



Thinking in Latent Space: Progressive Multimodal Simplification for Visual Reasoning

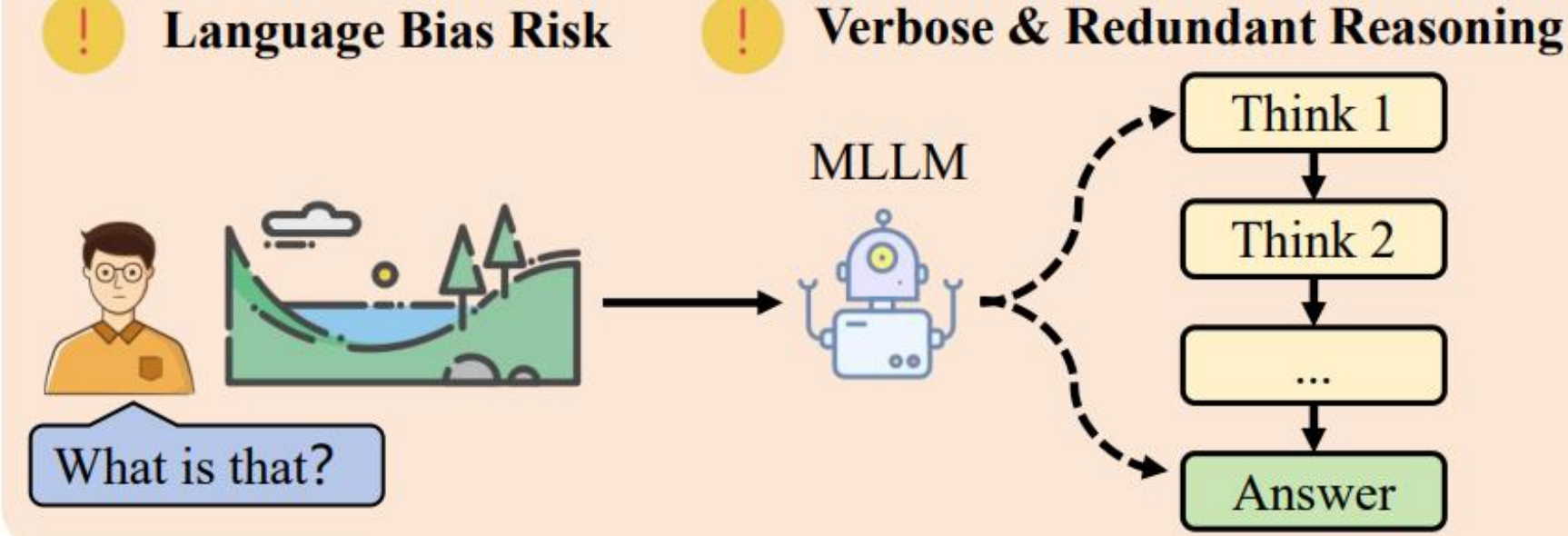
Yuesen Tang^{*1}, Yiming Yang^{*1}, Tengfei Bao^{*1}, Yu Tong^{2,3}

^{*}Equal contribution ¹Southeast University ²Wuhan University ³Xiaohongshu

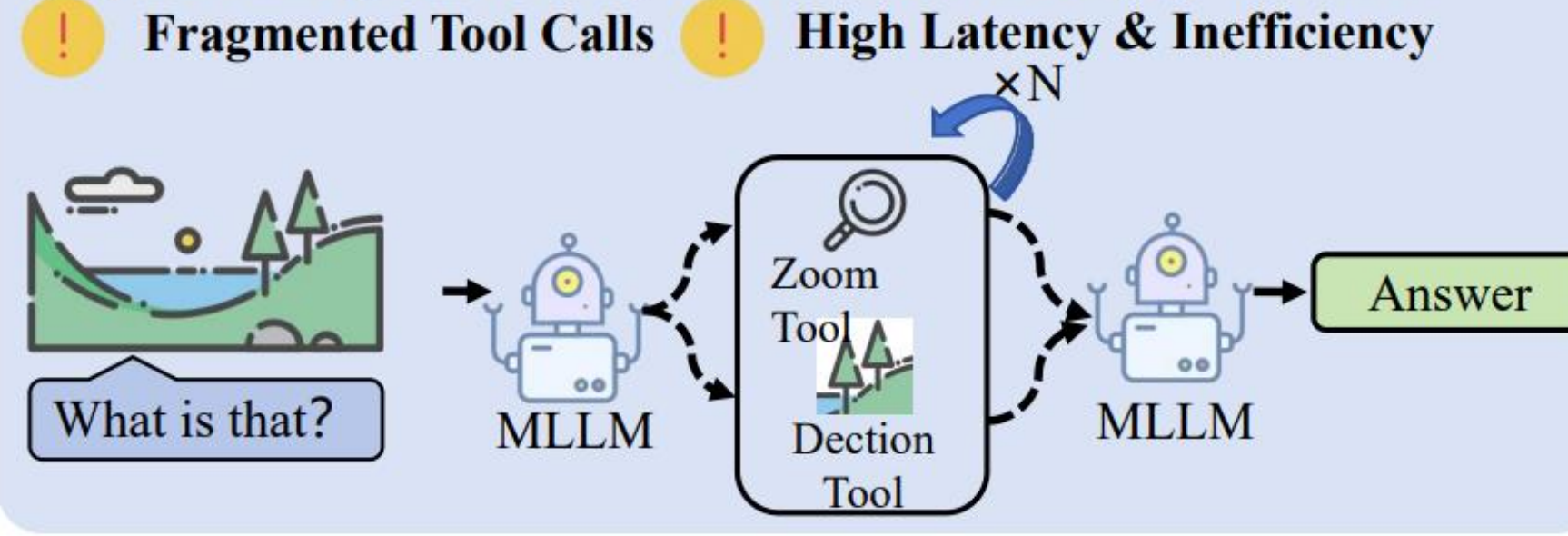


1. Introduction

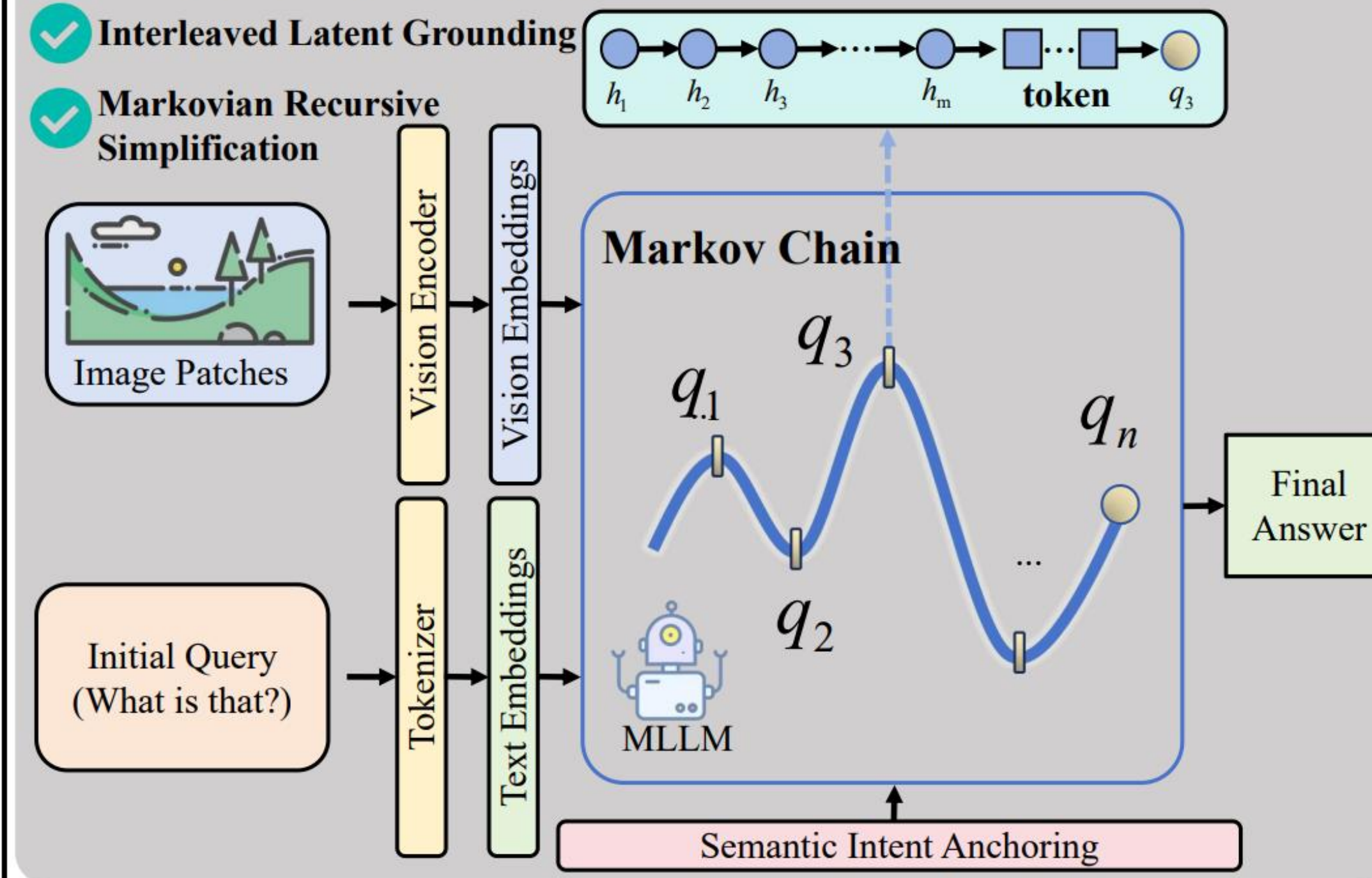
(A) Standard CoT (Textual-only)



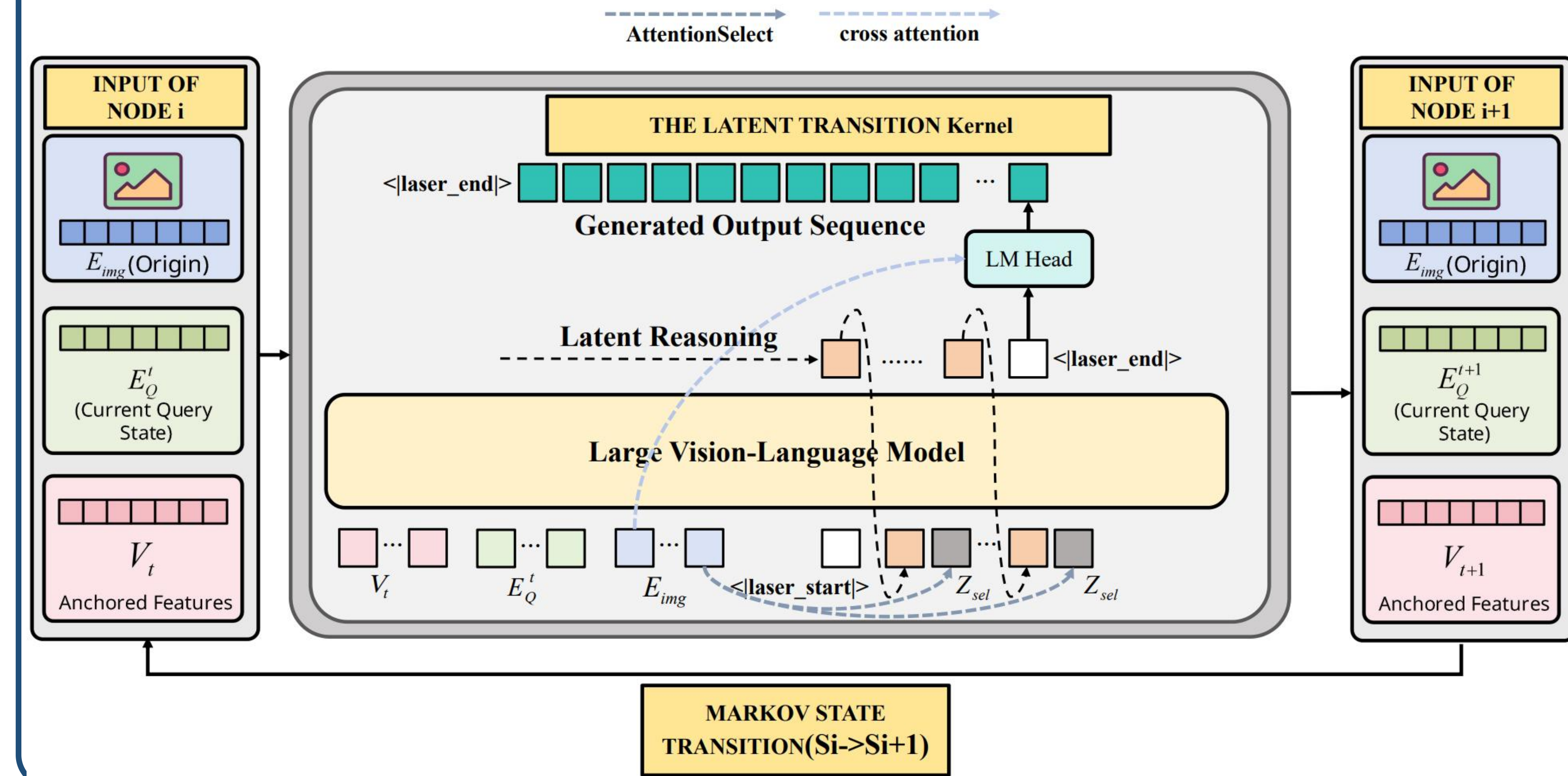
(B) Tool-Augmented (Think With Image)



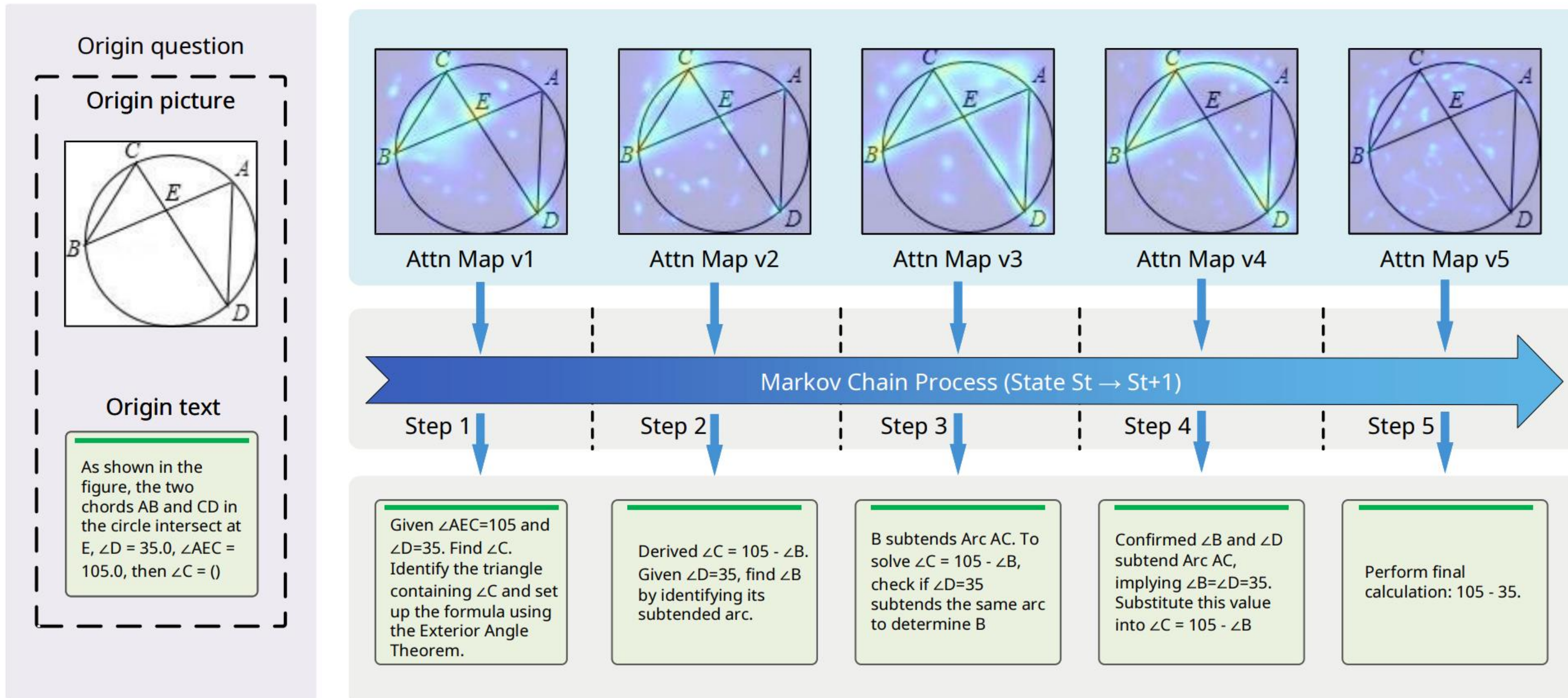
(C) Markov Latent Reasoning Chain



2. Method: LDPVR



3. Visualization of the process



4. Main Results

Method	Model	Visual Perception		Math Reasoning		Cross-modal	
		HalluBench	MMVP	MathVista	MM-Math	MMStar	ScienceQA
No-CoT	Qwen2.5 VL 3B	63.8	54.7	48.4	29.3	48.9	44.6
Multimodal COT		63.8	54.4	47.3	28.5	48.5	42.9
CCOT		64.0	55.5	48.0	30.2	49.3	44.5
SCAFFOLD		64.3	56.3	50.7	31.4	50.2	46.1
LDPVR		64.7 $\uparrow 0.9\%$	56.8 $\uparrow 2.1\%$	51.0 $\uparrow 2.6\%$	33.3 $\uparrow 4.0\%$	51.2 $\uparrow 2.3\%$	46.9 $\uparrow 2.3\%$
No-CoT	VLAA Thinking 7B	63.1	69.3	60.7	41.6	57.2	51.2
Multimodal COT		62.3	68.1	59.5	40.1	56.9	47.8
CCOT		64.6	68.0	60.5	41.8	59.0	49.4
SCAFFOLD		66.1	68.8	62.4	42.9	59.1	51.1
LDPVR		68.2 $\uparrow 5.1\%$	70.1 $\uparrow 0.8\%$	63.1 $\uparrow 2.4\%$	44.1 $\uparrow 2.5\%$	60.1 $\uparrow 2.9\%$	52.3 $\uparrow 1.1\%$
No-CoT	R1 OneVision 7B	62.6	67.4	51.8	41.1	52.6	51.2
Multimodal COT		62.2	68.4	52.7	39.1	51.1	51.2
CCOT		63.5	69.1	53.5	41.1	53.0	51.1
SCAFFOLD		64.1	70.1	55.9	42.1	54.6	55.3
LDPVR		65.2 $\uparrow 2.6\%$	72.4 $\uparrow 5.0\%$	58.7 $\uparrow 6.9\%$	43.7 $\uparrow 2.6\%$	56.8 $\uparrow 4.2\%$	56.1 $\uparrow 4.9\%$
No-CoT	Qwen3 VL 4B	71.8	71.1	64.9	65.9	57.5	51.1
Multimodal COT		70.3	69.2	62.1	64.2	57.3	52.8
CCOT		71.1	71.8	63.9	65.1	59.0	52.4
SCAFFOLD		72.8	72.9	64.9	65.4	59.7	54.1
LDPVR		73.1 $\uparrow 1.3\%$	72.9 $\uparrow 1.8\%$	66.1 $\uparrow 1.2\%$	66.2 $\uparrow 0.3\%$	60.8 $\uparrow 3.3\%$	55.2 $\uparrow 4.1\%$

