	75.3	–
UniLSeg	UniLSeg-20	✓	80.5	81.8	78.4	72.7	77.0	67.0	78.4	79.5	–
UniLSeg	UniLSeg-100	✓	81.7	83.2	79.9	73.2	78.3	68.2	79.3	80.5	–
PSALM	Phi1.5 1.3B	✓	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4	–
SAM 3 Agent	Qwen2.5-VL 7B	×	59.4	64.3	55.0	51.4	57.0	44.9	57.2	58.8	59.7
SAM 3 Agent	Qwen2.5-VL 72B	×	71.6	74.9	66.0	64.9	70.8	57.9	68.8	70.2	70.4
SAM 3 Agent	Llama4 Maverick	×	71.7	76.5	66.5	65.1	71.2	57.6	67.7	68.7	68.4
SAM 3 Agent	Qwen3-VL 8B Thinking	×	68.6	72.3	63.9	59.3	64.5	55.6	66.4	66.9	67.8
SAM 3 Agent	Qwen3-VL 235 Thinking	×	73.9	77.0	69.7	66.9	70.9	62.2	72.1	72.9	72.5
SAM 3 Agent	Gemini2.5 Pro	×	75.5	77.6	71.0	67.3	71.1	63.4	73.4	74.0	74.6



While the latest SAM models excel at pixel-level **perception**, they lack precise text-to-pixel **reasoning**.

Reliable Visual CoT Reasoning

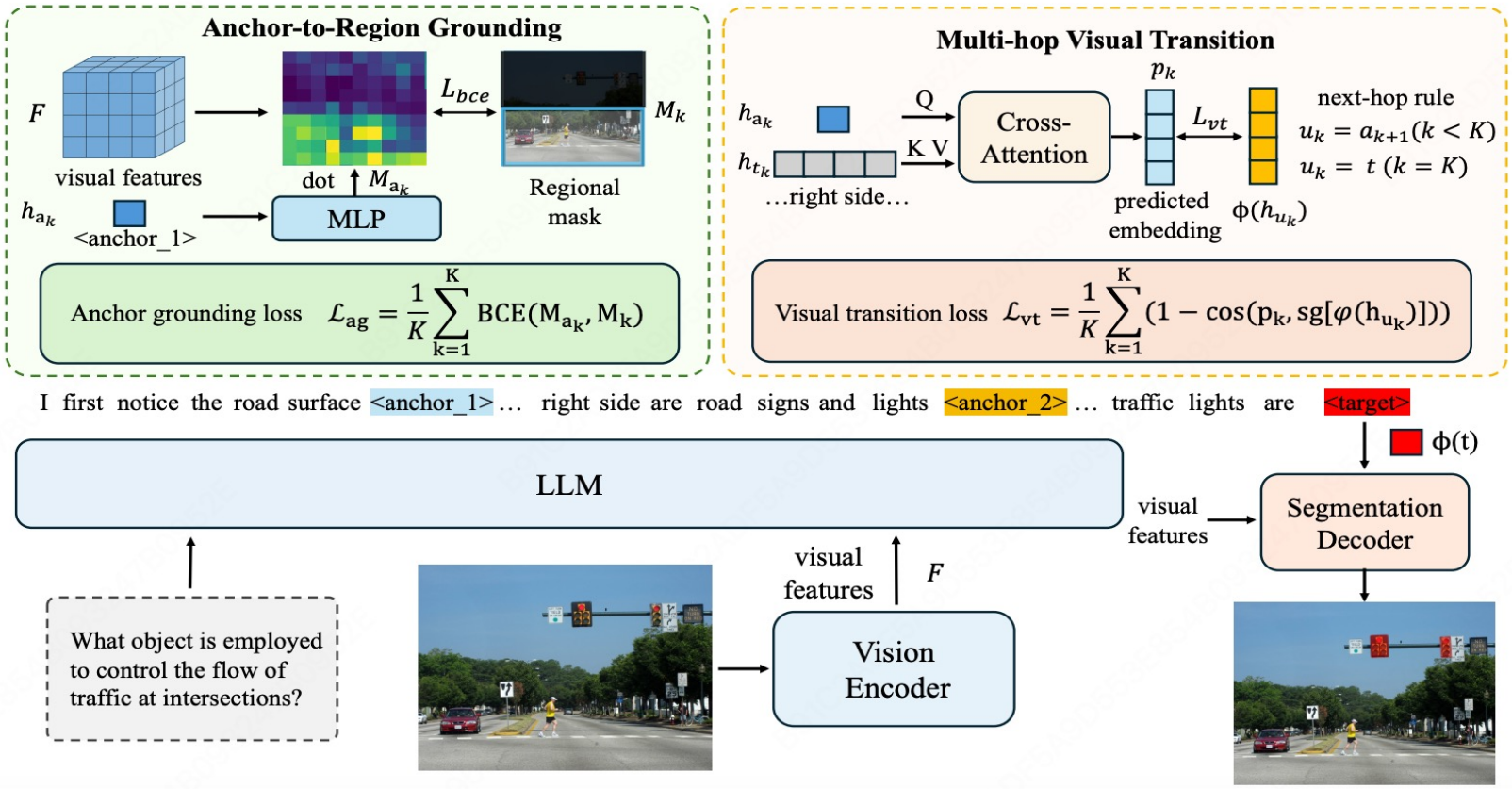
See to Believe: Segmentation by Reasoning with Visual Evidence, Neurips 2026 submission



w/o CoT: Directly generating masks using [SEG] bypasses intermediate reasoning, wasting the model's reasoning capabilities.

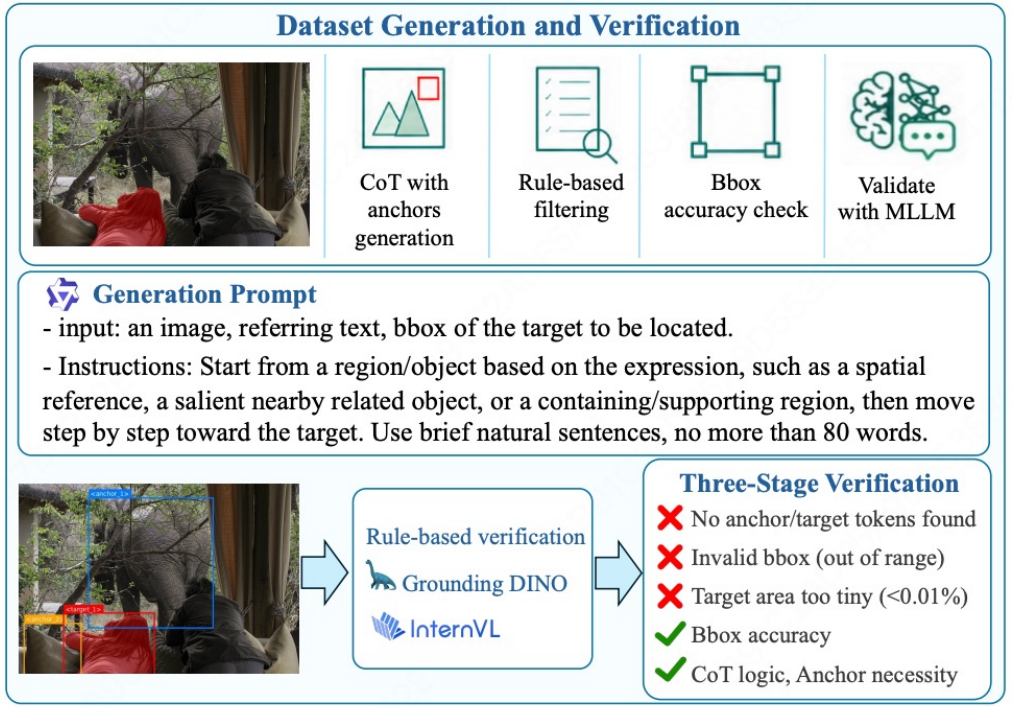
w textual CoT: Leading to ambiguous descriptions when multiple similar objects are present. Since intermediate steps cannot be **visually verified**, errors propagate to the final prediction.

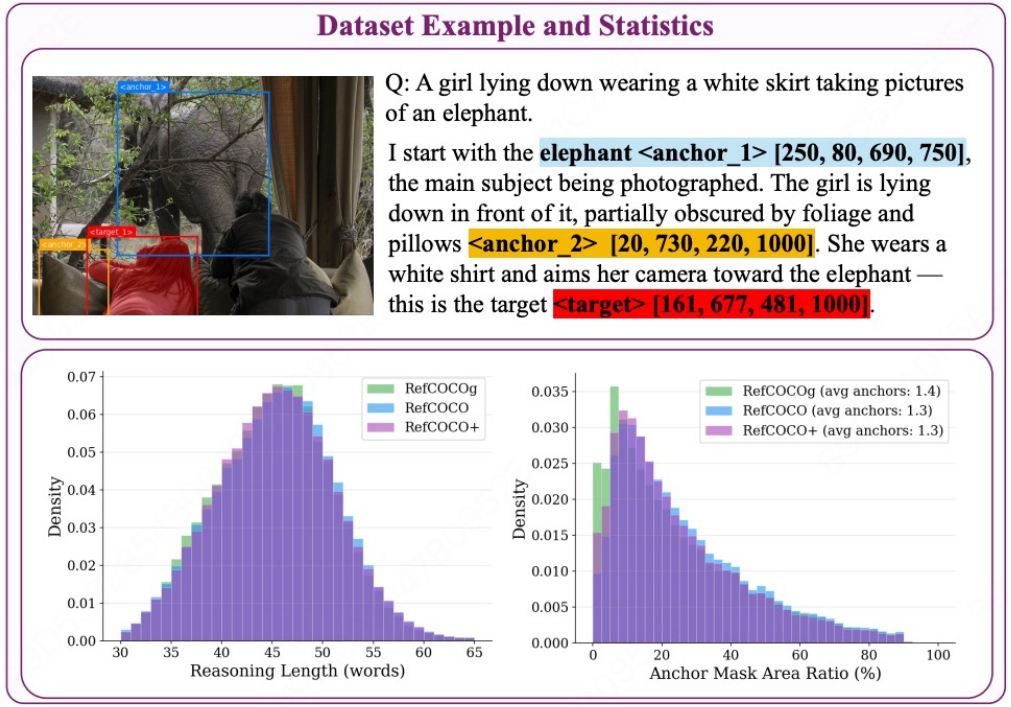
Reliable Visual CoT Reasoning



Insert <anchor> at each CoT step, projecting hidden states of each <anchor> to align with image regions, **anchoring each step for visual verification**.

Reliable Visual CoT Reasoning

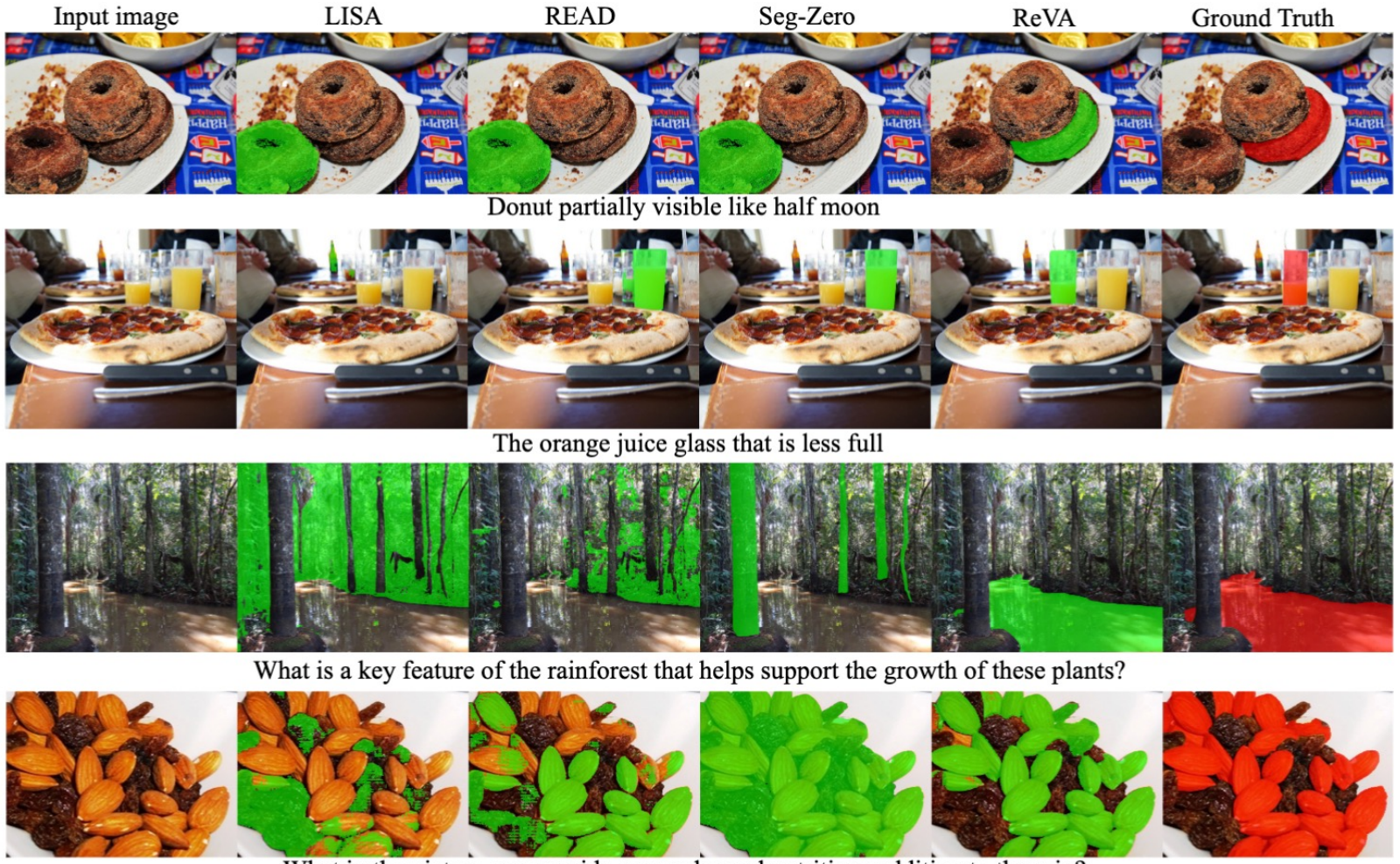




Data generation: Leveraging RefCOCO series, Qwen3-VL-235B generates reasoning trajectories with <anchor> region coordinates.

Three-stage filtering: Rule checks→ Grounding DINO validates anchor bbox accuracy → InternVL3.5 judges reasoning logic, yielding 245K high-quality samples.

Reliable Visual CoT Reasoning



Visualization