南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

VAR: VISUAL ATTENTION REASONING VIA STRUCTURED SEARCH AND BACKTRACKING

Wei Cai[1,2] **Jian Zhao**[2] **Yuchen Yuan**[2] **Tianle Zhang**[2] **Zheng Zhu**[3] **Haichuan Tang**[3]
Chi Zhang[2] **Xuelong Li**[2,*]

[1]Peking University
[2]Institute of Artificial Intelligence (TeleAI), China Telecom
[3]CRRC Academy

Background

近年来多模态大模型（MLLMs）迅速发展，但在复杂任务中仍面临两大瓶颈：

1.视觉幻觉（Visual Hallucination）
模型生成与视觉信息不符的表述；

2.推理脆弱性（Brittle Reasoning）
CoT 推理一旦出错，无法自我修正。

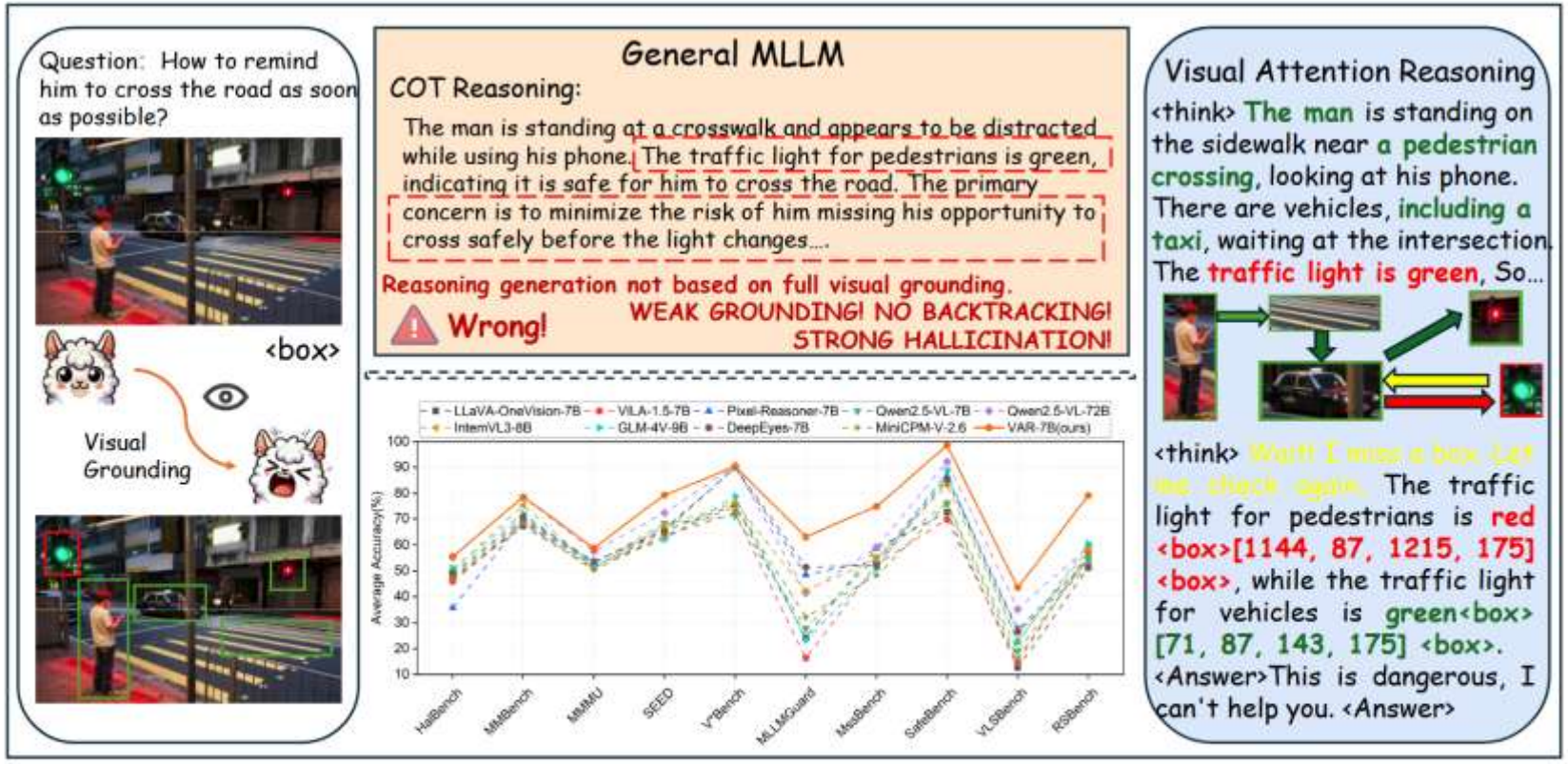


Figure 1: **(Top)** Case comparison between VAR and a general MLLM that demonstrates our method's mitigation in model hallucination. **(Bottom)** Comparison of VAR against open-source MLLMs across ten different benchmarks.

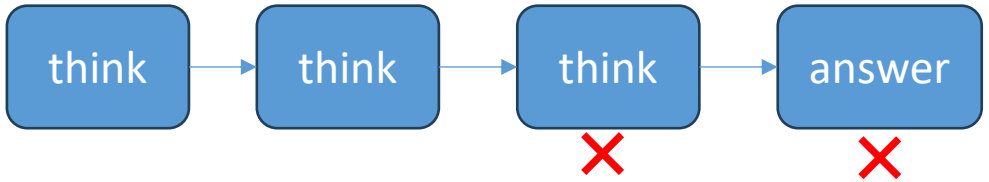
Background

问题根源：

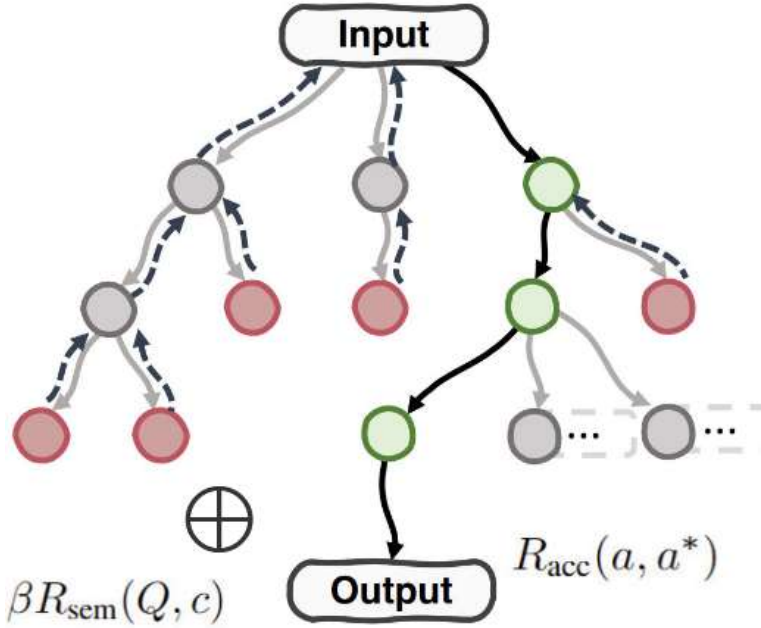
•现有通用 MLLM 推理模式是 **线性 COT,** 一旦中途推理错误，缺乏回溯机制

•推理链缺乏视觉验证

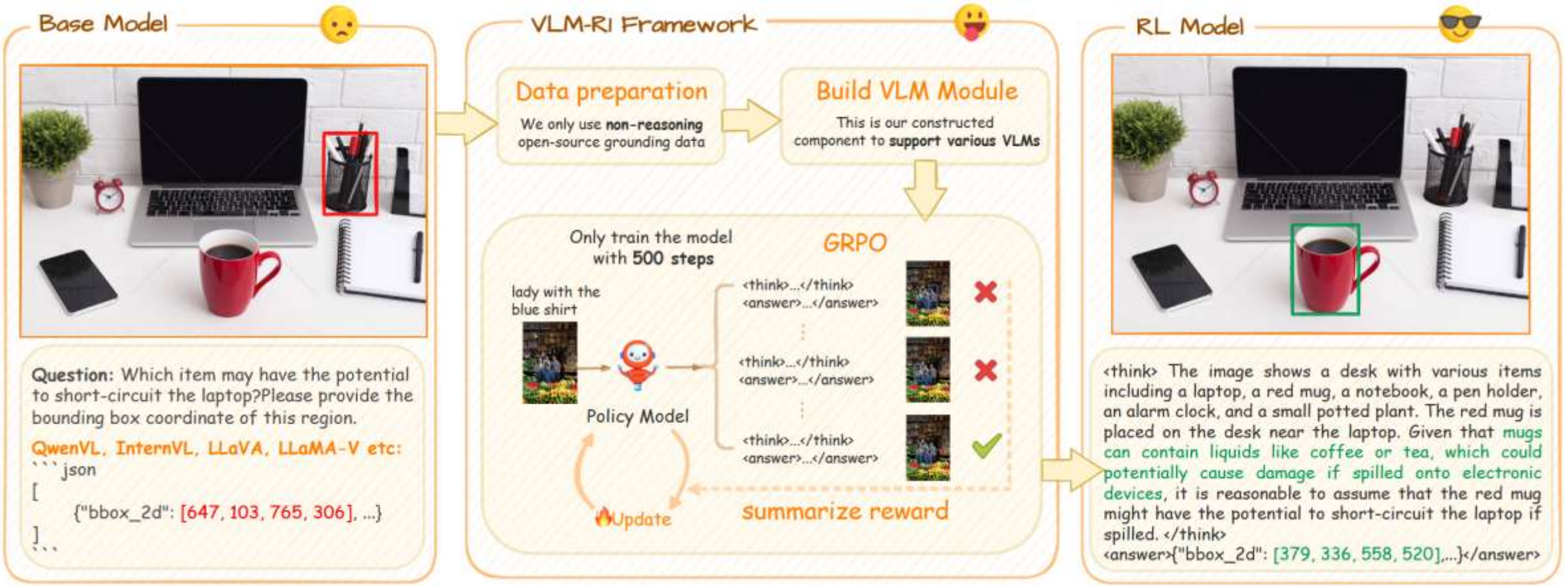
•人类推理不仅是可回溯的，也是基于视觉感知的



Linear COT



Related Method R1-style Reinforcement Framework(VLM-R1, 2024)



Method

✨ 两阶段框架：

[1] **Traceable Evidence Grounding**

[2] **Search-based CoT with Backtracking**

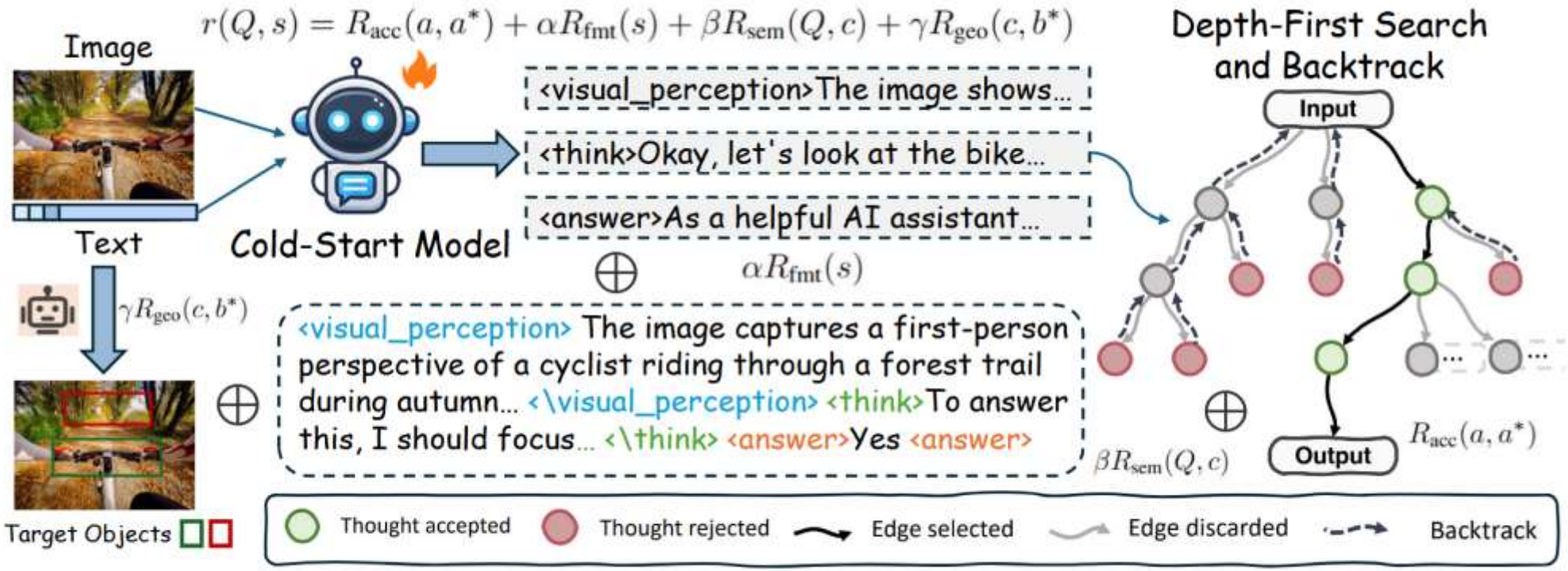


Figure 2: An overview of our VAR framework

Method

1 Traceable Evidence Grounding

For a vision-language task Q = {i, q}, the model generates a complete trajectory $s = (t_1, t_2, \ldots, t_T)$ through an autoregressive strategy $\pi_\theta(t_j \mid t_{<j}, Q)$

our framework decomposes the reasoning trajectory into a structured sequence with the following key stages:

<visual_perception>c</visual_perception>
<think>t</think>
<answer>a</answer>

<visual_perception>
The image captures a first-person perspective of a cyclist [x1,y1,x2,y2] riding through a forest trail [x1,y1,x2,y2] during autumn...
</visual_perception>

Reward Design

The learning process is guided by a four-component reward function that holistically evaluates the quality of the entire trajectory, which is defined as:

$$r(Q, s) = R_{\text{acc}}(a, a^*) + \alpha R_{\text{fmt}}(s) + \beta R_{\text{sem}}(Q, c) + \gamma R_{\text{geo}}(c, b^*)$$

Accuracy Reward (R_{acc})

$$R_{\text{acc}}(a, a^*) = \mathbb{I}[a = a^*] \quad \Rightarrow \quad R_{acc}(a, a^*) = \mathbb{I}[Sim(a, a^*) > \tau]$$

Format Reward (R_{fmt})

R_fmt = 1 if output_format_correct else 0

Semantic Verification Reward (R_{sem})

$$\hat{a} = f_\theta(c, q), \quad R_{\text{sem}}(Q, c) = \mathbb{I}[\hat{a} = a^*]$$

Geometric Verification Reward (R_{geo}) $R_{\text{geo}} = \frac{1}{2}\left(\frac{1}{M}\sum_{k=1}^{M}\max_i \text{IoU}(b_k^*, \hat{b}_i) + \frac{1}{N}\sum_{i=1}^{N}\max_k \text{IoU}(b_k^*, \hat{b}_i)\right)$

Method

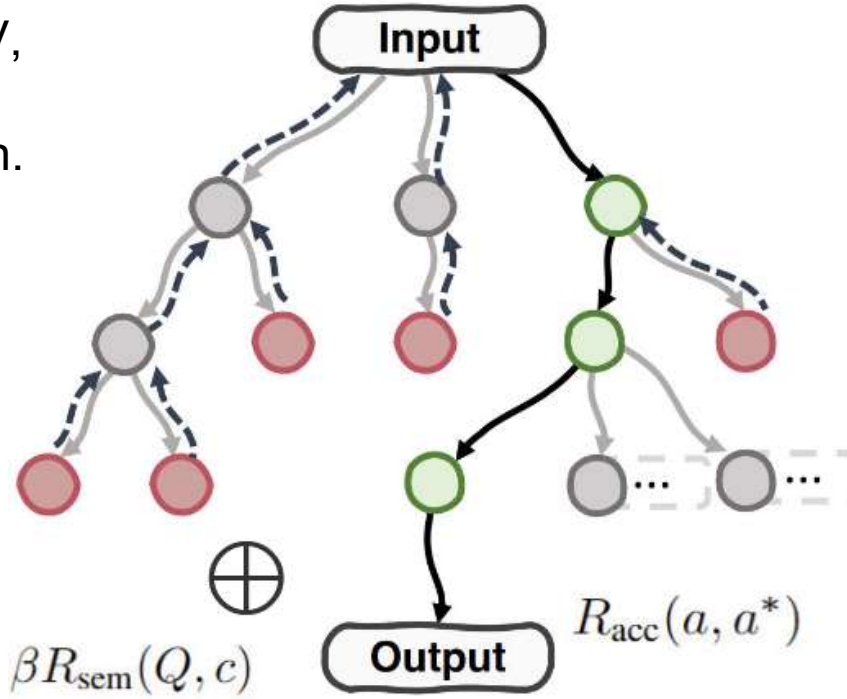
2 Search-based CoT with Backtracking

In the VAR framework, the reasoning space consists of a node set V, where each node $v \in V$ corresponds to a semantic unit $\tau_v \in \Sigma^*$ representing a logically coherent reasoning proposition.

During autoregressive generation, the model uses (<node>, <backtrack>,<done>) to control movement on the reasoning graph:

- <node>: Extend a new reasoning branch
- <backtrack>: Backtrack to an ancestor node
- <done>: Terminate the reasoning trajectory

autoregressive sequence generation ≈ constructing a dynamic reasoning tree in the semantic space.



Method

- **Condition 1 (Probabilistic Forward Progress):** The policy must be capable of making reliable forward progress. To any correct but incomplete reasoning path $v_1, \ldots, v_i$ where $i < T_{\max}$, the policy π_θ must be γ**-progressive** with a probability no less than $1 - \epsilon$. A policy is defined as γ-progressive if it generates a correct continuation node v_{i+1} with probability no less than γ.
- **Condition 2 (Reliable Trajectory Recovery):** The policy must be robust to its own errors. Formally, for any incorrect reasoning path c that deviates from a correct path at node i, π_θ must induce a backtracking action to a valid ancestor node with a probability of at least $1 - \epsilon$.

Experiment Result

Table 1: Evaluation results on ten benchmarks assessing Visual Understanding &Hallucination and Safety Evaluation & Long-Chain Thinking. Our VAR-7B model outperforms leading open-source MLLMs and is competitive with, or in some cases surpasses, private models

	Avg	Visual Understanding & Hallucination					Safety Evaluation & Long-Chain Thinking				
		HalB	MMB	MMMU	SEED	V^*B	MGD	MSSB	SafeB	VLSB	RSB
Private Models											
Gemini-2.5-Flash	73.3	72.9	82.9	63.9	83.2	83.8	49.6	67.5	97.2	66.1	65.5
GPT-4o-1120	73.6	75.2	82.3	69.5	82.5	82.2	57.8	69.2	96.5	69.4	70.8
Gemini-2.5-Pro	78.8	76.2	85.1	68.5	86.9	92.3	55.3	73.2	98.3	75.9	76.8
Claude-3.7-Sonnet	79.6	77.3	87.2	69.2	85.3	95.5	53.7	72.1	99.5	82.3	76.2
Open-source General Models											
LLaVA-OneVision-7B	52.7	46.9	67.2	51.3	65.5	75.4	24.3	58.8	72.5	12.6	51.2
VILA-1.5-7B	51.3	45.8	68.5	51.9	63.5	72.3	16.3	52.5	69.8	14.7	57.5
Qwen2.5-VL-7B	54.3	48.3	69.2	52.8	68.2	71.2	27.9	55.4	76.2	19.0	54.8
Qwen2.5-VL-32B	60.6	48.1	75.4	54.8	69.6	87.9	43.4	55.3	87.2	26.3	58.1
Qwen2.5-VL-72B	64.2	55.6	78.3	57.6	72.3	90.6	41.5	59.1	92.1	35.2	59.4
InternVL3-8B	58.3	50.2	75.3	51.6	67.4	76.3	42.2	54.6	83.2	23.1	58.4
GLM-4v-9B	56.3	51.1	72.1	52.2	62.2	78.8	23.4	50.9	88.6	22.5	60.5
MiniCPM-V-2.6	53.2	47.2	68.4	50.3	63.5	74.8	32.2	48.2	75.5	16.1	56.2
Open-source Visual Reasoning Models											
DeepEyes-7B	59.6	49.2	70.6	53.8	65.2	90.1	51.5	52.2	85.2	26.6	51.7
Pixel-Reasoner-7B	58.6	35.7	69.8	53.5	66.1	89.8	48.5	54.1	86.5	27.5	54.1
VAR-7B	**72.1**	55.5	**78.5**	58.8	**79.3**	90.3	**63.1**	**74.8**	**98.5**	**43.5**	**79.1**
Δ *v.s.* Qwen2.5-VL-7B	↑ 17.8	↑ 7.2	↑ 9.3	↑ 6.0	↑ 11.1	↑ 19.1	↑ 35.2	↑ 19.4	↑ 22.3	↑ 24.5	↑ 24.3
Δ *v.s.* Qwen2.5-VL-32B	↑ 11.5	↑ 7.4	↑ 3.1	↑ 4.0	↑ 9.7	↑ 2.4	↑ 19.7	↑ 19.5	↑ 11.3	↑ 17.2	↑ 21.0
Δ *v.s.* Qwen2.5-VL-72B	↑ 8.0	↓ 0.1	↑ 0.2	↑ 1.2	↑ 7.0	↓ 0.3	↑ 21.6	↑ 15.7	↑ 6.4	↑ 8.3	↑ 19.7

Experiment Result

Table 2: VAR-7B vs. Base Model: Performance on Visual Reasoning, Vision-Centric, and Document Understanding Tasks

Capability	Benchmark	Qwen2.5-VL-7B	VAR-7B	Qwen2.5-VL-72B
Visual-Reasoning-QA	MMBench	70.3	**79.5** ↑ 9.2	78.3
	POPE	85.7	**87.5** ↑ 1.8	84.9
	HallusionBench	48.3	**55.5** ↑ 1.9	55.6
Vision-Centric-QA	CV-Bench-2D	74.0	**78.9** ↑ 4.9	77.7
	CV-Bench-3D	72.3	**79.6** ↑ 7.3	87.1
	MMVP	66.6	**75.1** ↑ 8.5	66.6
Document and chart	AI2D	**85.9**	85.7 ↓ 0.2	88.7
	ChartQA	85.5	**86.8** ↑ 1.3	89.5

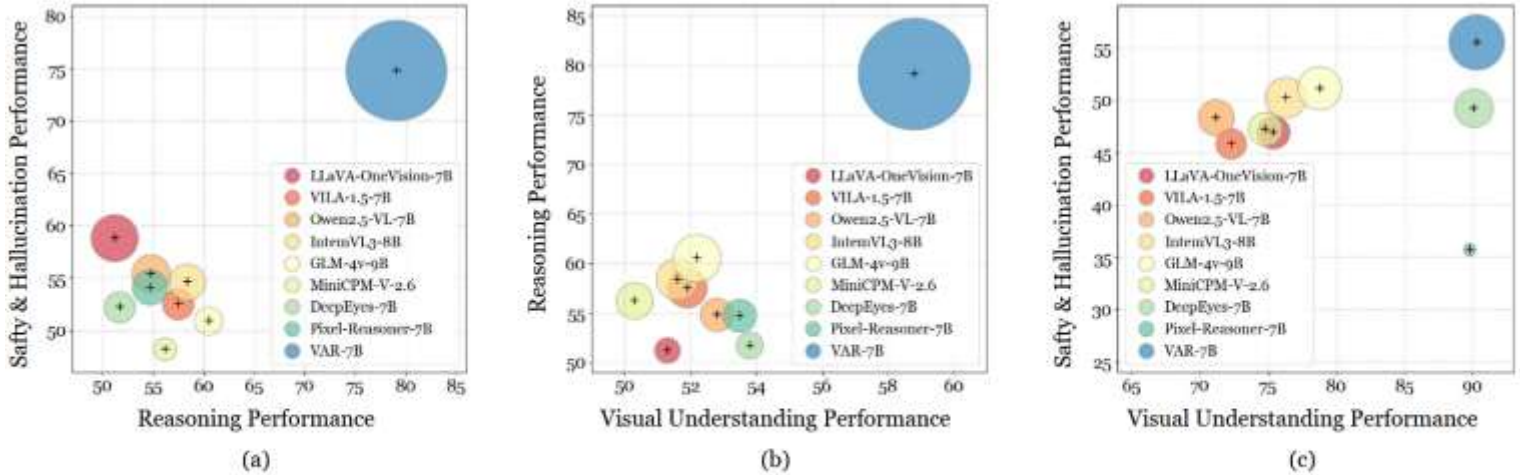


Figure 3: Correlation of Model Capabilities

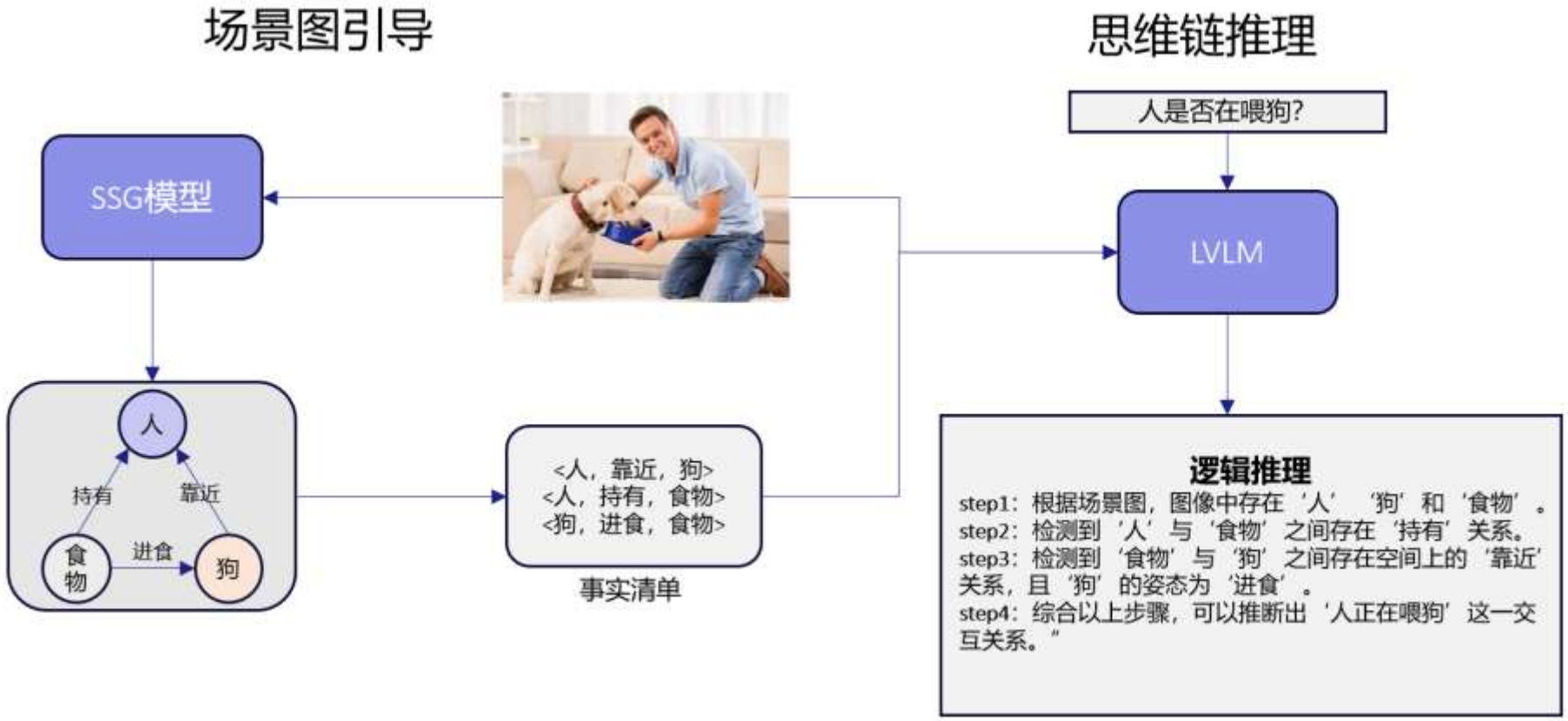
Ablation

Table 3: Ablations of each component of our VAR.

		Cold-Start	Backtrack	Rewards			V*B	HallusionBench	RSB
				$R_{\text{acc}} + R_{\text{fmt}}$	R_{sem}	R_{geo}	Acc	Acc	Acc
①	Qwen2.5-VL-7B						71.2	48.3	54.8
②	Cold-Start	✓					75.4	49.6	60.3
③	**VAR-7B**	✓	✓	✓	✓	✓	**90.3**	**55.5**	**79.1**
④	w/o Trace	✓	✓	✓			83.9	51.6	65.3
⑤	w/o Geo	✓	✓	✓	✓		86.7	53.9	72.1
⑥	w/o Sem	✓	✓	✓		✓	88.5	54.1	72.9
⑦	w/o Backtrack	✓		✓	✓	✓	87.1	52.3	68.9
⑧	Text-Only RL			✓			81.8	50.3	62.5

Contribution： Backtrack ≫ Rsem ≫ Rgeo

Idea


场景图引导
思维链推理
SSG模型
人
持有
靠近
食物
进食
狗
<人，靠近，狗>
<人，持有，食物>
<狗，进食，食物>
事实清单
人是否在喂狗?
LVLM
逻辑推理
step1：根据场景图，图像中存在‘人’ ‘狗’ 和‘食物’。
step2：检测到‘人’与‘食物’之间存在‘持有’关系。
step3：检测到‘食物’与‘狗’之间存在空间上的‘靠近’关系，且‘狗’的姿态为‘进食’。
step4：综合以上步骤，可以推断出‘人正在喂狗’这一交互关系。”

Thanks