DocThinker: Explainable Multimodal Large Language Models with Rule-based Reinforcement Learning for Document Understanding

Wenwen Yu[1], Zhibo Yang[2], Yuliang Liu[1], Xiang Bai[1]✉

[1]Huazhong University of Science and Technology, [2]Alibaba Group

ICCV 2025

Presenter：Wenwen Yu

2025-10-22

DocThinker: Explainable Multimodal Large Language Models with Rule-based Reinforcement Learning for Document Understanding

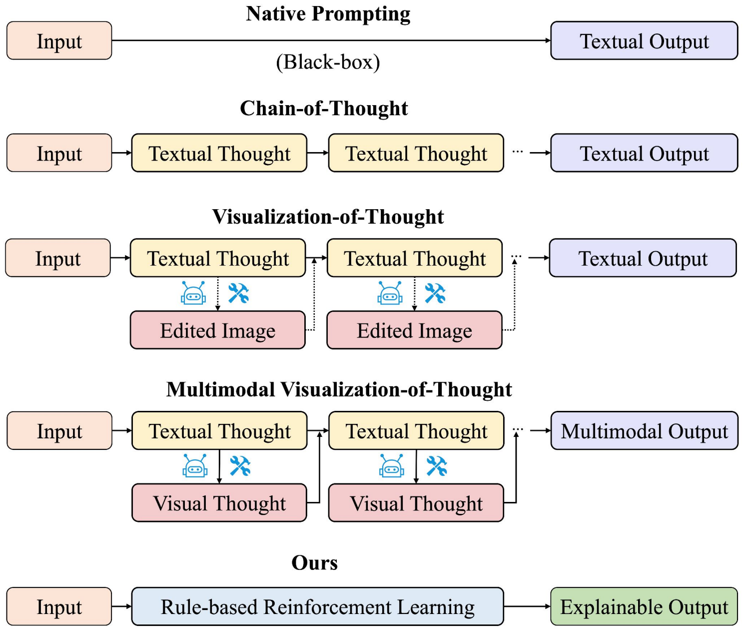
Wenwen Yu[1] · Zhibo Yang[2] · Yuliang Liu[1] · Xiang Bai[1]

[1]*Huazhong University of Science and Technology* [2]*Alibaba Group*

ICCV HONOLULU HAWAII OCT 19-23, 2025

Introduction:

- Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in document understanding. However, their reasoning processes remain largely black-box, making it difficult to ensure reliability and trustworthiness, especially in high-stakes domains such as legal, financial, and medical document analysis.
- Existing methods use fixed Chain-of-Thought (CoT) reasoning with supervised fine-tuning (SFT) but suffer from catastrophic forgetting, poor adaptability, and limited generalization across domain tasks.

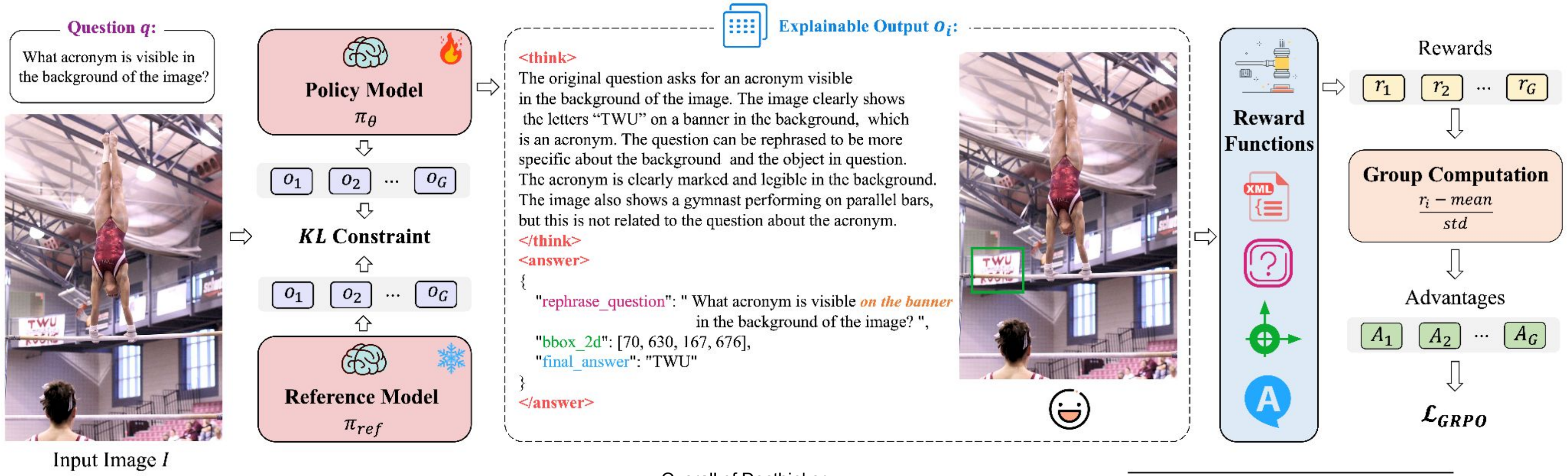


Comparison of different approaches for improving model's explainability and transparency in MLLM-based document understanding.

- We propose Docthinker, a rule-based Reinforcement Learning (RL) framework for dynamic inference-time reasoning.

Methods:

- IInstead of relying on static CoT templates, \ourmodel autonomously refines reasoning strategies via policy learning, generating explainable intermediate results, including structured reasoning processes, rephrased questions, regions of interest (RoI) supporting the answer, and the final answer. By integrating multi-objective rule-based rewards and KL-constrained optimization, our method mitigates catastrophic forgetting and enhances both adaptability and transparency.



Overall of Docthinker

Model	Template1	Template2
Specialist Models		
TransVG [6]	50.1	54.0
MAttNet [48]	52.3	60.5
QRNet [45]	52.7	59.1
MDETR [14]	54.4	63.3
TAMN [9]	77.8	80.8
DocThinker-7B	82.4	

Experiments:

- DocThinker significantly improves generalization while producing more explainable and human-understandable reasoning steps.

MLLM	Res.	Data	Str.	Document-oriented Understanding						General Multimodal Understanding				
				Doc/Text					Chart	General VQA		Relation Reasoning		
				DocVQA	TextCaps	TextVQA	DUDE	SROIE	InfoQA	F30k	V7W	GQA	OI	VSR
LLaVA-1.5-7B [22]	336^2	-	SFT	0.244	0.597	0.588	0.290	0.136	0.400	0.581	0.575	0.534	0.412	0.572
LLaVA-1.5-13B [22]	336^2	-	SFT	0.268	0.615	0.617	0.287	0.164	0.426	0.620	0.580	0.571	0.413	0.590
SPHINX-13B [18]	224^2	-	SFT	0.198	0.551	0.532	0.000	0.071	0.352	0.607	0.558	0.584	0.467	0.613
VisCoT-7B [35]	224^2	438k	SFT	0.355	0.610	0.719	0.279	0.341	0.356	0.671	0.580	0.616	**0.833**	0.682
VisCoT-7B [35]	336^2	438k	SFT	0.476	0.675	0.775	0.386	0.470	0.324	0.668	0.558	0.631	0.822	0.614
Qwen2.5VL-7B† [1]	336^2	-	-	0.350	0.642	0.735	0.202	0.472	0.325	0.603	0.556	0.455	0.347	0.616
Qwen2.5VL-7B† [1]	1536^2	-	-	0.773	0.710	0.792	0.492	0.708	0.663	0.685	0.604	0.457	0.371	0.603
Qwen2.5VL-7B* [1]	336^2	4k	SFT	0.355	0.658	0.740	0.215	0.489	0.334	0.624	0.563	0.467	0.405	0.619
Qwen2.5VL-7B* [1]	1536^2	4k	SFT	0.784	0.725	0.801	0.498	0.714	0.674	0.680	0.609	0.472	0.427	0.624
DocThinker-3B	336^2	4k	RL	0.460	0.663	0.746	0.213	0.486	0.335	0.664	0.572	0.486	0.485	0.625
DocThinker-3B	1536^2	4k	RL	0.751	0.691	0.762	0.469	0.735	0.566	0.682	0.583	0.490	0.517	0.637
DocThinker-7B	336^2	4k	RL	0.579	0.682	0.802	0.408	0.495	0.347	0.674	0.580	0.546	0.542	0.656
DocThinker-7B	1536^2	4k	RL	0.795	0.738	0.827	0.515	0.806	0.689	0.701	0.625	0.694	0.686	0.721
DocThinker-7B	1536^2	8k	RL	**0.802**	**0.757**	**0.836**	**0.568**	**0.814**	**0.697**	**0.734**	**0.641**	**0.737**	0.784	**0.768**

Conclusion

- This paper introduced DocThinker, a reinforcement learning-based framework designed to enhance explainability, adaptability, and reasoning ability in multimodal document understanding. DocThinker achieves state-of-the-art or highly competitive performance on standard benchmarks