VisMem: Latent Vision Memory Unlocks Potential of Vision-Language Models

Xinlei Yu[1] Chengming Xu[2] Guibin Zhang[1] Zhangquan Chen[3] Yudong Zhang[5]
Yongbo He[4] Peng-Tao Jiang[6] Jiangning Zhang[4] Xiaobin Hu[1*] Shuicheng Yan[1]
[1]National University of Singapore [2]Fudan University [3]Tsinghua University [4]Zhejiang University
[5]University of Science and Technology of China [6]vivo

Abstract

*Despite the remarkable success of Vision-Language Models (VLMs), their performance on a range of complex visual tasks is often hindered by a "visual processing bottleneck": a propensity to lose grounding in visual evidence and exhibit a deficit in contextualized visual experience during prolonged generation. Drawing inspiration from human cognitive memory theory, which distinguishes short-term visually-dominant memory and long-term semantically-dominant memory, we propose **VisMem**, a cognitively-aligned framework that equips VLMs with dynamic latent vision memories, a short-term module for fine-grained perceptual retention and a long-term module for abstract semantic consolidation. These memories are seamlessly invoked during inference, allowing VLMs to maintain both perceptual fidelity and semantic consistency across thinking and generation. Extensive experiments across diverse visual benchmarks for understanding, reasoning, and generation reveal that VisMem delivers a significant average performance boost of 11.8% relative to the vanilla model and outperforms all counterparts, establishing a new paradigm for latent-space memory enhancement. The code will be available: https://github.com/YU-deep/VisMem.git.*

1. Introduction

Visual-Language Models (VLMs) have demonstrated impressive capabilities in visual understanding, reasoning and generation [31, 49]. Latest flagship models, both closed-sourced [2, 11, 38] and open-sourced [1, 4, 54, 55, 62], represent a significant leap towards a general-purpose intelligent model that can both perceive and think about the visual world. Despite their success, VLMs still face significant inherent challenges when tackling complicated tasks that require advanced visual abilities, such as fine-grained perception, multi-step reasoning, or maintaining fidelity over long generative sequences [17, 25]. A fundamental limitation stems from the pervasive propensity, exhibited during deep autoregressive decoding, toward a deficit in visual memory, which prioritizes accumulated textual context over the initial visual evidence and lacks visual semantic knowledge [51, 89]. It manifests as a "visual processing bottleneck" that impairs performance in fine-grained visual understanding, efficient reasoning, and robust generation.

Prior efforts to overcome this limitation have explored several distinct strategic axes, which can be primarily categorized into four paradigms, as illustrated in Fig. 1. One intuitive paradigm is the **(a) direct training paradigm**, which optimizes model parameters via fine-tuning or reinforcement learning [26, 35, 43, 65]. This relatively brute-force approach often sacrifices generalization for task-specific performance, leading to catastrophic forgetting. Another axis concerns the representation space of the intervention, **(b) image-level paradigm**, operating in the pixel space by explicitly synthesizing new visual inputs, which offers image-level thinking but at a prohibitive computational cost [13, 24, 29, 47, 48, 86]. Conversely, **(c) token-level paradigm** constrains operations to visual tokens, which is more efficient but fundamentally non-generative, limiting the model to merely re-surfacing what it has already encoded [8, 16, 28, 74]. Recently, a promising direction lies in the **(d) latent space paradigm**, which introduces continuous latent contexts in the sequential inference process. Unfortunately, existing latent space methods either rely solely on the language space [21, 30, 46, 67, 80] or require auxiliary visual data [69], limiting their application in VLMs.

To overcome this problem, we resort to cognitive psychology, specifically the *Dennis Norris Theory* [37]:

> *Short-term memory and long-term memory are two distinct storage systems that can be modeled on their neural underpinnings, the former is governed by vision, while the latter holds sway over abstract semantics.*

While this cognitive theory reveals the essence of human cognition, it can be smoothly translated into an architectural principle of VLMs: short-term memory is visually-dominant, enhancing perception of the current

*Corresponding authors.

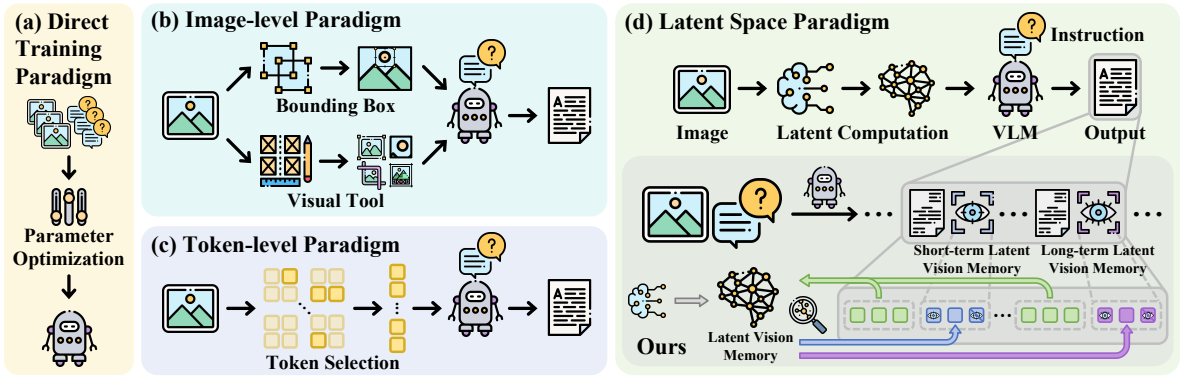


Figure 1. Four primary paradigms for enhancing visual capabilities: (a) the direct training paradigm, (b) the image-level paradigm, (c) the token-level paradigm, and (d) the latent space paradigm. Our VisMem belongs to the last one, featuring latent vision memory.

visual scenes, while long-term memory is semantically-dominant, providing generalized knowledge and contextualized semantic, completing the full cognitive chain.

Based on such inspiration, we propose VisMem, a novel and cognitively-aligned framework that systematically incorporates short- and long-term latent vision memory into VLMs. VisMem functions by non-intrusively extending the vocabulary of VLMs with special tokens that trigger on-demand latent vision memory invocation during autoregressive generation. Upon generating an invocation token, a lightweight query builder assesses the hidden states, which contains the current multi-modal cognition, to formulate a contextual-aware query which is then dispatched to one of two specialized, lightweight memory formers: short-term memory former that generates latent tokens encoding fine-grained, perceptual evidences of current visual inputs; long-term memory former that synthesizes tokens representing abstract, high-level semantic knowledge. These generated latent memory tokens are seamlessly inserted into the generation stream, enriching the contexts and enabling it to output with a seamless integration of detailed visual information and generalized semantic knowledge.

With a two-stage training paradigm based on reinforcement learning tailored for our proposed framework, the model learns to first generate effective memory contents, based on which the optimal patterns for invoking the memory is then learned. Our extensive experiments across a wide range of benchmarks spanning visual understanding, reasoning, and generation demonstrate that our approach can substantially enhance the comprehensive visual capabilities on various base models, while also improving cross-domain generalization and mitigating the problem of catastrophic forgetting. Our contributions are listed as follows:

- We propose a novel paradigm to proactively harness vision memory, alleviating the "visual processing bottleneck" and augmenting advanced visual capabilities.
- We propose a short- and long-term latent vision memory system with distinct purposes and mechanisms, which is analogous to the cognitive psychology.
- We propose a dynamic memory invocation mechanism for seamlessly invoking and inserting latent memory tokens into the autoregressive inference process.
- We evaluate the framework on extensive benchmarks, showcasing significant improvements in advanced visual capacities, cross-domain generalization, catastrophic forgetting mitigation, and compatibility across base models.

2. Related Work

2.1. Visual Capacities Enhancement

As demonstrated in Fig. 1, existing methods to alleviate "visual processing bottleneck" of VLMs broadly fall into four main categories: **(a) direct training paradigm**, which directly optimizes model parameters for target visual tasks, as in SFT, Visual-RFT [35], VLM-R1 [43], Vision-R1 [26], and PAPO [65]. Nonetheless, these methods suffer from catastrophic forgetting, specifically manifested as the degradation of general capabilities and over-specialization in specific visual cognition tasks [73, 88]; **(b) image-level paradigm**, which either leverages bounding boxes to denote visual evidence, represented by methods as Visual CoT [41], DeepEyes [86], SpatialVTS [33], VGR [57], and GRIT [13], or externally generate the iterative visual inputs via predefined tools, as seen in Sketchpad [24], VPRL [68], PyVision [84], OpenAI o3 [39], Pixel-Reasoner [47], MVoT [29], and OpenThinkImg [48]. Nevertheless, modifying visual inputs incurs extremely high computational costs, accompanied by high latency and reliance on external tools and concretized images; **(c) token-level paradigm**, which select original representations and cannot modify visual evidences, thus restricted by insufficiently refined information and suboptimal selection strategies, as in ICoT [16], MINT-CoT [8], SCAFFOLD [28], LLaVA-AURORA [6], VPT [74], Chameleon [53], **(d) latent space paradigm**, which employs latent states to optimize autoregressive generation, but its focus remains on pure language models, *e.g.*, Coconut [21], MemGen [80],

LatentSeek [30], SoftCoT [67], CODI [46]. Although Mirage [69] attempts to construct a latent vision space, requiring substantial manually labeled images. Our VisMem also belongs to this paradigm, but differs from existing methods by integrating latent vision memory within generation processes, characterized by a short and long memory system.

2.2. Memory Empowerment

Another mechanism closely tied to our approach involves endowing models with memory functionality. One intuitive strategy entails directly optimize models on prior trajectories, exemplified by [14, 44, 79], or to store them into the external memory repositories [52, 60]. Besides, some models inject persistently stored, retrieval-augmented knowledge from external environments, such as Expel [82] and MemoryBank [87], others, such as SkillWeaver [85] and Alita [40], distill prior knowledge as reusable tools. Currently, latent memory, as an implicit memory representation with better cross-domain generalization, efficiently encodes deep semantic associations, including M+ [64] and MemGen [80]. Nevertheless, these memory paradigms fail to ideally accommodate visual information, which manifests as a continuous, high-dimensional perceptual input. Consequently, the exploration of efficient visual memory mechanisms remains a largely uncharted territory. Thus, we propose a more human-aligned latent vision memory paradigm.

3. Methodology

3.1. Preliminary

Problem Formulation. Based on the interaction process of VLMs, we formulate the problem and introduce the notations used. We first define a policy model $\mathcal{P}$, which is powered by a base VLM. Given a visual task to be solved, feeding a instruction-vision pair (I, V) sampled from a task distribution $\mathcal{D}$, the policy model unfolds a corresponding trajectory τ at a timestep t, including pairs of current state s_t of the environment and the action a_t performed by the model. Here, the state of the environment includes textual contexts and visual observations. Internally, the action is generated sequentially by the token-by-token autoregressive decoding of the model, yielding the output token sequence $\{x_{t,1}, x_{t,2}, \ldots, x_{t,l}\}$. The generation of *i-th* output token $x_{t,i}$ could be presented as:

$$x_{t,i} \sim \mathcal{P}\left(\cdot \mid s_t, x_{<i}\right), \tag{1}$$

where the prediction is conditioned on the current environment state and previously generated tokens. To endow the model with vision memory, a vision memory system $\mathcal{M}$ is adhered to the policy model, thus, the objective is to optimize the memory-enhanced model jointly and to maximize its expected performance:

$$\max_{\mathcal{P},\mathcal{M}} \mathbb{E}_{(I,V)\sim\mathcal{D},\tau\sim(\mathcal{P},\mathcal{M})}[S(\tau)], \tag{2}$$

where $S(\cdot)$ denotes the quantifiable performance results, *e.g.*, accuracy or signal from a reward model.

Motivation. Building on the *Dennis Norris Theory* [37], which aligns with contemporary models of human memory, the coordinated operation of short- and long-term visual memories surmounts the "visual processing bottleneck". Short-term latent visual memory maintains fine-grained detail for immediate use and is thus visually dominant; by contrast, long-term latent visual memory abstracts across experiences to enable flexible reuse and is therefore semantically dominant. Taking the task illustrated in Fig. 2 as a case in point, "find the classic Lay's on the shelf" entails the deployment of short-term vision memory, retaining visual details for immediate perceptual demands, while "get in the promotion" triggers generalized semantic knowledge about the "promotion label" acquired from historical scenarios, which is grounded in long-term latent memory, to facilitate the comprehension of the task-based sight. Existing paradigms for enhancing visual capabilities fail to adequately consider vision memory, thus, our VisMem proposes a latent memory method to bridge this gap. More theoretical foundations are in Appendix 6.

Memory System. Based on previous contents, the task could be further disassembles into two main interactive parts: ***memory invocation*** (Sec. 3.2): related to "where and how to invoke the short- or long-term vision memory"; ***memory formation*** (Sec. 3.3): related to "what content should the short- or long-term vision memory convey". Additionally, these two decomposed processes interact closely with each other, with distinct priorities and objectives, requiring a meticulously designed ***training recipe*** (Sec. 3.4).

3.2. Memory Invocation

As illustrated in Fig. 2, our latent vision memory invocation strategy largely aligns with the standard generation pipeline of VLMs, thereby preserving their robust fundamental visual capabilities. Typically, VLMs generate rationales and answers; however, such pure text sequences lack the granularity to capture fine-grained visual perceptions and semantics, which poses challenges to accurate visual understanding, reasoning, and generation. This limitation arises because during inference, VLMs tend to prioritize accumulated textual context over visual evidence, a phenomenon particularly pronounced in long sequences [17, 25, 71, 77]. To address this, we extend the vocabulary $\mathcal{V}$ of VLMs by incorporating four additional memory-operation tokens, resulting in $\mathcal{V}' = \mathcal{V} \cup \{<m_I^s>, <m_E^s>, <m_I^l>, <m_E^l>\}$. Here, $<m_I>$ and $<m_E>$ form paired invocation and end tokens, where the superscripts s and l denote short- or long-term memory, respectively. Specifically, we register these as indivisible special tokens in the tokenizer and enlarge the embedding matrix from $\mathbb{R}^{|\mathcal{V}|\times d}$ to $\mathbb{R}^{(|\mathcal{V}|+4)\times d}$, where d is the dimension of the model. Furthermore, we initialize the em-

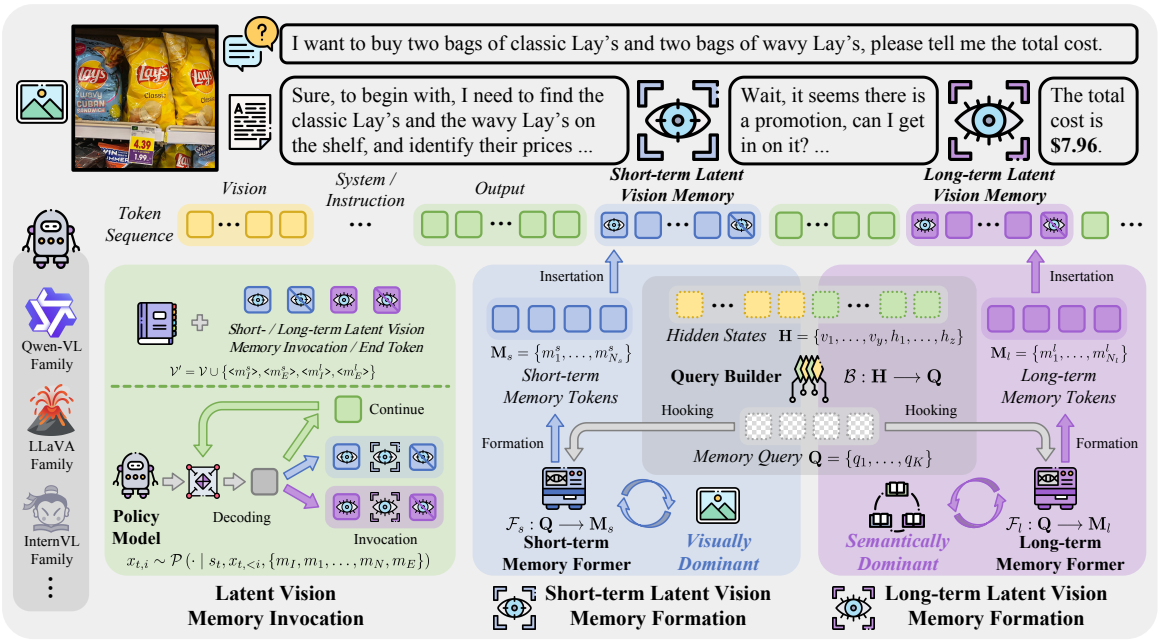


Figure 2. The overview of our proposed VisMem.

beddings of the invocation tokens (<m^s_I> and <m^l_I>) using the embedding vector of a delimiter token with small perturbations, and update these embeddings during training to facilitate faster convergence. The two end tokens (<m^s_E> and <m^l_E>) are treated as structural markers; they are initialized analogously with a lower learning rate. In practice, we also employ constrained decoding to encourage well-formed invocation-end pairs.

Specifically, the latent vision memory invocation tokens function as triggers for initiating memory insertion, based on the continuous internal cognitive states. During autoregressive generation (see Eq. (4)), upon the output of an invocation token, the memory former immediately initiates the latent vision memory formation procedure:

$$x_{t,i} \rightarrow \begin{cases} \text{invocation}, & x_{t,i} \in \{<m^s_I>, <m^l_I>\} \\ \text{continue}, & otherwise \end{cases}. \quad (3)$$

The resulting latent vision memory, whether short- or long-term as dictated by the specific token type, is subsequently inserted right after the already output invocation token. Following this insertion, the corresponding end token for short (<m^s_E>) or long memory (<m^l_E>) is automatically appended to resume token-by-token decoding:

$$x_{t,i} \sim \mathcal{P}\left(\cdot \mid s_t, x_{t,<i}, \{m_I, m_1, \ldots, m_N, m_E\}\right). \quad (4)$$

3.3. Memory Formation

To activate the vision memory capability of VLMs, we integrate two memory components: short-term vision memory, which encodes rich visual evidence, and long-term vision memory, which primarily encodes high-level, knowledge-based visual pertinent semantics, without modifying the core VLM and damaging general abilities. This integration leverages short-term memory to enhance advanced visual perception and comprehension, while long-term memory enables the generalization of semantic experiences during reasoning, thus comprehensively enhancing the overall visual performance. As illustrated in Fig. 2, the memory formation process hinges on two core components: a query builder $\mathcal{B}$, which is responsible for generating queries to hook memory; and memory formers $\mathcal{F}_s$ and $\mathcal{F}_l$, which are dedicated to constructing latent visual memories.

Query Builder. Through this process, we transform hidden states incorporating current cognition into a more efficient and accurate memory query. Initially, we instantiate a lightweight transformer encoder denoted as $\mathcal{B}$ and a learnable memory query $\mathbf{Q}_{init} = \{q_1, \ldots, q_K\}$, where K represents the length of the query sequence and each $q \in \mathbb{R}^d$. Given the state at a particular time, $\mathcal{B}$ encodes the query sequence based on internal visual and contextual hidden states to retrieve the corresponding latent memory contents. During each invocation, as the policy model generates the current output token sequence, *i.e.*, the token sequence starting from the initial position or from the end of the previous invocation, it accordingly produces a sequence of hidden state vectors $\{h_1, \ldots, h_z\}$. Similarly, visual encoder produces visual hidden state vectors $\{v_1, \ldots, v_y\}$. Thus, the combination of them $\mathbf{H} = \{v_1, \ldots, v_y, h_1, \ldots, h_z\} \in \mathbb{R}^{(y+z)\times d}$, characterizing the multi-modal cognitive state at the time, where y and z denote the lengths. Subsequently, we concatenate the initialized memory query to the rear of these hidden states to update the queried semantic information:

$$\mathbf{Q} = \mathcal{B}\left([\mathbf{H}, \mathbf{Q}_{init}]\right)[-K:], \quad (5)$$

where we select the output of the last layer of the encoder (see Eq. (10)), and take the last K encoded vectors as the memory query $\mathbf{Q} \in \mathbb{R}^{K\times d}$ to hook latent memory. Furthermore, we employ a masked attention to exclusively enable attention propagation from the query to the hidden states $\mathbf{H}$, while suppressing attention in the reverse direction, *i.e.*, from $\mathbf{H}$ to $\mathbf{Q}$ (see Eq. (11)). Here, both short- and long-term memory share the same query builder $\mathcal{B}$.

Latent Memory Former. Distinct from many existing paradigms [26, 43, 69], we internalize the latent vision memory into lightweight formers, preserving the general abilities of base VLMs and ensuring the compatibility of our paradigm. We initialize two lightweight LoRA adapters, which are respectively designated as the short-term memory former $\mathcal{F}_s$ and long-term memory former $\mathcal{F}_l$, attached to the vision encoder and the final language model of the VLM, without directly tampering with the core parameters. More precisely, we first append the generated memory query $\mathbf{Q}$ along with a set of learnable memory tokens after the corresponding target token sequence $\mathbf{X}$. Then we process it by short-term or long-term memory former, which contextualizes and embeds the latent memory information:

$$\mathbf{M}_{s/l} = \mathcal{F}_{s/l}\left([\mathbf{X}, \mathbf{Q}, \mathbf{M}_{init}]\right)\left[-N_{s/l}:\right], \tag{6}$$

where short- and long-term latent vision memory $\mathbf{M}_{s/l} \in \mathbb{R}^{N_{s/l}\times d}$, while N_s and N_l are the predetermined lengths of memory tokens, which can be taken from $\{2, 4, 8, 16, 32\}$. For the short-term pathway, the resultant memory representation is concatenated with the visual token stream, and pass through the original projector to align it with the representation space of the language model. The two memory formers serve as dedicated memory carriers, exclusively storing visual evidences and semantic knowledge within themselves. When the policy model executes a memory invocation, the incoming memory query triggers externalization of useful short- or long-term memory. These memories are seamlessly inserted into the token generation process alongside the invocation and end signals and barely interfere with the original generation, as specified in Eq. (4).

3.4. Training Recipe

We design a two-stage training procedure based on GRPO [42], whose optimization objectives are to optimize the effective formation and invocation of latent memory. The first stage enhances the utility of memory, while the second stage maximizes the reward of each invocation, thereby accelerating the convergence of different components steadily. More detailed algorithms and implementations are present in Appendix 7.2 and 8.3.

Stage I: Memory Formation Optimization. In this stage, we update the query builder $\mathcal{B}$, and memory formers $\mathcal{F}_{s/l}$ while keeping the policy model $\mathcal{P}$ frozen. Initially, during the autoregressive generation process, we randomly invoke either short- or long-term memory upon detecting the delimiter, thereby acquiring initial memory capabilities. Then, the scope of memory invocations is extended to the intervals between delimiters, this not only provides a richer trajectory of memory interactions but also enables memory invocation at arbitrary positions within the generation sequence. The core objective is to maximize the performance improvement relative to trajectory without memory integration $\Delta S(\tau) = S(\tau) - S(\tau_{base})$, thereby enhancing the quality of the memory formation (full function in Eq. (14)):

$$\max_{\mathcal{F}_{s/l},\mathcal{B}} \mathbb{E}_{\tau\sim\mathcal{P}\left(\cdot|x,\mathbf{M}_{s/l}\right),\mathbf{M}_{s/l}\sim\mathcal{F}_{s/l}(\mathbf{Q}),\mathbf{Q}\sim\mathcal{B}(\mathbf{H})}[\Delta S(\tau)]. \tag{7}$$

Stage II: Memory Invocation Optimization. In this process, we update part parameters θ of the policy model $\mathcal{P}$, and keeps all the memory formation components frozen. At this stage, the policy model $\mathcal{P}$ is required to invoke memory efficiently and accurately, which entails two core requirements: selecting the correct memory type and avoiding invalid invocations. Thus, we add two penalties to the objective, which could be optimized by (full function in Eq. (15)):

$$\max_{\theta} \mathbb{E}_{\tau\sim\mathcal{P}\left(\cdot|x,\mathbf{M}_{s/l}\right)}[\Delta S(\tau) - \alpha(p_{type} + p_{neg})], \tag{8}$$

where α denotes the penalty intensity. The type penalty, $p_{\text{type}} = \max\left(0, S(\tau_{\text{rev}}) - S(\tau)\right)$, serves to penalize the erroneous selection of memory types, where τ_{rev} represents the invocation of an alternative memory type. In parallel, the negative penalty $p_{\text{neg}} = \max\left(0, \overline{S} - S(\tau)\right)$ is designed to penalize invocations with negative returns, aiming to enhance efficiency. Here, $\overline{S}$ denotes the mean of quantifiable scores across candidate trajectories.

4. Experiments

4.1. Settings

Benchmarks. We select 12 benchmarks to comprehensively evaluate three main abilities of VLMs, *i.e.*, understanding, reasoning and generation [31]. These benchmarks include: **(1) Understanding:** MMStar [7], MMVet [75], MMT [72], BLINK [15], MuirBench [56]; **(2) Reasoning:** MMMU [78], LogicVista [66], Math-V [58], MV-Math [61]; **(3) Generation:** HallBench [19], MultiTrust [81], MMVU [34]. Details are in Appendix 8.2.

Baselines. We compare our VisMem against 15 baselines, falling into four categories: **(a) direct training methods**: SFT, Visual-RFT [35], VLM-R1 [43], Vision-R1 [26] and PAPO [65]; **(b) image-level methods**: GRIT [13], Sketchpad [24], MVoT [29], OpenThinkImg [48] and DeepEyes [86]; **(c) token-level methods**: Scaffold [28], MINT-CoT [8], ICoT [16], and VPT [74]; **(d) latent space methods**: Mirage [69]. Details are in Appendix 8.3.

Implementation Details. All experiments (except for Tab. 2) are implemented on Qwen2.5-VL-7B [4] based on 8

Table 1. Results on 12 benchmarks to evaluate visual understanding, reasoning and generation abilities. The **best** and second best values are emphasized, and the average values are calculated for both specific capabilities and overall results.

Method	MM Star	MM Vet	MMT	BLINK	Muir Bench	*Avg.*	MMMU	Logic Vista	Math -V	MV -Math	*Avg.*	Hall Bench	Multi Trust	MMVU	*Avg.*	*Avg.*
Vanilla [4]	62.6	66.0	54.0	55.4	57.4	*59.3*	56.0	43.5	23.8	18.9	*35.6*	52.3	64.8	55.4	*57.7*	*51.0*
SFT	64.7	67.5	56.8	54.5	58.7	*60.3*	57.7	46.1	25.9	22.8	*38.1*	53.6	67.0	59.1	*59.9*	*52.8*
Visual-RFT [35]	65.6	70.5	59.1	58.0	62.9	*63.6*	62.4	51.7	33.4	26.5	*45.2*	55.8	70.7	63.2	*63.2*	*57.4*
VLM-R1 [43]	66.3	73.0	59.4	60.6	63.8	*64.6*	63.4	53.0	39.5	34.6	*47.6*	54.2	69.9	61.7	*61.9*	*58.3*
Vision-R1 [26]	67.1	71.7	60.2	60.8	64.0	*65.0*	63.2	53.9	42.5	38.7	*49.8*	56.4	72.6	63.6	*64.2*	*59.7*
PAPO [65]	64.2	69.8	57.9	53.3	56.7	*60.4*	61.2	52.5	34.1	34.8	*45.7*	50.3	67.7	56.5	*58.2*	*55.0*
Sketchpad [24]	62.1	64.5	57.0	54.9	52.8	*58.3*	57.9	47.4	27.5	24.6	*39.4*	52.1	66.2	57.2	*58.5*	*52.1*
GRIT [13]	65.8	67.8	57.9	52.5	51.0	*59.0*	59.4	51.6	30.8	22.4	*43.6*	53.7	67.3	60.1	*60.4*	*54.2*
PixelReasoner [47]	65.3	67.1	58.7	56.8	60.5	*61.7*	58.9	49.3	32.0	25.9	*43.0*	55.9	69.9	61.5	*62.4*	*55.6*
DeepEyes [86]	66.4	70.5	60.3	60.4	63.0	*64.1*	60.3	49.1	34.6	31.5	*43.9*	57.4	72.6	64.6	*64.9*	*57.6*
OpenThinkImg [48]	66.0	71.6	60.8	59.2	61.7	*63.9*	61.4	52.8	37.8	28.0	*45.8*	54.9	74.0	64.3	*64.4*	*58.0*
Scaffold [28]	63.9	67.0	58.5	52.5	52.9	*59.0*	58.1	51.0	29.3	21.0	*41.9*	54.8	68.5	60.6	*61.3*	*53.9*
ICoT [16]	65.6	67.9	60.5	54.3	57.0	*61.1*	58.6	49.8	42.9	30.8	*46.0*	57.0	69.1	62.0	*62.7*	*56.5*
MINT-CoT [8]	66.2	69.5	57.3	55.4	58.9	*61.5*	57.7	51.5	**47.3**	39.2	*48.9*	56.7	71.4	60.8	*63.0*	*57.7*
VPT [74]	64.2	70.8	59.0	58.6	63.5	*63.2*	59.1	53.0	37.6	34.7	*46.1*	52.3	64.7	61.4	*59.5*	*56.6*
Mirage [69]	64.5	71.8	56.1	56.3	59.0	*61.5*	59.4	50.6	36.1	35.4	*45.4*	50.9	66.1	60.3	*59.1*	*55.5*
VisMem (Ours)	**68.9**	**75.1**	**62.5**	**64.5**	**69.8**	***68.2***	**63.9**	**55.7**	46.9	**41.4**	***52.0***	**59.6**	**77.0**	**68.2**	***68.3***	***62.8***

NVIDIA H200 141G GPUs. The length of memory query K is set to 8, and the lengths of short-term N_s and long-term latent vision memory N_l are 8 and 16, respectively. More implementation details are listed in Appendix 8.4.

4.2. Main Results

The main experimental results demonstrate that our proposed memory system VisMem unlocks the untapped potentials with three key ***enh***ancements: ***[Enh.1]*** advanced visual capabilities, ***[Enh.2]*** cross-domain generalization, ***[Enh.3]*** catastrophic forgetting alleviation.

***[Enh.1]* VisMem enables advanced and comprehensive visual capabilities.** As presented in Tab. 1, our proposed method demonstrates distinct superiority over other baseline models. Compared with the vanilla model, VisMem achieves a notable average improvement of **11.8%** across all benchmarks. When compared with the top three baselines (*i.e.*, Vision-R1 [26], VLM-R1 [43], and OpenThinkImg [48]), our method still maintains improvements of 3.1%, 4.5%, and 4.8%, respectively. Furthermore, it consistently enhances performance across the three core domains of visual tasks, namely, understanding, reasoning, and generation. Our latent vision memory mechanism yields comprehensive enhancements in visual capabilities, with specific gains of +8.9% in visual understanding, +16.4% in reasoning, and +10.6% in generation, relative to the vanilla model. It is also noteworthy that direct RL-based methods (*e.g.*, VLM-R1 [43] and Vision-R1 [26]) also achieve relatively better performance than most other paradigms. However, this approach of directly modifying parameters relies on incremental parameter updates, which may lead to the overwriting of prior general knowledge and result in catastrophic forgetting.

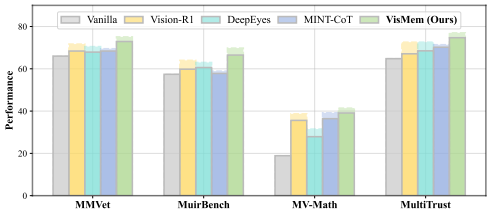


Figure 3. Results of the cross-domain generalization study. Models are only trained on Visual CoT [41] and Mulberry [70]. Dashed bar indicates the results with full training data.

As illustrated in Tab. 5 and 6, we conduct additional evaluations on selected subsets of MuirBench [56] and LogicVista [66]. Endowed with short- and long-term vision memory, our VisMem outperforms all baseline methods by a substantial margin in tasks demanding fine-grained visual evidence, such as counting (+7.0%), visual retrieval (+9.4%), and grounding (13.1%), while also yielding notable improvements in visual reasoning tasks, including inductive (+5.7%) and deductive (+7.1%) learning.

***[Enh.2]* VisMem showcases great cross-domain generalization.** To evaluate the cross-domain generalization capability of our model, specifically whether its stored latent visual memory can transfer across diverse unseen tasks, we exclusively train our VisMem and comparative baseline models on two datasets: Visual CoT [41] and Mulberry [70], then subsequently assess their performance on four unseen target benchmarks. As demonstrated in Fig. 3, 7, and Tab. 7, VisMem not only consistently achieves significant performance gains on out-of-domain tasks (+6.9% on MMVet [75], +9.1% on MuirBench [56], +20.2% on MV-Math [61], and +9.9% on MultiTrust [81]), but also main-

Table 2. Results on nine base models with various sizes and sources, including Qwen2.5-VL-3B/7B/32B [4], LLaVA-OV-1.5-4B/8B [1], InternVL-3.5-4B/8B/14B/38B [62]. ↑ indicates the performance enhancement compared with the base model.

Base Model	MM Star	MM Vet	MMT	BLINK	Muir Bench	MMMU	Logic Vista	Math -V	MV -Math	Hall Bench	Multi Trust	MMVU
Qwen2.5-VL-3B [4]	52.9	61.5	49.8	46.0	46.1	52.6	39.7	19.8	13.2	46.3	56.9	48.4
+ VisMem (Ours)	61.0 ↑8.1	72.5 ↑11.0	59.3 ↑9.5	58.6 ↑12.6	64.4 ↑18.3	61.9 ↑9.3	53.1 ↑13.4	38.7 ↑18.9	31.7 ↑18.5	58.0 ↑11.7	70.3 ↑13.4	60.6 ↑12.2
Qwen2.5-VL-7B [4]	62.6	66.0	54.0	55.4	57.4	56.0	43.5	23.8	18.9	52.3	64.8	55.4
+ VisMem (Ours)	68.9 ↑6.3	75.1 ↑9.1	62.5 ↑8.5	64.5 ↑9.1	69.8 ↑11.4	63.9 ↑7.9	55.7 ↑12.2	46.9 ↑23.1	41.4 ↑22.5	59.6 ↑7.3	77.0 ↑12.2	68.2 ↑12.8
Qwen2.5-VL-32B [4]	67.1	68.7	64.7	59.9	63.5	70.6	47.9	35.5	29.0	53.6	64.5	55.5
+ VisMem (Ours)	73.9 ↑6.8	77.9 ↑9.2	72.0 ↑7.3	68.6 ↑8.7	73.3 ↑9.8	75.9 ↑5.3	63.5 ↑15.6	65.0 ↑29.5	54.9 ↑25.9	60.2 ↑6.6	77.7 ↑13.2	68.4 ↑12.9
LLaVA-OV-1.5-4B [1]	62.5	60.4	54.4	38.2	42.6	49.4	39.3	20.1	11.0	41.8	47.5	44.2
+ VisMem (Ours)	69.0 ↑6.5	70.1 ↑9.7	62.7 ↑8.3	56.9 ↑18.7	59.6 ↑17.0	59.7 ↑10.3	53.7 ↑14.4	36.2 ↑16.1	27.2 ↑16.2	52.8 ↑11.0	66.4 ↑18.9	61.9 ↑17.7
LLaVA-OV-1.5-8B [1]	65.3	67.1	57.8	49.8	50.5	55.3	46.5	24.9	15.7	50.1	54.7	50.6
+ VisMem (Ours)	70.8 ↑5.5	75.7 ↑8.6	64.7 ↑6.9	61.0 ↑11.2	62.6 ↑12.1	63.0 ↑7.7	59.5 ↑13.0	45.1 ↑20.2	34.5 ↑18.8	55.5 ↑5.4	69.4 ↑14.7	67.0 ↑16.4
InternVL-3.5-4B [62]	62.8	73.1	62.7	57.1	52.8	59.9	53.2	51.8	17.5	43.0	56.2	44.7
+ VisMem (Ours)	70.2 ↑7.4	80.3 ↑7.2	69.0 ↑6.3	65.2 ↑8.1	63.9 ↑11.1	68.5 ↑8.6	64.6 ↑11.4	63.2 ↑11.4	30.8 ↑13.3	52.4 ↑9.4	70.4 ↑14.2	61.9 ↑17.2
InternVL-3.5-8B [62]	67.0	80.1	64.6	58.4	55.7	68.6	54.8	54.0	27.1	54.5	65.9	52.3
+ VisMem (Ours)	71.8 ↑4.8	85.4 ↑5.3	69.5 ↑4.9	66.1 ↑7.7	65.3 ↑9.6	73.3 ↑4.7	64.7 ↑9.9	66.4 ↑12.4	44.7 ↑17.6	60.9 ↑6.4	78.5 ↑12.6	68.3 ↑16.0
InternVL-3.5-14B [62]	67.3	79.0	66.1	56.9	57.7	68.8	55.9	56.5	29.4	54.1	69.6	54.8
+ VisMem (Ours)	72.4 ↑5.1	85.7 ↑6.7	70.6 ↑4.5	66.1 ↑9.2	67.0 ↑9.3	73.8 ↑5.0	65.5 ↑9.6	67.4 ↑10.9	46.6 ↑17.2	60.5 ↑6.4	77.8 ↑8.2	68.3 ↑13.5
InternVL-3.5-38B [62]	72.2	79.7	70.5	60.3	64.0	72.1	61.1	60.3	35.7	60.2	71.5	58.0
+ VisMem (Ours)	75.1 ↑2.9	86.4 ↑6.7	73.3 ↑2.8	67.5 ↑7.2	69.9 ↑5.9	75.8 ↑3.7	68.7 ↑7.6	70.2 ↑9.9	56.9 ↑21.2	65.8 ↑5.6	79.0 ↑7.5	69.9 ↑11.9

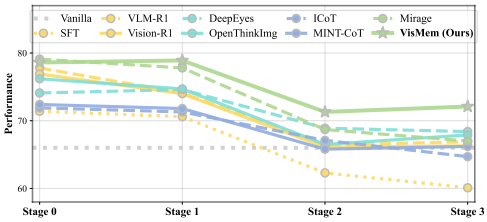


Figure 4. Results of four-stage continual learning on MMVet [75]. Stage 0 only includes itself, while stage 1, 2, 3 sequentially train models on different additional training data combinations.

tains leading performance relative to all baselines. Notably, our method outperforms the second-ranked model by a substantial margin of 2.7–6.8% across all four benchmarks, while narrowing the performance gap relative to results obtained with full training data. This observation underscores its robust cross-domain knowledge transfer capability.

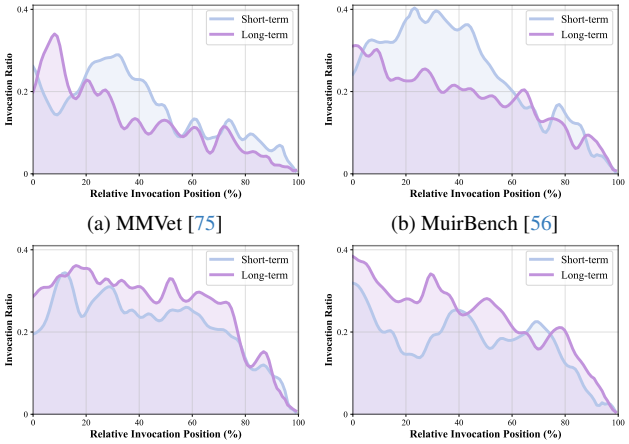
***[Enh.3]* VisMem alleviates catastrophic forgetting.** As illustrated in Fig. 4, 8, and Tab. 8, we conduct sequential training of the models across four stages, with performance assessed on MMVet [75] after each stage. At stage 0, the model was trained exclusively on the base task, and in subsequent stages, we incrementally incorporated selected benchmarks into the training process. From the continual learning results, our VisMem demonstrates significantly stronger knowledge retention capabilities. Although direct training paradigms yield relatively excellent overall performance in offline learning tasks with once-off training, they suffer from severe catastrophic forgetting. For instance, SFT exhibits over 10% performance degradation throughout the training process, the highest among all baselines. Additionally, at stage 0, VLM-R1 [43] and Vision-R1 [35] achieve performance improvements of 11.8% and 10.9% respectively compared to the vanilla model, however, these improvements are retained by less than 0.5% at stage 4. In contrast, our method effectively mitigates catastrophic forgetting, exhibiting the smallest performance gap relative to original full-data training among all baselines. It is further worth noting that our latent vision memory enhances performance at stages 1 and 3 without any degradation, reflecting superior cross-task generalization.

4.3. Additional Analyses

Through additional analyses, we derive three key research ***obs***ervations pertaining to VisMem: ***[Obs.1]*** compatibility across base models, ***[Obs.2]*** dynamic and adaptive memory invocation, ***[Obs.3]*** relatively low inference latency.

***[Obs.1]* VisMem is robustly compatible across various base models.** As detailed in Tab. 2 and Fig. 11, to evaluate the generalizability of our approach across diverse base models, we assess nine widely used base models, encompassing Qwen2.5-VL-3B/32B [4], LLaVA-OV-1.5-4B/8B [1], InternVL-3.5-4B/8B/14B/38B [62], with parameter scales ranging from 3B to 38B. The results indicate that our latent vision memory paradigm exhibits strong compatibility across various models, yielding significant performance improvements across most visual tasks.

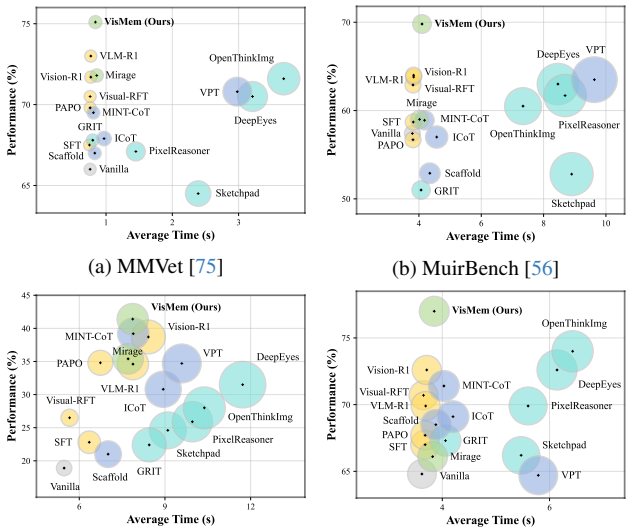
***[Obs.2]* The memory invocations are dynamic and self-adaptive.** To elaborate on the effectiveness of our dual latent memory system, we characterize the properties of the short- and long-term memories it forms. As illustrated in Fig. 5, we first analyze the type-specific invocation ratios and their relative positions within the output sequence across four benchmarks. In summary, invocation ratios are self-adaptive across tasks, while both memory types exhibit a dynamic downward trend in invocation frequency throughout the output sequence. Task-specific comparisons



(a) MMVet [75] (b) MuirBench [56]

(c) MV-Math [61] (d) MultiTrust [81]

Figure 5. Results of memory invocation ratio and invocation relative position across four benchmarks.



(a) MMVet [75] (b) MuirBench [56]

(c) MV-Math [61] (d) MultiTrust [81]

Figure 6. Results of average inference time and performance across four benchmarks. The size is proportional to its y-value.

in Fig. 9 further reveal that short-term latent memories are invoked more frequently to retrieve fine-grained details during visual information acquisition and understanding, particularly in multi-image scenarios, such as MuirBench [56]. Conversely, long-term latent vision memories play a more critical role in reasoning, *e.g.*, in MV-Math [61], by providing abstract semantic knowledge relevant to the current task. Furthermore, Tab. 5 and 6, which detail the sub-task performance of MuirBench [56] and LogicVista [66] respectively, further illustrate that short-term and long-term latent visual memories are complementary. Their dynamic invocation yields superior performance compared to relying on a single memory type or the absence of vision memory.

***[Obs.3]* VisMem incurs minimal inference latency while yielding substantial performance gains.** As showcased in Fig. 6 and Tab. 12, we compare the average inference time and task performance on four benchmarks to quantify the efficiency-performance trade-off of our method. Our VisMem, by harnessing the capabilities of dual vision memory, attains the best performance while incurring insignificant inference latency. Notably, image-level paradigms significantly elevate inference latency, particularly for tasks involving long thinking paths. In contrast, our VisMem exhibits remarkable effectiveness while maintaining average inference latency comparable to that of direct training optimization and token-level methods.

Ablation Study and Sensitivity Analysis. As reported in Tab. 3, we conduct ablative studies on the memory invocation and dual memory formation. The results reveal that both short-term and long-term memory components contribute to performance across diverse visual tasks, while their complementarity synergistically drives the optimal performance. Additionally, as detailed in Tab. 9, our design achieves a favorable balance between effectiveness and efficiency, with accurate and non-redundant memory invocation. As shown in Fig. 10 and Tab. 10, 11, we conduct sensitivity analyses of the sequence lengths of the memory query K, short-term N_s and long-term N_l latent memory tokens. As observed, performance generally improves with increasing sequence lengths within a reasonable range. Notably, our selected hyper-parameters achieve a favorable balance between performance and computational efficiency.

Table 3. Ablations of latent vision memory invocation and dual latent vision memory formation.

Ablation	MMVet	MuirBench	MV-Math	MultiTrust
Vanilla	66.0	57.4	18.9	64.8
Random Invocation (25%)	69.2	59.4	29.8	69.4
Random Invocation (50%)	71.9	63.2	26.1	68.5
Random Invocation (75%)	73.6	62.7	21.9	63.7
Full Invocation (100%)	73.4	56.0	17.5	62.6
Short-term Memory	71.5	65.6	29.6	73.6
Long-term Memory	69.4	60.2	36.1	69.8
Complete VisMem (Ours)	**75.1**	**69.8**	**41.4**	**77.0**

5. Conclusion

To address "visual processing bottleneck" of VLMs that impairs advanced visual capacities, we propose VisMem in this work, a cognitively inspired framework embedding dynamic latent vision memory, which integrates dual specialized memory formers guided by human patterns, with a non-intrusive memory invocation mechanism. Extensive experiments validate VisMem achieves an obvious performance improvement across various benchmarks, and exhibits strong cross-domain generalization, catastrophic forgetting mitigation, compatibility, and efficient inference, unlocking comprehensive and advanced visual potentials.

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 1, 7, 4

[2] Anthropic. Introducing claude haiku 4.5, 2025. 1

[3] Alan Baddeley. Working memory: Theories, models, and controversies. *Annual review of psychology*, 63(1):1–29, 2012. 1

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 5, 6, 7, 2, 3, 4, 8

[5] Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL: Long Papers)*, pages 15230–15250, Vienna, Austria, 2025. Association for Computational Linguistics.

[6] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 3836–3845, 2025. 2

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems (NeurIPS)*, 37:27056–27087, 2024. 5, 2, 6

[8] Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025. 1, 2, 5, 6, 4, 8

[9] Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *arXiv preprint arXiv:2505.16782*, 2025.

[10] Zhangquan Chen, Ruihui Zhao, Chuwei Luo, Mingze Sun, Xinlei Yu, Yangyang Kang, and Ruqi Huang. Sifthinker: Spatially-aware image focus for visual reasoning. *arXiv preprint arXiv:2508.06259*, 2025.

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[12] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 9062–9072, 2025.

[13] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025. 1, 2, 5, 6, 8

[14] Dayuan Fu, Keqing He, Yejie Wang, Wentao Hong, Zhuoma GongQue, Weihao Zeng, Wei Wang, Jingang Wang, Xunliang Cai, and Weiran Xu. Agentrefine: Enhancing agent generalization through refinement tuning. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[15] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, pages 148–166. Springer, 2024. 5, 2, 6

[16] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19520–19529, 2025. 1, 2, 5, 6, 8

[17] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024. 1, 3

[18] Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025.

[19] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, 2024. 5, 2, 6

[20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[21] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024. 1, 2

[22] Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. Multimodal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 18180–18187, 2024.

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[24] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:139348–139379, 2024. 1, 2, 5, 6, 8

[25] Jen-Tse Huang, Dasen Dai, Jen-Yuan Huang, Youliang Yuan, Xiaoyuan Liu, Wenxuan Wang, Wenxiang Jiao, Pinjia He, and Zhaopeng Tu. Visfactor: Benchmarking fundamental visual cognition in multimodal large language models. *arXiv preprint arXiv:2502.16435*, 2025. 1, 3

[26] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 1, 2, 5, 6, 3, 4, 8

[27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[28] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024. 1, 2, 5, 6, 8

[29] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *International Conference on Machine Learning (ICML)*, 2025. 1, 2, 5

[30] Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, et al. Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space. *arXiv preprint arXiv:2505.13308*, 2025. 1, 3

[31] Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*, 2024. 1, 5

[32] Zixu Li, Zhiheng Fu, Yupeng Hu, Zhiwei Chen, Haokun Wen, and Liqiang Nie. Finecir: Explicit parsing of fine-grained modification semantics for composed image retrieval. *arXiv preprint arXiv:2503.21309*, 2025.

[33] Xun Liang, Xin Guo, Zhongming Jin, Weihang Pan, Penghui Shang, Deng Cai, Binbin Lin, and Jieping Ye. Enhancing spatial reasoning through visual and textual thinking. *arXiv preprint arXiv:2507.20529*, 2025. 2

[34] Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, et al. Seeing clearly, answering incorrectly: A multimodal robustness benchmark for evaluating mllms on leading questions. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 9087–9097, 2025. 5, 2, 6

[35] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 1, 2, 5, 6, 7, 8

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[37] Dennis Norris. Short-term memory and long-term memory are still different. *Psychological bulletin*, 143(9):992, 2017. 1, 3

[38] OpenAI. Gpt 5, 2025. 1

[39] OpenAI. Think with image, 2025. 2

[40] Jiahao Qiu, Xuan Qi, Tongcheng Zhang, Xinzhe Juan, Jiacheng Guo, Yifu Lu, Yimin Wang, Zixin Yao, Qihan Ren, Xun Jiang, et al. Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution. *arXiv preprint arXiv:2505.20286*, 2025. 3

[41] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:8612–8642, 2024. 2, 6, 4, 5

[42] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 5, 1

[43] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 1, 2, 5, 6, 7, 3, 4, 8

[44] Qianli Shen, Yezhen Wang, Zhouhao Yang, Xiang Li, Haonan Wang, Yang Zhang, Jonathan Scarlett, Zhanxing Zhu, and Kenji Kawaguchi. Memory-efficient gradient unrolling for large-scale bi-level optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:90934–90964, 2024. 3

[45] Ruolin Shen, Xiaozhong Ji, Kai Wu, Jiangning Zhang, Yijun He, HaiHua Yang, Xiaobin Hu, and Xiaoyu Sun. Align and surpass human camouflaged perception: Visual refocus reinforcement fine-tuning. *arXiv preprint arXiv:2505.19611*, 2025.

[46] Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025. 1, 3

[47] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025. 1, 2, 6, 8

[48] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025. 1, 2, 5, 6, 4, 8

[49] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025. 1

[50] Guohao Sun, Hang Hua, Jian Wang, Jiebo Luo, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Latent chain-of-thought for visual reasoning. *arXiv preprint arXiv:2510.23925*, 2025.

[51] Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long CoT reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL: Long Papers)*, pages 5158–5171, Vienna, Austria, 2025. Association for Computational Linguistics. 1

[52] Jihoon Tack, Jaehyung Kim, Eric Mitchell, Jinwoo Shin, Yee Whye Teh, and Jonathan Richard Schwarz. Online adaptation of language models with a memory of amortized contexts. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:130109–130135, 2024. 3

[53] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[54] GLM-V Team. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 1

[55] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 1

[56] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations (ICLR)*, 2025. 5, 6, 8, 2, 3

[57] Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025. 2

[58] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:95095–95169, 2024. 5, 2, 6

[59] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[60] Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:53764–53797, 2024. 3

[61] Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19541–19551, 2025. 5, 6, 8, 2

[62] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 7, 4

[63] Yujun Wang, Jinhe Bi, Soeren Pirk, Yunpu Ma, et al. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. *arXiv preprint arXiv:2506.14766*, 2025.

[64] Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogerio Feris, and Zexue He. M+: Extending memoryllm with scalable long-term memory. *arXiv preprint arXiv:2502.00592*, 2025. 3

[65] Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025. 1, 2, 5, 6, 8

[66] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024. 5, 6, 8, 2, 4

[67] Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*, 2025. 1, 3

[68] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025. 2

[69] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025. 1, 3, 5, 6, 2, 4, 8

[70] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024. 6, 2, 4, 5

[71] Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 14625–14634, 2025. 3

[72] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 5, 2, 6

[73] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23219–23230, 2024. 2

[74] Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal large language model. *arXiv preprint arXiv:2502.17425*, 2025. 1, 2, 5, 6, 8

[75] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning (ICML)*, 2024. 5, 6, 7, 8, 2, 4

[76] Xinlei Yu, Zhangquan Chen, Yudong Zhang, Shilin Lu, Ruolin Shen, Jiangning Zhang, Xiaobin Hu, Yanwei Fu, and Shuicheng Yan. Visual document understanding and question answering: A multi-agent collaboration framework with test-time scaling. *arXiv preprint arXiv:2508.03404*, 2025.

[77] Xinlei Yu, Chengming Xu, Guibin Zhang, Yongbo He, Zhangquan Chen, Zhucun Xue, Jiangning Zhang, Yue Liao, Xiaobin Hu, Yu-Gang Jiang, et al. Visual multi-agent system: Mitigating hallucination snowballing via visual flow. *arXiv preprint arXiv:2509.21789*, 2025. 3

[78] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024. 5, 2, 6

[79] Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*, 2025. 3

[80] Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents. *arXiv preprint arXiv:2509.24704*, 2025. 1, 2, 3, 6

[81] Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. In *The Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 5, 6, 8, 2

[82] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19632–19642, 2024. 3

[83] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1702–1713, 2025.

[84] Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025. 2

[85] Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025. 3

[86] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 1, 2, 5, 6, 4, 8

[87] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19724–19731, 2024. 3

[88] Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47(6):4489–4504, 2025. 2

[89] Yucheng Zhou, Zhi Rao, Jun Wan, and Jianbing Shen. Rethinking visual dependency in long-context reasoning for large vision-language models. *arXiv preprint arXiv:2410.19732*, 2024. 1

VisMem: Latent Vision Memory Unlocks Potential of Vision-Language Models

Supplementary Material

6. Theoretical Foundations

As the mainstream position in anthropological cognitive psychology since the 20th century, short-term memory and long-term memory are two distinct storage systems that can be differentiated based on their functional and neural underpinnings [3, 37]. Specifically, the *Dennis Norris Theory* [37] proposes that short-term memory requires processing new visual information, temporarily storing multiple tokens, and enabling variable signals. It relies neurologically on vision-specific brain regions, *e.g.*, the visual cortex and the posterior superior temporal lobe associated with verbal short-term memory), exhibiting visual dominance; long-term memory, however, centers on abstract semantic representations and relies on semantic-related brain regions like the medial temporal lobe and mid-temporal lobe.

Thus, we propose a framework termed VisMem to invoke dual short and long latent memory during the token-by-token autoregressive generation. Aligned with *Dennis Norris Theory* [37], we instantiate these roles in a VLM backbone via latent vision memory invocation and latent vision memory formation, which together produce distinct short and long latent memory tokens and integrate them into the generation stream of the model.

7. Methodology Details

7.1. Query Builder

As described in Sec. 3.3, the we initialize a lightweight transformer-based encoder as memory builder $\mathcal{B}$. We feed the concatenated memory query $\mathbf{Q}$ and hidden states of vision and output $\mathbf{H}$ into the builder to encoder query as memory hook (see Eq. (5)). The transformer-based builder has L layers of encoders, the output process of the ℓ layer could be summarized as:

$$\mathrm{SA}(x) = \mathrm{SM}\left(\frac{(xW_q)(xW_k)^\top}{\sqrt{d_k}} + M\right)(xW_v), \tag{9}$$

$$x^\ell = \mathrm{FF}\left(\mathrm{LN}\left(x^{\ell-1} + \mathrm{SA}\left(\mathrm{LN}\left(x^{\ell-1}\right)\right)\right)\right) + x^{\ell-1}, \tag{10}$$

where we simplify the input sequence to x, and SM, MHA, FF, LN denote the softmax, multi-head self-attention, feed-forward layer, layer normalization operations, respectively. In addition, M is the mask which only allows attention from memory query $\mathbf{Q}$ to hidden states $\mathbf{H}$, and blocks the reverse direction:

$$M_{ij} = \begin{cases} -C, & i < K \text{ and } j \geq K \\ 0, & otherwise \end{cases}, \tag{11}$$

where $C \gg 0$ is constant, thus the attention is close to $-\infty$.

7.2. Training Recipe

As mentioned in Sec. 3.4, we design a two-stage training pipeline: at the first stage, the main objective is to optimize the memory formation process (see Eq. (7)); at the second stage, the main objective is to optimize the memory invocation (see Eq. (8)). We update the models based on reinforcement learning, *i.e.*, GRPO strategy [42]. Specifically, for each instruction-vision pair (I, V), the policy model $\mathcal{P}$ generates a group of G distinct candidate trajectories, termed as $\mathcal{T} = \{\tau_1, \ldots, \tau_G\}$. For each trajectory, we utilize a $S(\cdot)$ to quantify the performance. Then, a group-relative baseline is calculated via averaging and standardizing all trajectories within the candidate group G:

$$\overline{S} = \frac{1}{G}\sum_{i=1}^{G} S(\tau_i), \hat{S} = \sqrt{\frac{1}{G}\sum_{i=1}^{G}\left(S(\tau_i) - \overline{S}\right)^2}. \tag{12}$$

Consequently, the group-relative advantage of each trajectory could be formulated as:

$$\hat{A} = \frac{S(\tau) - \overline{S}}{\hat{S} + \epsilon}. \tag{13}$$

At the **Stage I**, the reinforcement learning optimizes the memory formation process, whose final objective function is:

$$\begin{aligned} \mathcal{J}_{GRPO}^{stage1}(\phi) = \mathbb{E}_{\tau, \mathbf{M}_{s/l}, \mathbf{Q}} \Bigg[& \frac{1}{G}\sum_{i=1}^{G} \\ & \min\left(\rho_i(\phi)\hat{A}_i, \mathrm{clip}\left(\rho_i(\phi), 1-\epsilon, 1+\epsilon\right)\hat{A}_i\right) \Bigg] \\ & - \beta D_{\mathrm{KL}}\left[\pi_\tau^\phi \| \pi_{\mathrm{ref}}^\phi\right], \end{aligned} \tag{14}$$

where ϵ controls the group-relative advantage $\hat{A}$, β regulates the KL divergence penalty, and the updated policy parameters $\pi^\phi = \pi^\phi(\mathbf{Q} \mid \mathbf{H}) \cdot \pi^\phi(\mathbf{M}_{s/l} \mid \mathbf{Q})$.

At the **Stage II**, the reinforcement learning optimizes the memory invocation process, whose final objective function is:

$$\begin{aligned} \mathcal{J}_{GRPO}^{stage2}(\theta) = \mathbb{E}_{\tau, x} \Bigg[& \frac{1}{G}\sum_{i=1}^{G} \\ & \min\left(\rho_i(\theta)\hat{A}_i, \mathrm{clip}\left(\rho_i(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i\right) \Bigg] \\ & - \beta D_{\mathrm{KL}}\left[\pi_\tau^\theta \| \pi_{\mathrm{ref}}^\theta\right]. \end{aligned} \tag{15}$$

8. Experiment Details

8.1. Training Data

During the two-stage training procedure, we use the same training data to optimize both the memory invocation and memory formation in the latent vision memory system. Initially, we include the training split dataset of the selected benchmarks and retain their original data division. For benchmarks without a training phase, we use them solely for evaluation. Additionally, we incorporate the Visual CoT [41] and Mullberry [70], improving the reasoning abilities.

8.2. Benchmarks

To comprehensively evaluate the performance of the selected baselines, we involve 12 benchmarks, consisting of 5 benchmarks of understanding, 4 benchmarks of reasoning, and 3 benchmarks of generation:

- MMStar [7] is a high-quality vision-centric benchmark meticulously curated by human experts. This benchmark assesses 6 core capabilities across 18 detailed axes of visual understanding.
- MMVet [75] establishes 6 core visual understanding capabilities and investigates 16 critical integrations derived from their combinations. It uses an evaluator tailored for open-ended outputs.
- MMT [72] consists of carefully curated multi-choice visual questions, covering 32 core meta-tasks and 162 subtasks within the field of visual understanding.
- BLINK [15] reconstructs 14 classic computer vision tasks into multiple-choice questions. Each question is paired with either single or multiple images and supplemented with visual prompting.
- MuirBench [56] covers 12 diverse multi-image tasks, which involve 10 categories of multi-image relations. Each standard instance is paired with an unanswerable variant that differs only minimally in semantics.
- MMMU [78] comprises meticulously curated visual questions sourced from college exams, quizzes, and textbooks spanning 30 subjects and 183 subfields, which focus on advanced reasoning grounded in domain-specific knowledge.
- LogicVista [66] evaluates general logical cognition abilities across 5 logical reasoning tasks, which encompass 9 distinct capabilities. Each question is annotated with the correct answer and the human-written reasoning behind the selection.
- Math-V [58] is comprised of high-quality mathematical problems with visual contexts, drawn from real mathematical competitions. It spans 16 distinct mathematical disciplines and is categorized into 5 difficulty levels, evaluating the mathematical reasoning.
- MV-Math [61] is a dataset comprising mathematical problems, integrating multiple images interleaved with text, and detailed annotations. It features multiple-choice, free-form, and multi-step questions across 11 subject areas at 3 difficulty levels.
- HallBench [19] consists of images paired with questions, designed by human experts to assess the hallucination level of generation.
- MultiTrust [81] covers five primary aspects: truthfulness, safety, robustness, fairness, and privacy, evaluating the trustworthiness of generation.
- MMVU [34] encompasses 12 categories, and designs evaluation metrics that measure the quality and error degree of generation.

8.3. Baselines

We select a total of 16 baselines, including the vanilla model [4], 5 direct training paradigms: SFT, Visual-RFT [35], VLM-R1 [43], Vision-R1 [26], and PAPO [65]; 5 image-level paradigms: Sketchpad [24], GRIT [13], PixelReasoner [47], DeepEyes [86], and OpenThinkImg [48]; 4 token-level paradigms: Scaffold [28], ICoT [16], MINT-CoT [8], and VPT [74]; and 1 latent space paradigm: Mirage [69].

Here, VLM-R1 [43] and Vision-R1 [26] follow the main GRPO [20] paradigm based on VLMs. To assess the effectiveness of different methods, our VisMem is trained on Qwen-2.5-VL-7B [4]. For strategies initially implemented on other base models, *e.g.*, GPT-4o [27] and Qwen2-VL [59], we transfer them to Qwen2.5-VL-7B [4] for fair comparison. Besides, we maintain identical training datasets across various counterparts; however, for those with curated datasets, we follow their original settings. For example, Mirage [69] requires additional labeled training images, so we follow its original training dataset; GRIT [13] uses a tailored training process with designed data; and MINT-CoT [8] curates high-quality mathematical samples with grids and annotations.

8.4. Implementations

The configurations and implementations of the experiments include three main parts: the core hyperparameters, the parameters of the LoRA adapter, and the parameters we use during training. The configurations and implementations of the experiments are listed in Tab. 4.

9. Additional Results

9.1. Benchmark Subset Results towards Visual Subcapacities

To precisely identify the capability boundaries and advantages of our VisMem, rather than relying solely on overall scores to judge its quality, we evaluate the results of subsets of MuirBench [56] and LogicVista [66] benchmarks.

Table 4. Configurations of parameters.

Configurations	Parameters	Values	
Core	K	8	
	N_s	4	
	N_l	8	
LoRA [23]	$rank$	16	
	α	32	
	$drop\_out\_rate$	0.1	
	$target\_module$	$[q\_proj, v\_proj]$	
Training		**Stage I**	**Stage II**
	$batch\_size$	8	
	$epoch$	2	
	$warmup\_ratio$	0.2	0.1
	$num\_iteration$	1	
	$learning\_rate$	$5e^{-5}$	$1e^{-5}$
	$optimizer$	AdamW [36]	
	$scheduler$	Cosine	
	$group\_size$	16	
	$clip\_ratio$	0.2	
	$kl\_penalty\_coefficient\ \beta$	0.015	0.030
	$target\_kl\_per\_token$	0.03	0.05
	$penalty\_intensity\ \alpha$	-	0.3

Table 5. Results on 9 selected subsets of MuirBench [56]. We compare our VisMem with the second and third best scored counterparts, and separately use the short or long latent memory to assess the improvements of each.

Method	Counting	Grounding	Matching	Scene	Difference	Cartoon	Diagram	Geographic	Retrieval
Vanilla [4]	44.1	34.2	80.9	70.5	53.2	52.9	82.4	53.7	76.1
VLM-R1 [43]	52.5	38.1	83.6	73.5	58.1	55.1	86.8	56.7	79.4
Vision-R1 [26]	53.8	39.2	**84.5**	73.1	57.4	57.2	87.4	57.9	78.9
VisMem (Short Memory)	**61.3**	<u>49.4</u>	82.7	72.1	<u>58.9</u>	54.0	<u>88.9</u>	61.8	<u>87.5</u>
VisMem (Long Memory)	46.3	42.6	83.2	<u>74.3</u>	55.4	<u>59.4</u>	87.4	<u>62.7</u>	78.3
VisMem	<u>60.8</u>	**52.3**	<u>84.0</u>	**76.2**	**60.6**	**59.7**	**90.1**	**65.5**	**89.8**

We select 9 subsets of the former benchmark, including: counting, grounding, matching, scene, difference, cartoon, diagram, geographic, and retrieval. While in the latter benchmark, we also select 10 subsets, including 5 reasoning skills: inductive, deductive, numerical, spatial, and mechanical, and 5 capacities: patterns, puzzles, OCR, graphs, and tables. It is worth noting that the selected subsets are only part of the benchmark, thus, the average values of the 10 subsets are not the results of the benchmarks.

As listed in Tab. 5, compared with VLM-R1 [43] and Vision-R1 [26], our VisMem achieves the best results on 7 subsets and ranks second on the remaining two subsets. Specifically, it has a generalized enhancement of at least 5% over the base model. Besides, VisMem improves the performance the vanilla model by 16.7% / 18.2% / 11.8% / 13.7% on the counting, grounding, geographic, and retrieval subtasks, vastly exceeding the second-best counterpart by 7.0-13.1%. These results indicate that our latent vision memory system significantly promote the fine-grained visual cognition and perception of the base VLMs.

As presented in Tab. 6, our VisMem outperforms two baseline models, *i.e.*, VLM-R1 [43] and Vision-R1 [26], by achieving the top performance across 8 subsets. Specifically, it delivers a generalized improvement of no less than 7% over the base model. Notably, on inductive, deductive, graph-based, and table-based sub-tasks, VisMem surpasses

Table 6. Results on 10 selected subsets (5 reasoning skills and 5 capabilities) of LogicVista [66]. We compare our VisMem with the second and third best scored counterparts, and separately use the short or long latent memory to assess the improvements of each.

Method	Inductive	Deductive	Numerical	Spatial	Mechanical	Patterns	Puzzles	OCR	Graphs	Tables
Vanilla [4]	44.6	45.0	39.7	37.9	48.7	30.1	32.5	41.6	34.4	36.8
VLM-R1 [43]	53.7	52.7	45.8	44.1	57.3	35.8	42.8	49.0	46.5	52.6
Vision-R1 [26]	53.5	51.4	46.7	44.8	**58.9**	36.5	43.6	**49.7**	48.2	53.8
VisMem (Short Memory)	49.8	50.1	44.7	45.2	54.3	35.2	42.0	47.6	50.3	54.1
VisMem (Long Memory)	57.5	58.4	42.8	40.0	52.0	35.7	38.0	47.4	48.9	51.3
VisMem	**59.4**	**59.8**	**46.9**	**47.2**	57.4	**38.9**	**44.6**	48.5	**52.8**	**57.9**

the vanilla model by 14.8%, 14.8%, 18.4%, and 21.1%, respectively, which exceeds the second-ranked model by a substantial margin of 5.3–7.1%. These results demonstrate that our latent visual memory system delivers contextualized semantic knowledge, thereby enhancing visual reasoning and robust generation capabilities.

9.2. Cross-domain Generalization

To evaluate the cross-domain generalization capability of our model, we train it exclusively on general datasets, namely, Visual CoT [41] and Mullberry [70]), to verify whether latent visual memory can be transferred to unseen domains. As shown in Tab. 7 and Fig. 7, our method demonstrates superior performance, which exhibits a smaller performance drop than the fully trained model across all four selected benchmarks, confirming strong cross-domain generalization. Despite being trained on only two datasets, our method achieves a significant performance improvement of 9.1–20.5% across the four benchmarks, with a mere 2% performance gap relative to the fully trained model. When compared to other baselines, it still maintains a performance lead of 3.4% / 6.7% / 2.7% / 4.7% across the four evaluations, respectively.

In general, the image-level, token-level, and latent space paradigms suffer from smaller performance degradation, whereas the direct training paradigm exhibits inferior generalization ability. For example, VLM-R1 [43] experiences a 5.3% performance drop; by contrast, this value is only 2.1% for OpenThinkImg [48], 1.1% for MINT-CoT [8], and 2.3% for our method. These results indicate that while direct training optimizations notably improve performance on specific tasks, they compromise generalization ability to some extent.

9.3. Catastrophic Forgetting Mitigation

To assess the extent of catastrophic forgetting, we conducted continual learning experiments with our VisMem and other baselines. As presented in Tab. 8 and Fig. 8, our method effectively mitigates forgetting of earlier tasks. It consistently achieves the best performance at each stage, demonstrating strong robustness against catastrophic forgetting. Following four-stage sequential continual training, it retains 72.1% performance on MMVet [75], outperforming 68.4% of DeepEyes [86] and 67.0% of Mirage [69].

While the direct training paradigm significantly improves performance on specific tasks, it adapts to new tasks via direct updates to core parameters. This introduces conflicts when parameter update directions contradict the storage of existing knowledge, compounded by a lack of constraints from prior knowledge. Consequently, in stage 3, the performance of most direct training methods even falls below that of the vanilla model. In contrast, methods such as OpenThinkImg [48] and our proposed VisMem exhibit stronger knowledge retention and forward transfer capabilities. For instance, in stage 3, training on additional datasets further improves their performance on MMVet [75].

9.4. Versatility across Various Base Models

As presented in Tab. 2 and Fig. 11, we incorporate our latent visual memory paradigm into 9 base models, including Qwen2.5-VL-3B/7B/32B [4], LLaVA-OV-1.5-4B/8B [1], and InternVL-3.5-4B/8B/14B/38B [62]. Our VisMem consistently enhances the visual capabilities of all base models, spanning 3B to 38B parameter sizes across three VLM families. For the widely used medium-sized models (*i.e.*, 7B or 8B parameter models), our latent visual memory delivers substantial performance gains, which brings a 6.3–23.1% improvement across all benchmarks for Qwen2.5-VL-7B [4], a 5.5–20.2% improvement for LLaVA-OV-1.5-8B [1], and a 4.8–17.6% improvement for InternVL-3.5-8B [62], respectively.

Furthermore, in most benchmarks, smaller-parameter base models yield greater performance gains than their medium- or large-sized counterparts. This phenomenon may stem from an imbalance in task difficulty, which makes it more challenging for models with higher baseline scores to achieve further improvements. In contrast, larger models exhibit more significant gains in dense reasoning benchmarks: the integration of latent visual memory overcomes bottlenecks in visual reasoning by providing fine-grained

Table 7. Results of various models with full training datasets and partial datasets (Visual CoT [41] and Mulberry [70]), and evaluated across four benchmarks.

Method	MMVet		MuirBench		MV-Math		MultiTrust	
	Full	Part	Full	Part	Full	Part	Full	Part
Vanilla [4]	66.0		57.4		18.9		64.8	
SFT	67.5	65.8	58.7	57.2	22.8	21.2	67.0	65.4
Visual-RFT [35]	70.5	65.3	62.9	57.8	26.5	24.2	70.7	66.0
VLM-R1 [43]	73.0	67.7	63.8	59.0	34.6	32.1	69.9	66.1
Vision-R1 [26]	71.7	68.4	64.0	59.8	38.7	35.6	72.6	67.1
PAPO [65]	69.8	68.6	56.7	56.4	34.8	32.8	67.7	66.4
DeepEyes [86]	70.5	67.9	63.0	60.6	31.5	27.9	72.6	68.5
OpenThinkImg [48]	71.6	69.5	61.7	59.7	28.0	25.9	74.0	68.3
ICoT [16]	67.9	67.1	57.0	56.4	30.8	28.3	69.1	68.4
MINT-CoT [8]	69.5	68.4	58.9	57.8	39.2	36.4	71.4	70.2
Mirage [69]	71.8	70.2	59.0	57.2	35.4	33.1	66.1	64.0
VisMem (Ours)	**75.1**	**72.9**	**69.8**	**66.4**	**41.4**	**39.1**	**77.0**	**74.9**

Table 8. Results of various models on MMVet [75] with four-stage continual learning. Stage 0: MMVet [75]; Stage 1: BLINK [15], and MuirBench [56]; Stage 2: LogicVista [66], and Math-V [58]; Stage 3: MultiTrust [81], and MMVU [34].

Method	Stage 0	Stage 1	Stage 2	Stage 3	Original
Vanilla [4]	66.0				
SFT	71.4	70.6	62.3	60.1	67.5
Visual-RFT [35]	74.0	72.2	67.3	65.7	70.5
VLM-R1 [43]	77.8	74.1	66.4	66.9	73.0
Vision-R1 [26]	76.9	74.0	66.1	66.3	71.7
PAPO [65]	75.0	74.5	63.4	62.9	69.8
DeepEyes [86]	74.1	74.6	68.9	68.4	70.5
OpenThinkImg [48]	76.2	74.7	66.5	67.9	71.6
ICoT [16]	71.9	71.3	67.1	64.7	67.9
MINT-CoT [8]	72.4	71.8	65.8	66.2	69.5
Mirage [69]	**79.1**	77.8	68.7	67.0	71.8
VisMem (Ours)	78.6	**78.9**	**71.3**	**72.1**	**75.1**

visual evidence and semantic knowledge. Notably, this model-agnostic approach, independent of specific model architectures or structures, bolsters the prospects for broad practical application.

9.5. Ablation Study

The vanilla model establishes a baseline characterized by the shortest inference time and highest speed across all benchmarks, yet exhibits the lowest performance. This confirms that latent vision memory is indispensable for enhancing task performance. For the random memory invocation variants, increasing the invocation probability (25%–100%) results in longer inference time and reduced speed. Performance peaks at a 75% probability before declining, indicating that excessive memory invocation impairs efficiency without yielding additional performance benefits. Ablation studies of the short-term and long-term memory components reveal task-specific advantages: the short-term memory component outperforms on MuirBench [56] and MultiTrust [81], while the long-term component demonstrates superior performance on MV-Math [61]. Notably, the complete VisMem framework achieves the highest performance across all benchmarks, validating the value of integrating dual-component vision memory for balanced and robust vi-

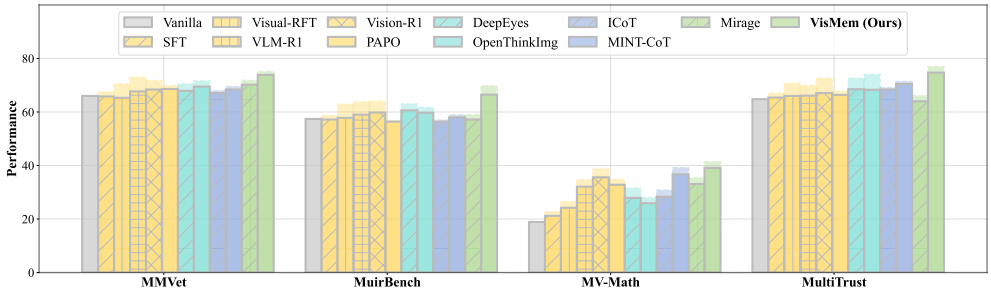


Figure 7. Results of various models of the cross-domain generalization study. Models are only trained on Visual CoT [41] and Mulberry [70], and are evaluated on four benchmarks.

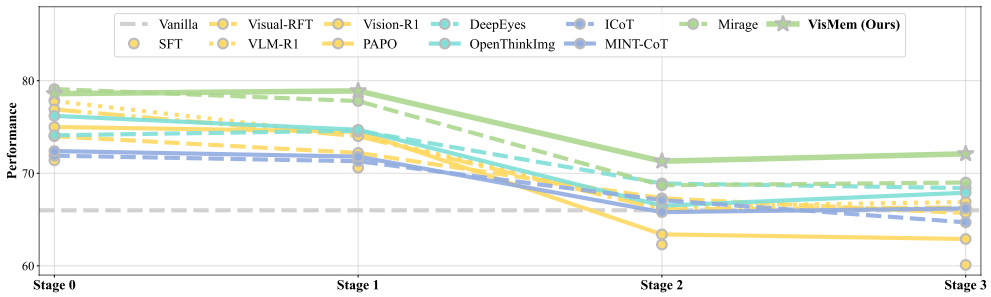


Figure 8. Results of four-stage continual learning on MMVet [75]. The model is sequentially trained on each training data combination (Stage 0 → Stage 1 → Stage 2 → Stage 3). Stage 0 only includes MMVet [75] as training data, while Stage 1, 2, 3 add data targeting visual understanding [7, 15, 56, 72], reasoning [58, 61, 66, 78],generation [19, 34, 81].

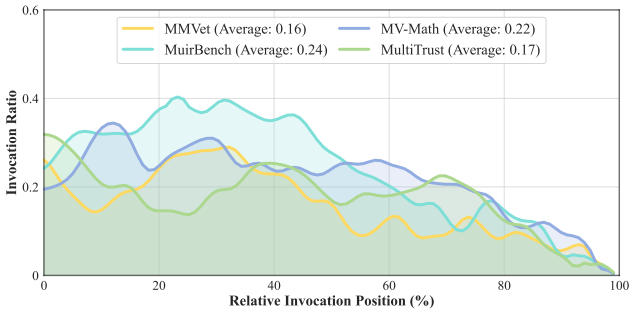
Table 9. Ablations of latent vision memory invocation and dual vision memory formation. Following [80], "Random Invocation" denotes that the latent memory is inserted into the output sequence with a certain probability when outputting delimiter symbol tokens, and short or long latent memory is inserted with equal probability. When only utilizing short or long latent memory, we directly skip the formation of the specific memory if invocation tokens are predicted and continue the process of decoding.

Ablation	**MMVet**			**MuirBench**			**MV-Math**			**MultiTrust**		
	Time	Speed	Perf.	Time	Speed	Perf.	Time	Speed	Perf.	Time	Speed	Perf.
Vanilla	**0.76**	**1.32**	66.0	**3.79**	**0.26**	57.4	**5.47**	**0.18**	18.9	**3.62**	**0.28**	64.8
Random Invocation (25%)	0.80	1.25	69.2	3.94	0.25	59.4	8.79	0.11	29.8	6.14	0.16	69.4
Random Invocation (50%)	0.83	1.20	71.9	4.12	0.24	63.2	11.68	0.09	26.1	8.62	0.12	68.5
Random Invocation (75%)	0.86	1.16	73.6	4.27	0.23	62.7	14.78	0.07	21.9	10.11	0.10	63.7
Full Invocation (100%)	0.88	1.14	73.4	4.43	0.23	56.0	17.87	0.06	17.5	13.43	0.07	62.6
Short-term Memory	0.79	1.27	71.5	4.00	0.25	65.6	7.64	0.12	29.6	4.96	0.20	73.6
Long-term Memory	0.81	1.23	69.4	3.95	0.25	60.2	7.61	0.12	36.1	4.80	0.21	69.8
Complete VisMem (Ours)	0.84	1.19	**75.1**	4.10	0.24	**69.8**	7.87	0.13	**41.4**	5.85	0.17	**77.0**

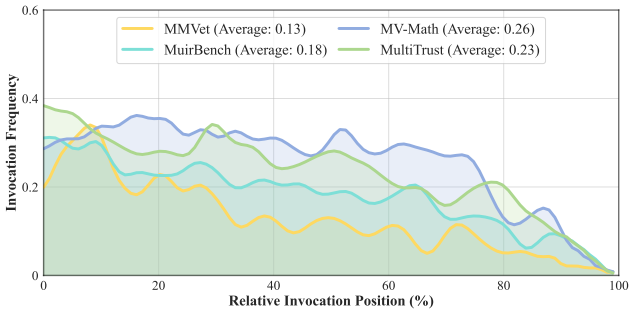
sual capacities.

9.6. Analysis of Latent Vision Memory

We visualize the invocation ratio and relative invocation position, as presented in Fig. 5 and 9: the former illustrates benchmark-specific differences between the two memory components, while the latter depicts type-specific variations across the four benchmarks. In addition, as reported in Tab. 5 and 6, the short- and long-term latent visual memory components exhibit task-specific advantages for different visual sub-tasks. For instance, the short-term memory provides supplementary visual information to support enhanced visual understanding, such as counting, grounding, and visual retrieval. By contrast, the long-term memory encodes contextualized semantic knowledge, which strength-



(a) Short Memory Invocation



(b) Long Memory Invocation

Figure 9. Results of memory invocation ratio and relative position across four benchmarks. The former denotes the proportion of invoked samples to all samples, while the relative position denotes the position in the whole output sequence when the invocation occurred. We apply gaussian smoothing to the curves to highlight their main trends.

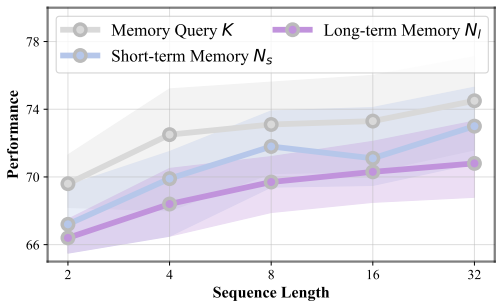


Figure 10. Results of sensitivity analysis on the sequence length of memory query K, short- and long-term memory N_s and N_l.

Table 10. Results of different length of memory query K.

K	MMVet	MuirBench	MV-Math	MultiTrust
Vanilla	66.0	57.4	18.9	64.8
2	69.6	66.0	34.7	71.9
4	72.5	68.9	40.6	74.8
8	73.1	69.8	41.1	77.0
16	73.3	70.0	**41.4**	77.7
32	**74.5**	**70.3**	40.9	**78.2**

ens complex visual reasoning. These results reveal that our proposed VisMem dynamically adjusts invocation position and frequency according to task characteristics, thereby balancing efficiency and performance.

9.7. Sensitive Analysis of Sequence Lengths

We conduct an analysis on MMVet [75] focused on the lengths of three key sequences: the memory query K, the short-term latent visual memory N_s, and the long-term latent visual memory N_l. It is observed that as the lengths of these three sequences increase from 2 to 32, model performance improves accordingly, but this is accompanied by increased computational costs.

Table 11. Results of different length of short latent vision memory N_s and the length of long latent vision memory N_l across four benchmarks.

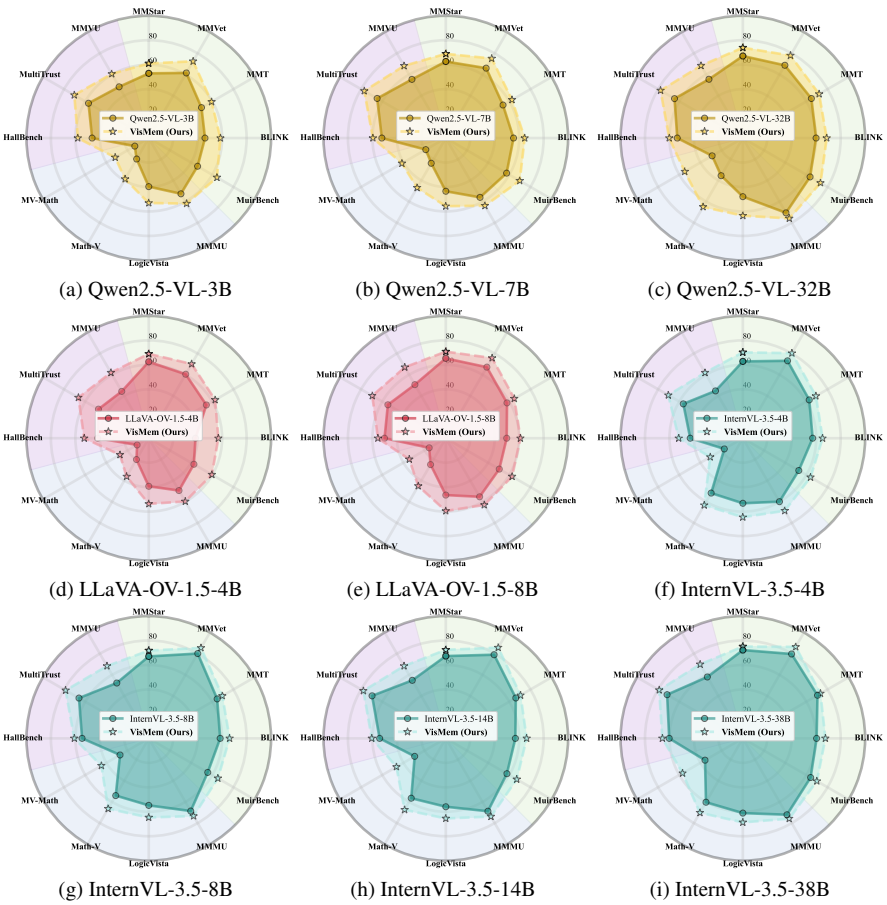
N_s	N_l	MMVet	MuirBench	MV-Math	MultiTrust
Vanilla		66.0	57.4	18.9	64.8
2	-	67.2	63.7	28.2	69.3
4	-	69.9	64.6	31.5	71.4
8	-	71.8	65.2	33.8	73.4
16	-	71.1	67.8	34.0	73.3
32	-	73.0	69.1	34.4	72.7
-	2	66.4	60.3	29.3	71.0
-	4	68.4	61.8	32.4	72.8
-	8	69.7	63.0	33.5	74.2
-	16	70.3	63.4	34.8	74.9
-	32	70.8	63.1	35.5	75.3
8	16	**75.1**	**69.8**	**41.1**	**77.0**

9.8. Inference Efficiency

As presented in Tab. 12 and the bubble plots in Fig. 6, we compare the average inference time, average inference speed, and task performance across the four benchmarks. Our approach achieves an optimal performance-efficiency balance, with minimal additional time overhead. For instance, image-level paradigms exhibit nearly twice the inference time of the vanilla model, resulting in significant latency and substantial inference overhead. In contrast, our VisMem introduces only controllable computational latency increments, ranging from 8.2% to 43.8% relative to the vanilla model, which are on par with those of other direct training and token-level paradigms.

Table 12. Average inference time per sample (seconds), average inference speed (samples / seconds), and task performances across four benchmarks on various methods. Perf. indicates Performance.

Method	MMVet			MuirBench			MV-Math			MultiTrust		
	Time	Speed	Perf.	Time	Speed	Perf.	Time	Speed	Perf.	Time	Speed	Perf.
Vanilla [4]	0.76	1.32	66.0	**3.79**	**0.26**	57.4	**5.47**	**0.18**	18.9	**3.62**	**0.28**	64.8
SFT	**0.75**	**1.33**	67.5	3.82	0.26	58.7	6.35	0.16	22.8	3.68	0.27	67.0
Visual-RFT [35]	0.76	1.32	70.5	3.81	0.26	62.9	5.66	0.17	26.5	3.65	0.27	70.7
VLM-R1 [43]	0.77	1.30	73.0	3.83	0.26	63.8	7.88	0.13	34.6	3.69	0.27	69.9
Vision-R1 [26]	0.77	1.30	71.7	3.83	0.26	64.0	8.42	0.12	38.7	3.71	0.27	72.6
PAPO [65]	0.76	1.32	69.8	3.81	0.26	56.7	6.74	0.15	34.8	3.68	0.27	67.7
Sketchpad [24]	2.39	0.42	64.5	8.90	0.11	52.8	9.10	0.11	24.6	5.47	0.18	66.2
GRIT [13]	0.80	1.25	67.8	4.07	0.25	51.0	8.45	0.12	22.4	4.06	0.25	67.3
PixelReasoner [47]	1.45	0.69	67.1	7.34	0.14	60.5	9.96	0.10	25.9	5.60	0.18	69.9
DeepEyes [86]	3.21	0.31	70.5	8.46	0.12	63.0	11.72	0.09	31.5	6.14	0.16	72.6
OpenThinkImg [48]	3.68	0.27	71.6	8.69	0.12	61.7	10.38	0.10	28.0	6.43	0.16	74.0
Scaffold [28]	0.83	1.20	67.0	4.35	0.23	52.9	7.01	0.14	21.0	3.88	0.26	68.5
ICoT [16]	0.97	1.15	67.9	4.57	0.22	57.0	8.94	0.11	30.8	4.20	0.24	69.1
MINT-CoT [8]	0.81	1.23	69.5	4.18	0.24	58.9	7.89	0.13	39.2	4.03	0.25	71.4
VPT [74]	2.98	0.34	70.8	9.63	0.10	63.5	9.59	0.10	34.7	5.79	0.17	64.7
Mirage [69]	0.86	1.16	71.8	4.02	0.25	59.0	7.71	0.13	35.4	3.82	0.26	66.1
VisMem (Ours)	0.84	1.19	**75.1**	4.10	0.24	**69.8**	7.87	0.13	**41.4**	3.85	0.26	**77.0**



(a) Qwen2.5-VL-3B (b) Qwen2.5-VL-7B (c) Qwen2.5-VL-32B

(d) LLaVA-OV-1.5-4B (e) LLaVA-OV-1.5-8B (f) InternVL-3.5-4B

(g) InternVL-3.5-8B (h) InternVL-3.5-14B (i) InternVL-3.5-38B

Figure 11. Results on different base models.