ICML 2026

ImgCoT: Compressing Long Chain of Thought into Compact Visual Tokens for Efficient Reasoning of Large Language Model

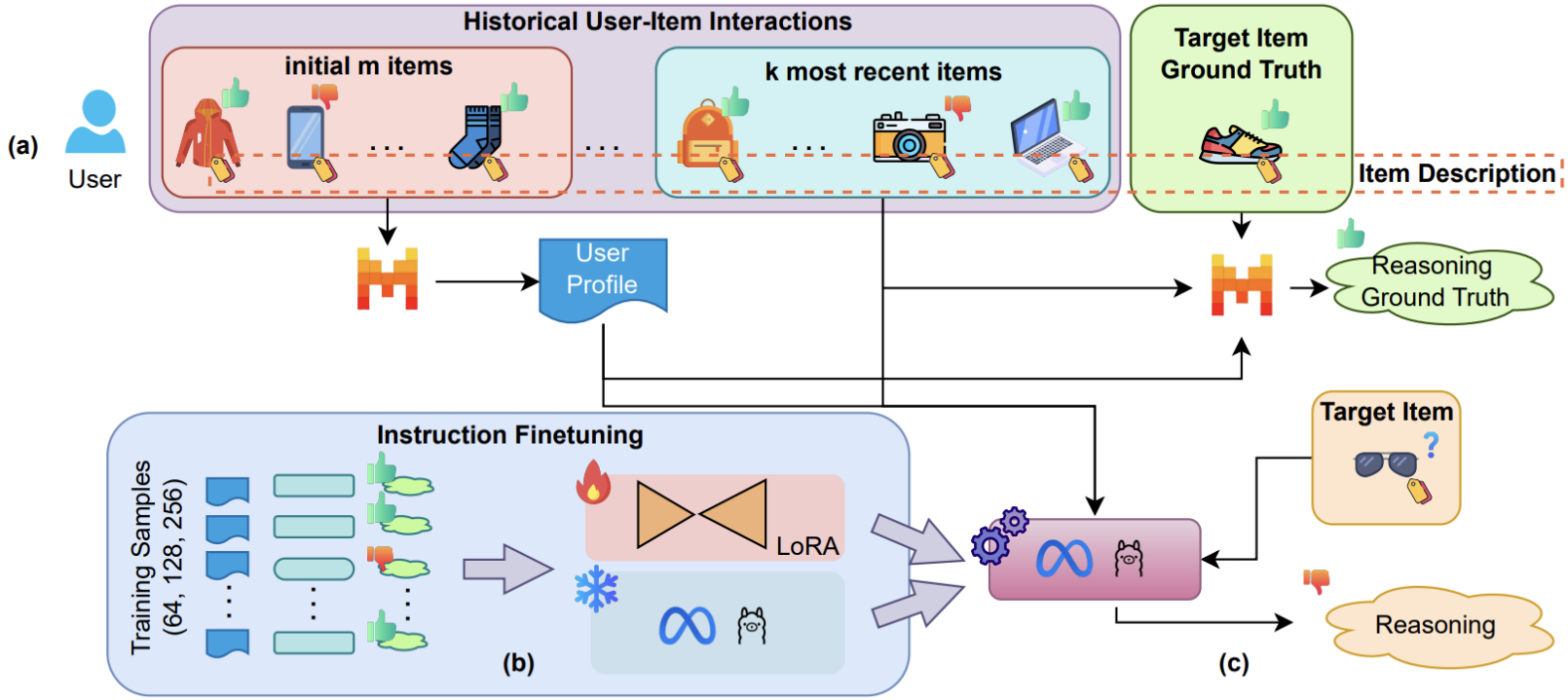
Xiaoshu Chen[1], Sihang Zhou[1], Ke Liang[1], Taichun Zhou[1], Yaohua Wang[1], Yang Gao[2], Xinwang Liu[1]

1National University of Defense Technology 2Nanjing University

CONTENTS

Background:

- The lengthy nature of Chain-of-Thought (CoT) reasoning in LLMs limits their applicability in time-sensitive real-world scenarios, such as recommendation and search systems.



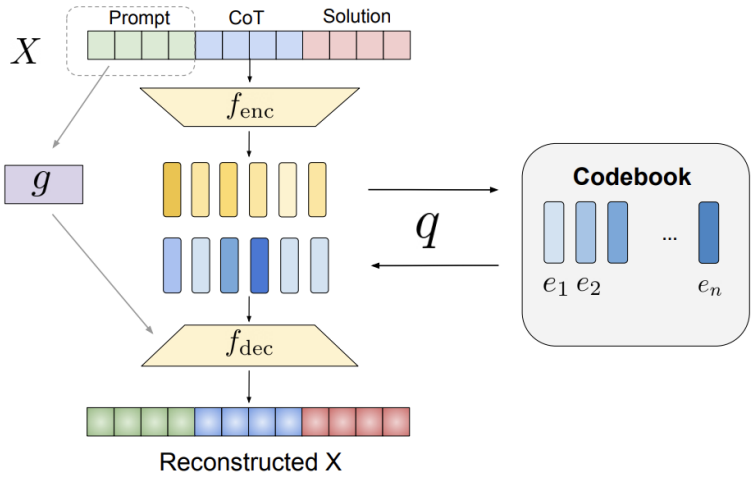
ReasoningRec framework[1]

An extra millisecond of response time causes massive number of usersexit interface

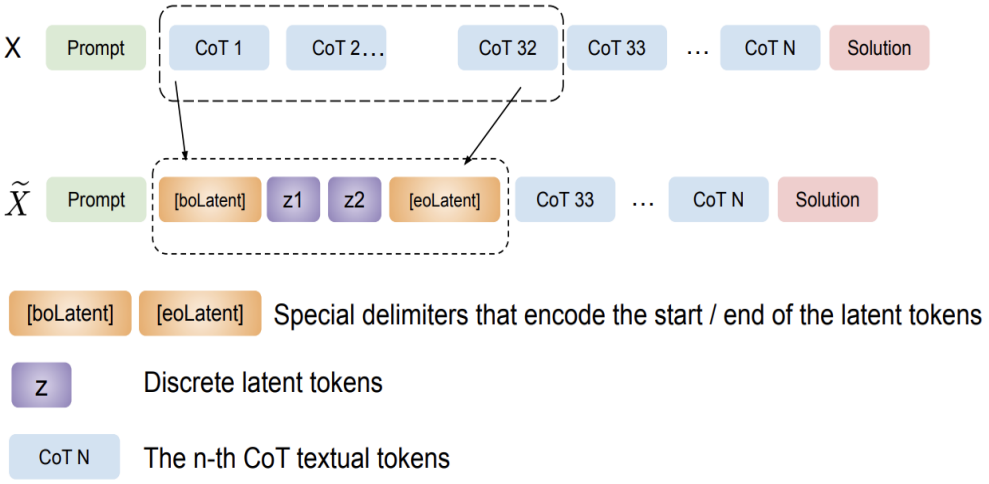
[1] ReasoningRec: Bridging Personalized Recommendations and Human-Interpretable Explanations through LLM Reasoning, in NAACL 2025

Related Work

- **Mainstream methods:** Accelerating Chain-of-Thought Reasoning through Latent CoT Reasoning.
- **Best performance:** Abort tokens[1]--Based on a Quantized Variational Autoencoder (VQVAE), it employs CoT text as its reconstruction target, compressing the verbose CoT into a latent space to yield compressed, latent CoT tokens.



Compression by VQVAE

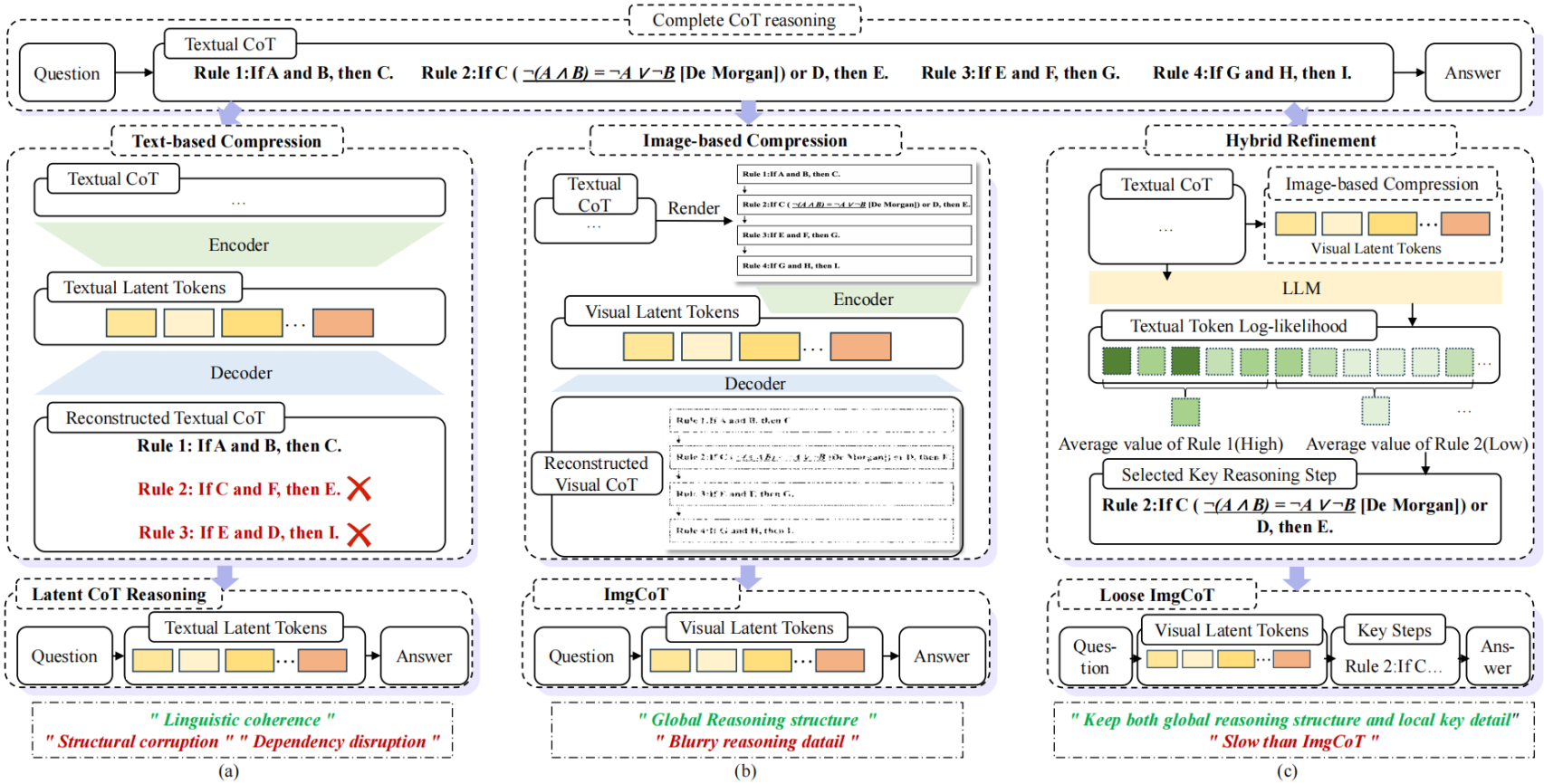


Reasoning with latent tokens

[1] Token Assorted: Mixing Latent and Text Tokens for Improved Language Model Reasoning, in ICLR2025

Motivation

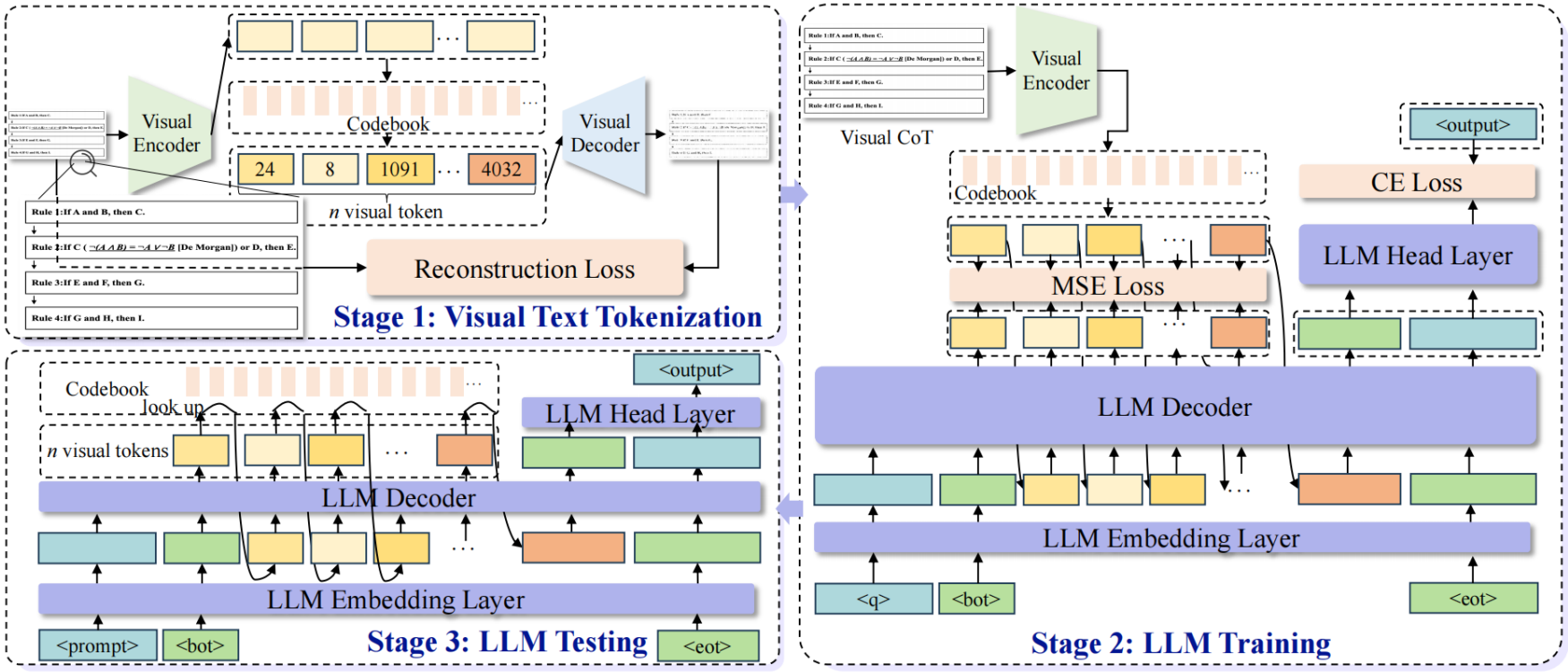
- Text reconstruction objectives exhibit linguistic inductive bias.
- The visual inductive bias inherent in visual reconstruction objectives is better suited for compressing the CoT into the latent space.
- Compressing into the latent space results in the loss of certain domain-specific reasoning details.



CONTENTS

Method

- ImgCoT
  - Compress the rendered CoT into the latent space using VQVAE.
  - Train an LLM to perform reasoning based on visual latent tokens.
  - During inference, utilize the latent tokens to perform reasoning in an autoregressive manner.
- L-ImgCoT: L-ImgCoT preserves key steps by leveraging log-likelihood scores.



CONTENTS

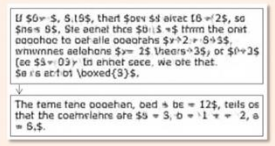
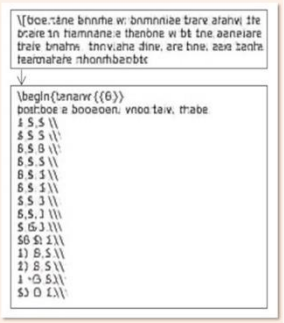
Main Results

- 8 latent tokens achieve an effect similar to full-CoT.

	Model	MATH		GSM		GPQA		ProsQA	
		Acc ↑	*# Tokens*↓	*Acc*↑	*# Tokens* ↓	*Acc* ↑	*# Tokens*↓	*Acc*↑	*# Tokens* ↓
Qwen2.5-0.5B -Instruction	Full-CoT	9.2	149.4	16.9	116.3	34.5	150.2	94.6	71.3
	Coconut (COLM'25)	3.5	6.0	3.7	6.0	26.0	6.0	73.8	6.0
	ICoT	3.3	0.0	3.9	0.0	**39.6**	0.0	54.9	0.0
	CODI (EMNLP'25)	0.4	6.0	0.5	6.0	27.1	6.0	0.0	6.0
	CoLaR (NeurIPS'25)	3.1	39.4	5.8	19.1	27.7	28.7	68.9	18.9
	ImgCoT(ours)	9.8	8.0	9.2	8.0	34.5	8.0	97.4	8.0
	w/o layout	9.1	8.0	9.2	8.0	32.7	8.0	96.8	8.0
	w/ textual tokens	9.0	8.0	8.3	8.0	27.3	8.0	96.2	8.0
	L-ImgCoT(ours)	**10.1**	102.8	**17.5**	64.7	38.1	89.3	**98.6**	40.9
Qwen2.5-1.5B -Instruction	Full-CoT	19.4	238.7	44.1	126.6	40.0	229.5	99.6	46.2
	Coconut (COLM'25)	8.8	6.0	5.4	6.0	31.4	6.0	99.6	6.0
	ICoT	12.5	0.0	22.3	0.0	30.7	0.0	98.6	0.0
	CODI (EMNLP'25)	4.3	6.0	3.6	6.0	42.8	6.0	99.6	6.0
	CoLaR (NeurIPS'25)	4.9	60.5	6.9	22.9	37.1	30.8	99.4	7.1
	ImgCoT(ours)	19.5	8.0	38.7	8.0	41.8	8.0	**100.0**	8.0
	w/o layout	19.1	8.0	38.4	8.0	36.4	8.0	**100.0**	8.0
	w/ textual tokens	17.4	8.0	37.5	8.0	36.4	8.0	99.2	8.0
	L-ImgCoT(ours)	**19.9**	127.4	**45.2**	71.3	**43.6**	130.8	**100.0**	21.7
LLama3.2-3B -Instruction	Full-CoT	23.8	205.9	60.5	105.6	**45.5**	179.3	**100.0**	41.3
	Coconut (COLM'25)	13.8	6.0	19.5	6.0	28.6	6.0	**100.00**	6.0
	ICoT	15.0	0.0	46.7	0.0	28.9	0.0	99.2	0.0
	CODI (EMNLP'25)	-	-	23.3	6.0	-	-	-	-
	CoLaR (NeurIPS'25)	19.1	41.2	27.4	20.1	20.4	33.9	**100.0**	6.8
	ImgCoT(ours)	24.1	8.0	56.8	8.0	43.6	8.0	**100.0**	8.0
	w/o layout	23.6	8.0	52.7	8.0	43.6	8.0	**100.0**	8.0
	w/ textual tokens	22.6	8.0	49.3	8.0	41.8	8.0	**100.0**	8.0
	L-ImgCoT(ours)	**24.3**	143.8	**61.0**	63.2	**45.5**	115.7	**100.0**	27.5

Ablation Study

- **Qualitative comparison**--between the original CoT and its reconstructions from different latent representations.
- **Performance comparison**--under varying numbers of visual versus textual latent tokens.

Original CoT	Reconstructed Textual CoT	Reconstructed Visual CoT
If $b=-1, d=5$, then $ac+b+d=ac+4=-2$, so $ac=-6$. We substitute $a=1-c$ from the first equation to get the quadratic $c^2-c-6=0$, which has solutions $c=-2$ (so $a=3$) or $c=3$ (so $a=-2$). In either case, we get that $a+b+c+d=\boxed{5}$. The remaining equation, $ad + bc = 17$, tells us that the coefficients are $a = 3, b = -1, c = -2, d = 5.$	Let A be a positive integer and B a positive integer. Let $A^2=\{1,2,3,4,5\}$ and $B^2=\{1,2,3,4,5\}$. Then $A^2=\{1,2,3,4,5\}$ and $B^2=\{1,2,3,4,5\}$. I know that $A^2=\{1,2,3,4,5\}$ and $B^2=\{1,2,3,4,5\}$. 	
\]Since the numbers less than 15 are small, we can easily apply this process to all the numbers from 1 to 15. Here is a table showing how many factors each number has: \begin{tabular}{c|c} number & how many factors \\ \hline 1 & 1 \\ 2 & 2 \\ 3 & 2 \\ 4 & 3 \\ 5 & 2 \\ 6 & 4 \\ 7 & 2 \\ 8 & 4 \\ 9 & 3 \\ 10 & 4 \\ 11 & 2 \\ 12 & 6 \\ 13 & 2 \\ 14 & 4 \\	The first digit of the first digit The second digit of the second digit The third digit of the third digit The third digit of the third digit The third digit of the third digit The third digit of the third digit The third digit of the third digit The third digit of the third digit The third digit of the third digit	

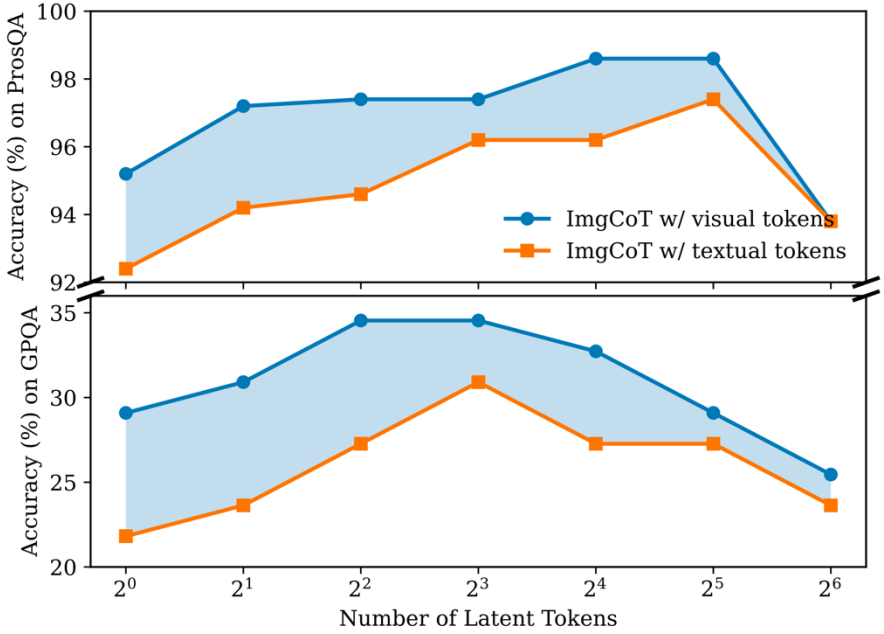


Image-based Compression vs. Text-based Compression

Generalization

- Out-of-domain generalization of visual and textual latent tokens at inference.

Model	In-Domain	Out-of-Domain			
	GSM	Gaokao	SVAMP	SingleEq	MultiArith
ImgCoT	11.4	3.1	9.6	12.0	5.8
w/o visual tokens	10.0	1.8	5.6	8.4	4.2
Difference	↓ 1.4	↓1.3	↓3.0	↓3.6	↓ 1.6

- Performance Comparison of L-ImgCoT With and Without Visual Latent Tokens. The LLM is Qwen2.5-0.5B-Instruction.

Method	MATH	GSM	GPQA	ProsQA
L-ImgCoT *w/o* visual tokens	9.1	15.8	29.1	94.2
L-ImgCoT *w/* visual tokens	10.1	17.5	38.1	98.6

CONTENTS

Takeaways

- Visual reconstruction objectives are better suited than text reconstruction objectives for compressing the CoT into the latent space.
- In certain scenarios, ImgCoT—utilizing just 8 visual latent tokens—achieves performance comparable to that of the full CoT.
- ImgCoT still faces challenges such as a lack of transparent in its reasoning process and training instability.

Thanks!