Dual Latent Memory for Visual Multi-agent System

Xinlei Yu, Chengming Xu, Zhangquan Chen, Bo Yin, Cheng Yang, Yongbo He, Yihao Hu, Jiangning Zhang, Cheng Tan, Xiaobin Hu, Shuicheng YAN

1 National University of Singapore, 2 Fudan University, 3 Tsinghua University, 4 DeepWisdom 5 Zhejiang University, 6 Shanghai AI Lab

ICML International Conference On Machine Learning

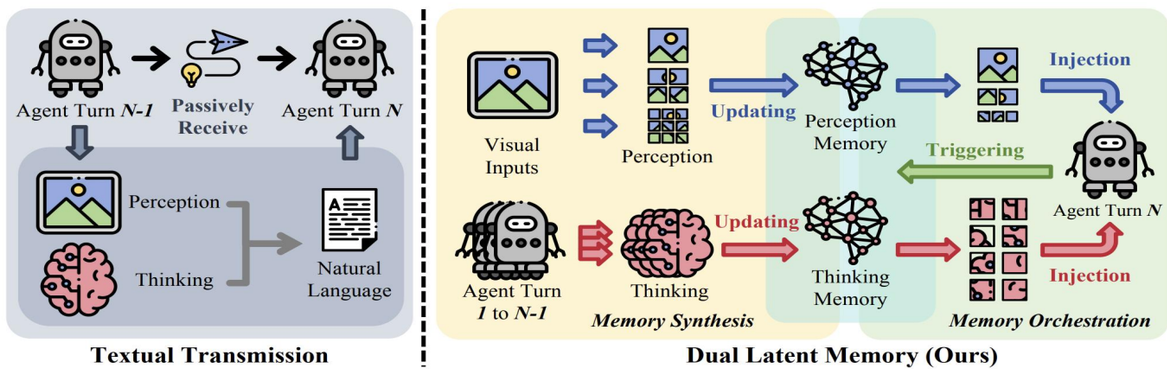
Wechat

Telegram

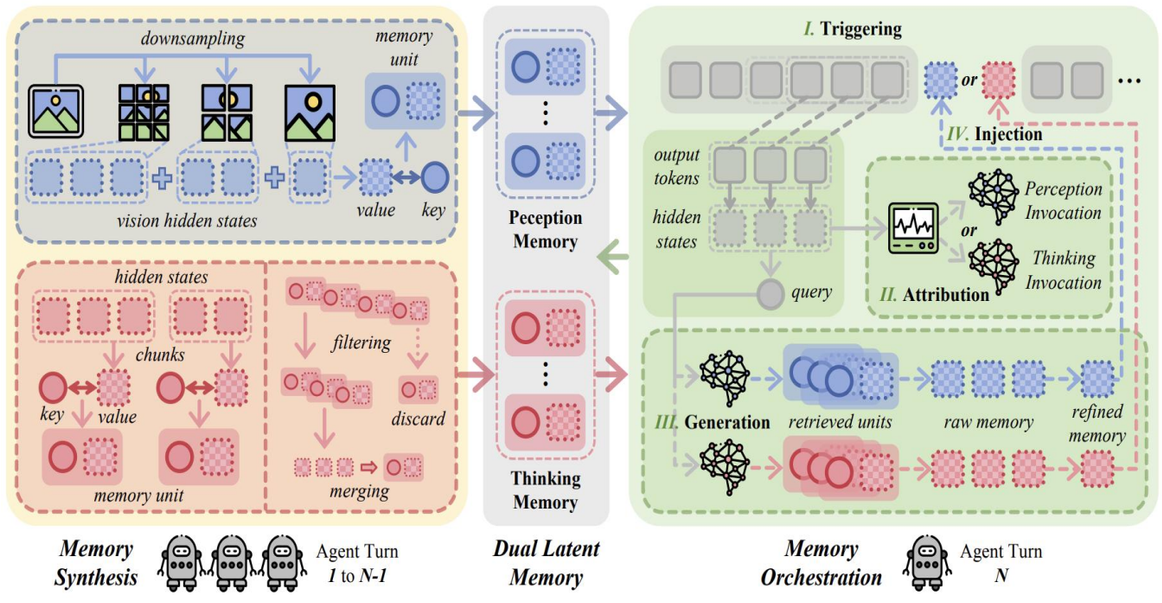
Contributions

※ Failure Analysis of VMAS
※ Dual Latent Memory Synthesis: perception and thinking memories
※ Proactive Memory Orchestration: triggering, attribution, generation, and injection
※ Results on models/sizes/structures with general performance improvements by **2.7-5.4%** and token consumption reduction by **21.3-44.8%**

Comparison



Methodology



Experiments

Table 1. Results of five base models based on dynamic structure, comparing both accuracy and total token usage. The best values are **bolded**, the rightmost column shows the average results, and the arrows indicate the improvement or reduction compared to VMAS.

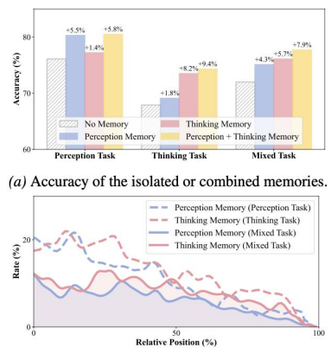
Backbone	Method	MMBench		MMStar		RealWorldQA		SimpleVQA		Average	
		Accuracy	*Token*	*Accuracy*	*Token*	*Accuracy*	*Token*	*Accuracy*	*Token*	*Accuracy*	*Token*
GLM-4.1V-9B-Thinking	Single	80.4	482	71.5	649	69.0	708	46.3	546	66.7	596
	VMAS	80.2	4416	71.8	6744	67.2	7446	47.5	5639	66.6	6061
	L^2-VMAS	**82.4** ↑2.2	**2487** ↓1929	**74.1** ↑2.3	**3451** ↓3293	**70.3** ↑3.1	**4314** ↓3132	**51.3** ↑3.8	**3125** ↓2514	**69.5** ↑4.3%	**3344** -44.8%
InternVL-3.5-8B	Single	79.5	416	69.0	454	67.4	523	41.1	452	64.3	3461
	VMAS	81.7	2732	72.5	2958	72.7	3360	45.0	2993	68.0	3011
	L^2-VMAS	**82.2** ↑0.5	**2024** ↓708	**74.7** ↑2.2	**2189** ↓769	**74.8** ↑2.1	**2534** ↓826	**47.6** ↑2.6	**2233** ↓760	**69.8** ↑2.7%	**3011** -25.4%
LLaVA-OV-1.5-8B	Single	82.6	340	67.0	388	68.3	445	44.6	389	65.6	391
	VMAS	84.7	2278	73.1	2552	72.8	2891	46.8	2567	69.4	2572
	L^2-VMAS	**84.4** ↓0.3	**1824** ↓454	**75.1** ↑2.0	**2067** ↓485	**76.2** ↑3.4	**2135** ↓756	**50.7** ↑3.9	**2069** ↓498	**71.6** ↑3.2%	**2024** -21.3%
Qwen3-VL-8B-Instruct	Single	84.0	318	70.4	363	71.0	402	50.1	358	68.9	360
	VMAS	84.9	2190	74.8	2467	74.7	2769	51.4	2492	71.4	2480
	L^2-VMAS	**87.4** ↑2.5	**1682** ↓508	**77.5** ↑2.7	**1907** ↓560	**77.6** ↑2.9	**2101** ↓668	**53.8** ↑2.4	**1865** ↓627	**74.1** ↑3.7%	**1889** -23.8%
Qwen3-VL-8B-Thinking	Single	84.8	557	75.5	687	72.9	730	49.4	604	70.9	645
	VMAS	85.3	5241	78.1	6670	76.2	7026	50.5	5679	72.5	6154
	L^2-VMAS	**88.8** ↑3.5	**2983** ↓2258	**81.4** ↑3.3	**3677** ↓2993	**80.2** ↑4.0	**3976** ↓3050	**55.3** ↑4.8	**3256** ↓2423	**76.5** ↑5.4%	**3473** -43.6%

Table 2. Results of various model size based on dynamic structure and Qwen3-VL family.

Backbone	Method	MMBench		MMStar		RealWorldQA		SimpleVQA		Average	
		Instruct	Thinking	Instruct	Thinking	Instruct	Thinking	Instruct	Thinking	Instruct	Thinking
Qwen3-VL-2B	Single	79.0	79.3	57.8	69.1	63.7	68.6	40.5	42.6	60.3	64.9
	VMAS	84.6	84.2	65.0	72.5	68.8	72.6	47.9	47.4	66.6	69.2
	L^2-VMAS	**84.2** ↓0.4	**85.8** ↑1.6	**66.8** ↑1.8	**73.6** ↑1.1	**68.6** ↓0.2	**73.7** ↑1.1	**50.0** ↑2.1	**49.6** ↑2.2	**67.4** ↑1.3%	**70.7** ↑2.2%
Qwen3-VL-4B	Single	83.4	84.1	69.4	74.3	70.5	72.5	48.2	48.0	67.9	69.7
	VMAS	85.7	85.4	73.9	75.3	75.9	75.3	50.4	49.5	71.5	71.4
	L^2-VMAS	**87.2** ↑1.5	**88.6** ↑3.2	**76.4** ↑2.5	**77.1** ↑1.8	**77.4** ↑1.5	**77.1** ↑1.8	**53.8** ↑3.4	**53.3** ↑3.8	**73.7** ↑3.1%	**74.0** ↑3.7%
Qwen3-VL-8B	Single	84.0	84.8	70.4	75.5	71.0	72.9	50.1	49.4	68.9	70.9
	VMAS	84.9	85.3	74.8	78.1	74.7	76.2	51.4	50.5	71.4	72.5
	L^2-VMAS	**87.4** ↑2.5	**88.8** ↑3.5	**77.5** ↑2.7	**81.4** ↑3.3	**77.6** ↑2.9	**80.2** ↑4.0	**53.8** ↑2.4	**55.3** ↑4.8	**74.1** ↑3.7%	**76.5** ↑5.4%
Qwen3-VL-32B	Single	87.2	89.0	76.8	79.7	79.1	78.4	56.7	55.3	75.0	75.6
	VMAS	89.1	87.3	77.8	77.6	78.0	77.6	56.8	57.8	75.4	75.1
	L^2-VMAS	**90.6** ↑1.5	**90.0** ↑2.7	**79.4** ↑1.6	**80.2** ↑2.6	**80.5** ↑2.5	**80.3** ↑2.7	**59.4** ↑2.6	**60.6** ↑2.8	**77.5** ↑2.7%	**77.8** ↑3.6%

Table 3. Results of various multi-agent structures based on Qwen3-VL-8B-Instruct/Thinking.

Structure	Method	MMBench		MMStar		RealWorldQA		SimpleVQA		Average	
		Instruct	Thinking	Instruct	Thinking	Instruct	Thinking	Instruct	Thinking	Instruct	Thinking
Linear	VMAS	83.1	82.8	72.6	74.2	72.9	73.0	48.9	49.5	69.3	69.9
	L^2-VMAS	**84.5** ↑1.4	**85.9** ↑3.1	**76.2** ↑3.6	**78.0** ↑3.8	**73.1** ↑0.2	**73.4** ↑0.4	**51.1** ↑2.2	**52.7** ↑3.2	**71.2** ↑2.7%	**72.5** ↑3.7%
Layered	VMAS	84.3	84.8	73.1	75.1	73.4	73.4	49.7	50.6	70.1	71.0
	L^2-VMAS	**85.7** ↑1.4	**87.4** ↑2.6	**76.4** ↑3.3	**78.6** ↑3.5	**74.8** ↑1.4	**74.9** ↑1.5	**51.6** ↑1.9	**53.8** ↑3.2	**72.1** ↑2.8%	**73.6** ↑3.8%
Centralized	VMAS	84.2	84.0	73.8	74.9	72.7	72.6	48.6	49.0	69.8	70.1
	L^2-VMAS	**86.5** ↑2.3	**87.3** ↑3.3	**76.9** ↑3.1	**79.4** ↑4.5	**76.3** ↑3.6	**77.2** ↑4.6	**52.2** ↑3.6	**54.3** ↑5.3	**73.0** ↑4.5%	**74.6** ↑6.3%
Random	VMAS	84.9	84.6	75.1	77.7	74.8	75.8	51.7	50.8	71.6	72.2
	L^2-VMAS	**86.9** ↑2.0	**87.6** ↑3.0	**77.8** ↑2.7	**81.3** ↑3.6	**77.3** ↑2.5	**80.0** ↑4.2	**53.5** ↑1.8	**54.8** ↑4.0	**73.9** ↑3.1%	**75.9** ↑5.1%
Complete	VMAS	84.8	83.9	75.3	77.4	75.3	75.0	52.0	50.6	71.9	71.7
	L^2-VMAS	**87.2** ↑2.4	**87.2** ↑3.3	**77.6** ↑2.3	**80.9** ↑3.5	**77.6** ↑2.3	**79.7** ↑4.7	**54.1** ↑2.1	**55.0** ↑4.4	**74.1** ↑3.1%	**75.7** ↑5.5%
Dynamic	VMAS	84.9	85.3	74.8	78.1	74.7	76.2	51.4	50.5	71.4	72.5
	L^2-VMAS	**87.4** ↑2.5	**88.8** ↑3.5	**77.5** ↑2.7	**81.4** ↑3.3	**77.6** ↑2.9	**80.2** ↑4.0	**53.8** ↑2.4	**55.3** ↑4.8	**74.1** ↑3.7%	**76.5** ↑5.4%



(a) Accuracy of the isolated or combined memories.

(b) Triggering rate of the two memories.

Figure 5. Effectiveness Analyses of the dual memory system based on perception, thinking, and mixed tasks.