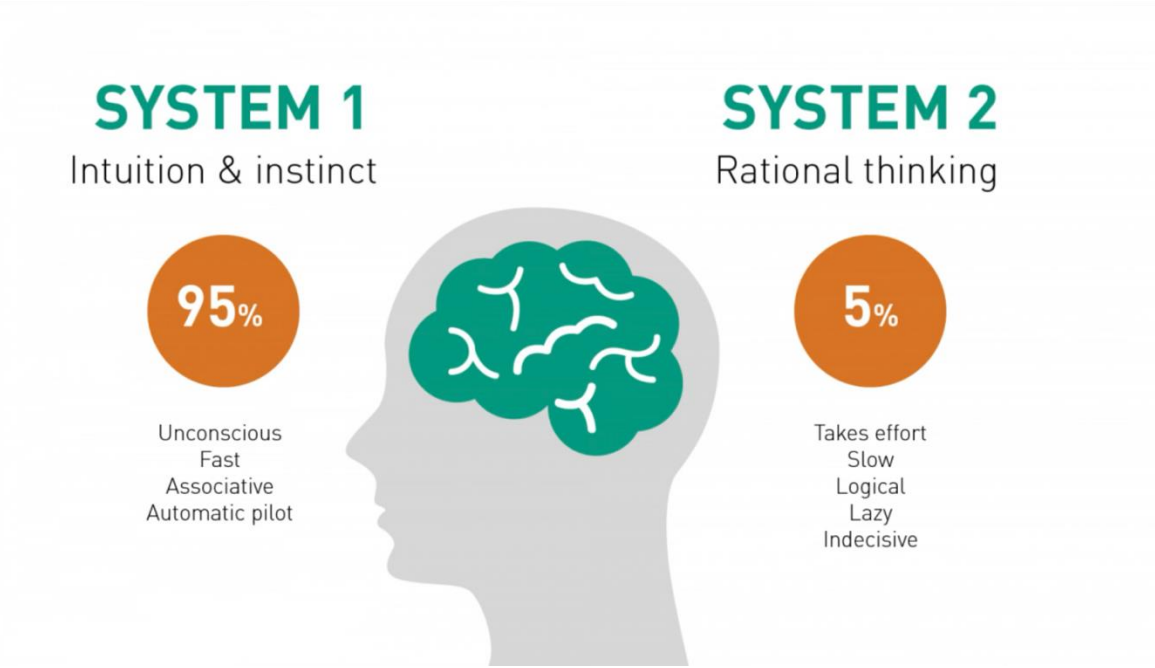
Visual Agents as Fast and Slow Thinkers

Guangyan Sun*, Mingyu Jin*, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, Dongfang Liu

ICLR 2025

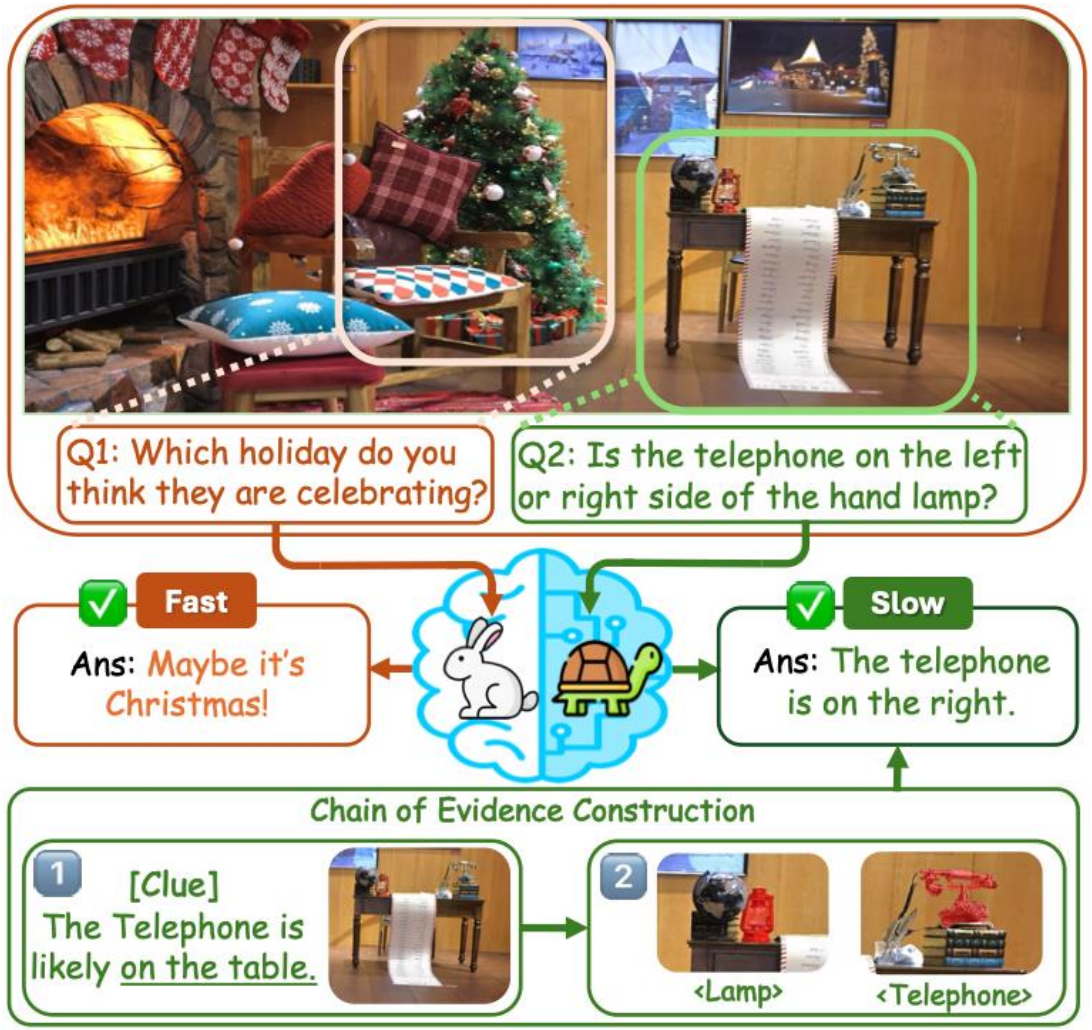
Background

- Achieving human-like intelligence requires System 1 (fast, intuitive) and System 2 (slow, analytical) thinking.
- Current Visual Agents Issues:
  - Over-reliance on heuristic reasoning (System 1)
  - Lack of explicit slow-thinking mechanisms (System 2)
  - Hallucinations and overconfidence in responses
- **How can we enable visual agents to adaptively switch between System 1 & 2?**

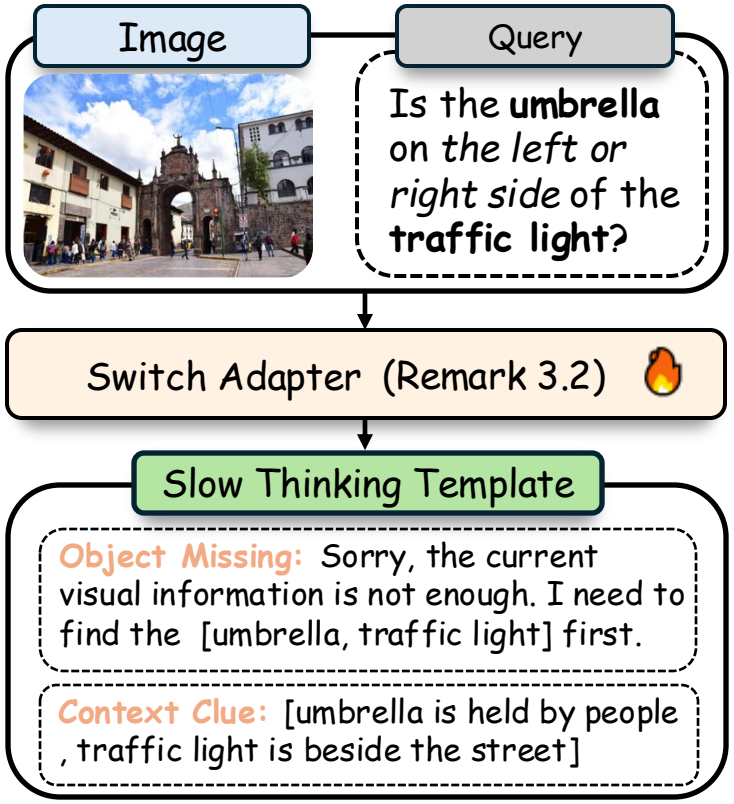


citation

FaST Framework



FaST Framework for Slow Thinking Mode

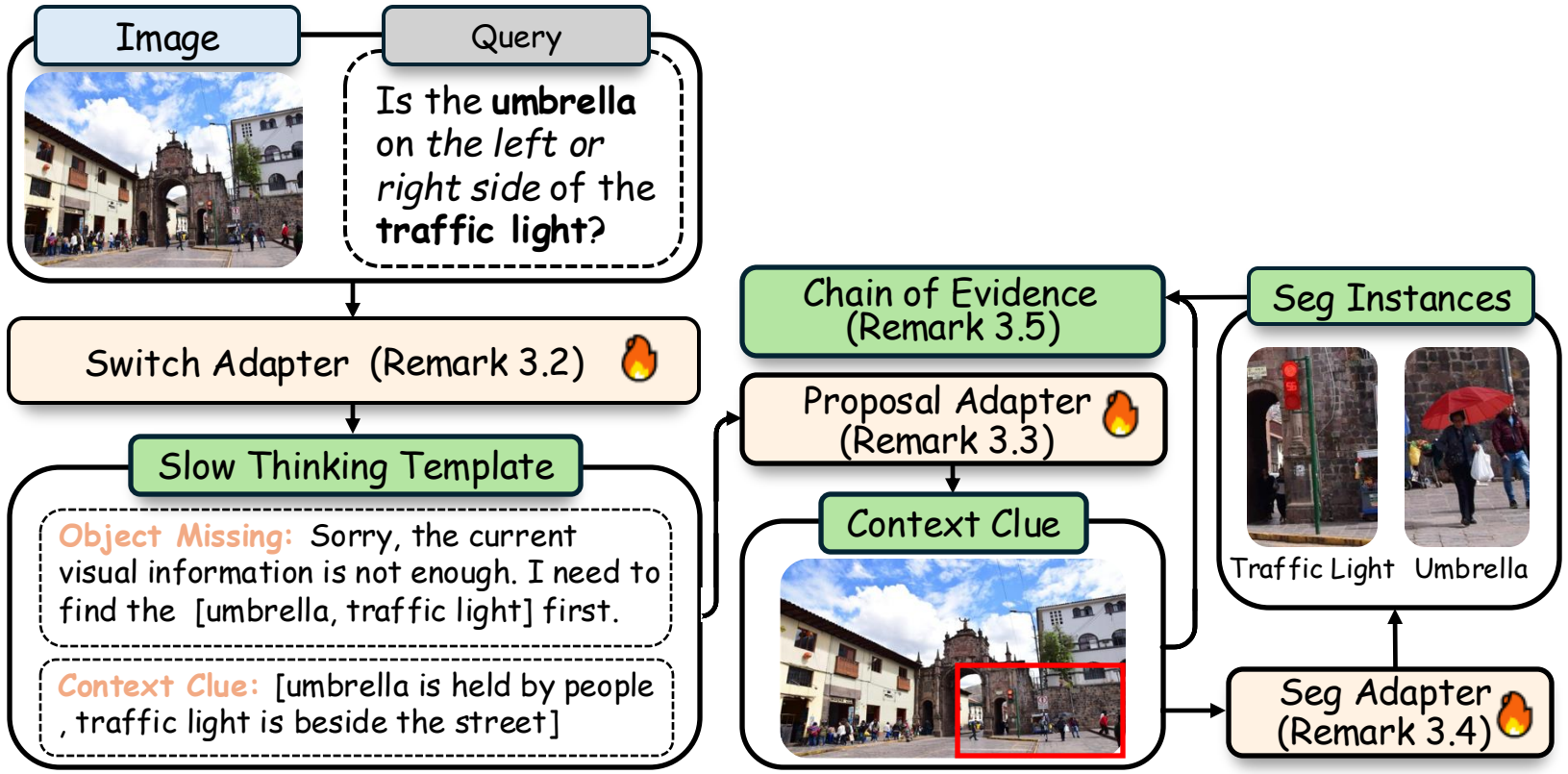


Negative Data for Slow Thinking

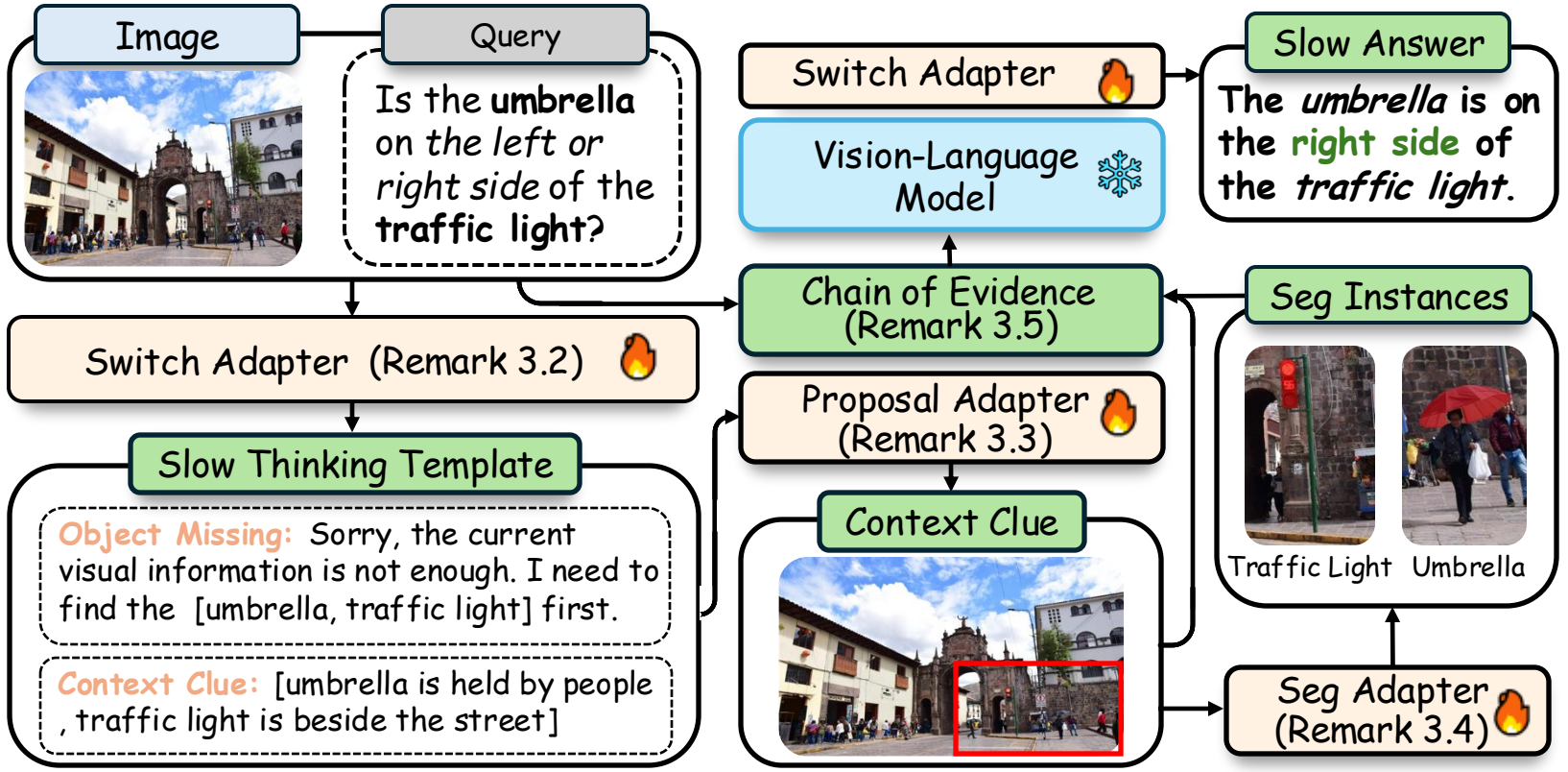
100k of (image, question, answer)

1. Objects too small
2. Complex reasoning required

FaST Framework for Slow Thinking Mode



FaST Framework for Slow Thinking Mode



Benchmark Results

Method	LLM	VQA Datasets				Multimodal Benchmarks			
		VQA^{v2}	GQA	VQA^{T}	SQA^{I}	POPE	MME	SEED	MM-Vet
BLIP-2[ICML23]	Vicuna-13B	65.0	32.3	42.5	61.0	85.3	1293.8	46.4	22.4
InstructBLIP[NeurIPS24]	Vicuna-13B	-	49.5	50.7	63.1	78.9	1212.8	53.4	25.6
Qwen-VL-Chat[arXiv23]	Qwen-7B	78.2	57.5	61.5	68.2	-	1487.5	58.2	-
mPLUG-Owl2[CVPR24]	LLaMA-7B	79.4	56.1	58.2	68.7	-	1450.2	**61.6**	**36.2**
Monkey[CVPR24]	Qwen-7B	80.3	60.7	-	**69.4**	67.6	-	-	-
LLaVA-v1.5[CVPR24]	Vicuna-7B	78.5	62.0	58.2	66.8	85.9	1510.7	58.6	30.5
Chain of Spot [arXiv24]	Vicuna-7B	80.7	63.7	60.9	68.2	86.4	1501.1	59.7	30.8
V*[CVPR24]	Vicuna-7B	-	-	-	-	82.4	1128.9	41.7	27.7
Visual CoT[arXiv24]	Vicuna-7B	-	63.1	**77.5**	-	-	-	-	-
FAST (Ours)	Vicuna-7B	**80.8**	**63.8**	60.7	68.9	**86.4**	**1517.4**	60.1	31.0
Δ (vs LLaVA-v1.5)	Vicuna-7B	+2.3	+1.8	+ 2.5	+2.1	+0.4	+6.7	+ 1.5	+ 0.5

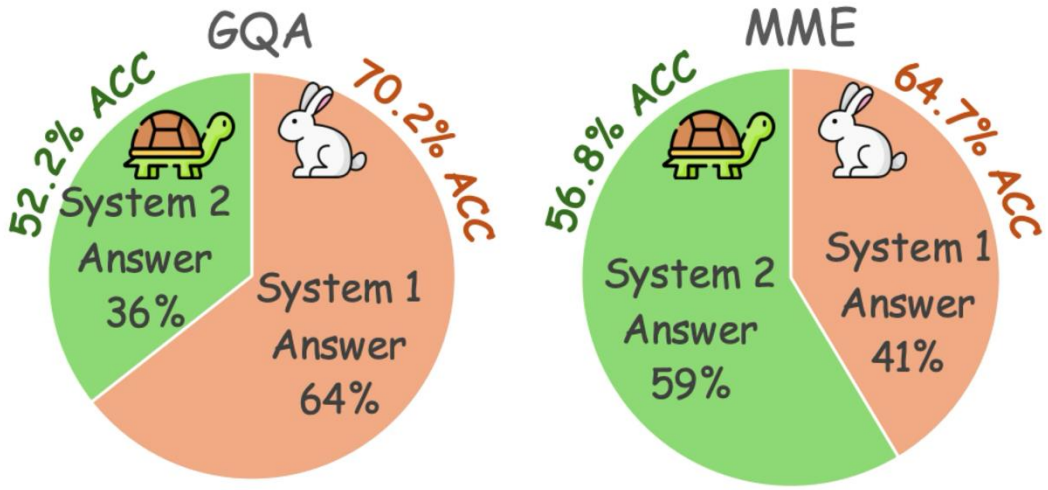
Main Results on VQA and Multimodal Benchmark

Method	Referring Segmentation			Reasoning Segmentation	
	refCOCO	refCOCO+	refCOCOg	ReasoSeg	
	CIoU	CIoU	CIoU	CIoU	GIoU
LAVT[CVPR22]	72.7	62.1	61.2	-	-
OVSeg[CVPR23]	-	-	-	28.5	18.6
GRES[CVPR23]	73.8	**66.0**	65.0	22.4	19.9
X-Decoder[CVPR23]	-	-	64.6	22.6	17.9
SEEM[NeurIPS24]	-	-	65.7	25.5	21.2
LISA-7B[CVPR24]	**74.1**	62.4	66.4	44.4	46.0
LLaVA w Seg Adapter	70.8	57.5	64.0	43.0	41.0
FAST (Ours)	73.3	64.4	**67.0**	**47.6**	**48.7**

Main Results on Referring and Reasoning Segmentation

Analysis of Switch Adapter

- Impact of Switch Adapter (MME)
  - 41% resolved with System 1
  - 59% need System 2 for deeper reasoning
  - High Accuracy for Easy Problems
  - Better Accuracy with System 2 Mode



System 1 Mode Analysis. We investigate the system switching ratio, along with fast thinking performance on easy or hard queries defined by the switch adapter.

Conclusion

- FaST is a novel framework integrating System 1 and 2 Thinking for Visual Agents
- Superior performance across multiple benchmarks
- Transparent decision-making through a chain of evidence construction

Paper

Code