

# Adapting Like Humans: A Metacognitive Agent with Test-time Reasoning

Yang Li<sup>1,\*</sup>Zhiyuan He<sup>1,\*</sup>Yuxuan Huang<sup>4,\*</sup>Zuhanling Xiao<sup>2</sup>Chao Yu<sup>3</sup>Meng Fang<sup>4</sup>Kun Shao<sup>1</sup>Jun Wang<sup>5,†</sup>

## Abstract

Recent Vision-Language Models (VLMs) exhibit strong perceptual reasoning abilities, yet they often struggle to adapt efficiently when encountering novel tasks at test time. In contrast, humans leverage the metacognitive model with memory, enabling continuous strategy refinement through metacognitive control when faced with new challenges. To bridge this gap, we propose metacognitive test-time reasoning (MCTR), a framework that equips models with the ability to learn, adapt, and improve during test time through metacognitive self-updating. Inspired by the dual structure of human metacognition, MCTR comprises meta-level and object-level VLM reasoning modules, each equipped with dedicated memory systems for hierarchical adaptive reasoning. Specifically, MCTR consists of (1) a meta-reasoning module which incrementally builds a structured memory by discovering and storing task-relevant rules, environmental patterns, and action-outcome relationships from test-time observations as natural language descriptions; and (2) an action-reasoning module that determines optimal actions through context-aware perception and strategic reasoning by dynamically retrieving and integrating knowledge from memory. The action-reasoning module continuously updates its policy through proposed metacognitive test-time reinforcement learning, adapting as knowledge memory evolves. We evaluate MCTR on 45 Atari games (33 seen, 12 unseen). MCTR demonstrates robust test-time adaptation, achieving 9/12 top-1 results on unseen games compared with baselines. Analyses through ablations, learning dynamics, and case studies reveal the complementary contributions of both components and show meta-reasoning evolving toward human-like adaptation strategies.

## 1. Introduction

Large language models (LLMs) have made substantial progress towards general reasoning, driven by large-scale

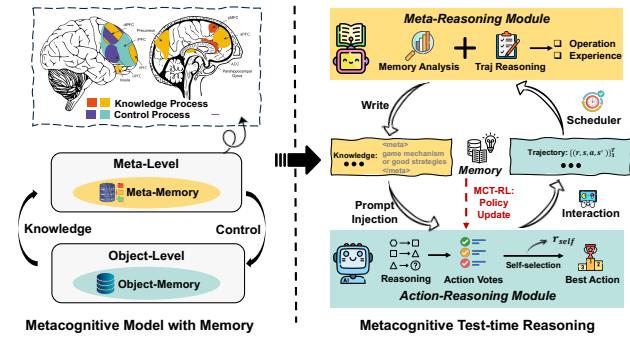


Figure 1. Comparison between a human metacognitive model and our MCTR framework. Left: A human metacognition model [17, 37, 44] comprises meta-level and object-level processes with dedicated memory systems, enabling bidirectional knowledge and control flow. Brain regions associated with knowledge (orange) and control (green) processes are shown. Right: MCTR attains human-like adaptation to novel tasks without prior knowledge through two synergistic modules: a meta-reasoning module (orange) that discovers operational knowledge under dynamic scheduling, and an action-reasoning module (green) that leverages this knowledge for multi-step reasoning while adapting via test-time reinforcement learning.

pre-training and neural scaling laws [10, 40, 43, 55]. Recent advances further improve reasoning through chain-of-thought prompting [57], tree-based search [61], and reinforced reasoning pipelines that use process reward models and trial-and-error self-play [9, 41, 53, 59], giving rise to large reasoning models that scale both train time and test time computation [59]. Beyond scaling, LLMs display emergent abilities such as role playing [29] and analogical reasoning [56], with applications in tool use, autonomous agents, and embodied decision-making. For adaptation at deployment, recent test-time approaches include inference-only methods such as in-context learning and prompt tuning [12, 34, 49], test-time training [3, 51, 64], and memory-based online reinforcement learning [63] that refines policies using environmental feedback [50].

Despite this progress, current vision-language reasoning models remain limited in fluid intelligence, the capacity to adapt to novel, complex tasks without prior knowledge [6, 18]. Their abilities are largely crystallised, relying on accumulated knowledge and familiar patterns. Ex-

\*These authors contributed equally to this work. † Corresponding author. <sup>1</sup>Huawei Noah’s Ark Lab. <sup>2</sup>University of Oxford. <sup>3</sup>Tsinghua University. <sup>4</sup>University of Liverpool. <sup>5</sup>University College London.

isting test time adaptation methods face further drawbacks. Test-time training [3, 51] incurs non-trivial computational cost due to gradient updates during inference; prompt-based techniques [49] depend on augmentation or retrieval that transfers poorly to unfamiliar task structure; and recent test-time reinforcement learning (TTRL) [62, 64], although able to discover strategies by trial and error without external rewards, remains limited in interpretability. Crucially, these methods do not exhibit fluid intelligence because they rely on implicit optimisation signals rather than constructing structured knowledge about task dynamics and reasoning metacognitively about when and how to apply strategies.

Humans facing novel tasks adapt from limited experience by observing successes, forming hypotheses, and refining strategies without extensive retraining. Cognitive science attributes this ability to metacognition [17, 37, 44] (left panel, Fig. 1), a two-level architecture with memory in which object-level processes execute tasks while meta-level processes monitor and control them in real time. Empirical studies show that effective learners revise reasoning strategies mid-task without external feedback [11, 15, 45], achieving rapid adaptation with high sample efficiency [32]. This perspective motivates endowing vision-language models with analogous meta-level capabilities: to reflect on their own traces, evaluate effectiveness online, and adjust strategies from limited experience.

Motivated by this, we propose **MetaCognitive Test-Time Reasoning (MCTR)**, a framework that equips vision–language models with human-like adaptive capabilities through structured meta-reasoning and experience-driven learning during inference. As shown in Fig. 1, MCTR comprises two modules aligned with the metacognitive view. The meta-reasoning module performs retrospective analysis of accumulated trajectories, discovers operational patterns and strategies for new tasks, and stores them as natural language entries in a knowledge memory. A scheduler regulates meta-reasoning module invocation based on knowledge stability, prioritising frequent summarisation when knowledge is sparse and reducing overhead as the memory matures. The action-reasoning module conditions on this knowledge to conduct mul-step vision–language reasoning and updates its policy through metacognitive test-time reinforcement learning with self-consistency rewards, enabling online adaptation when external supervision is absent or feedback is weak.

We evaluate MCTR on 45 Atari games (33 seen, 12 held-out) to assess generalisation. MCTR achieves 9/12 top-1 results on unseen games, demonstrating robust test-time adaptation. To understand what drives this improvement, we conduct three complementary analyses. Ablation studies confirm both metacognitive reasoning and MCT-RL are essential for adaptation. Learning dynamics analysis reveals MCTR balances reasoning stability through multi-

path consensus with genuine policy adaptation, evidenced by increasing agreement among current reasoning traces while declining alignment with historical actions. Besides, qualitative analysis shows meta-reasoning evolves from exploratory hypotheses to concrete, goal-directed strategies that mirror human adaptation patterns. Together, these results indicate that metacognitive test-time reasoning provides a promising route to flexible and robust decision-making in open-ended domains.

The main contributions of this paper are threefold:

- We introduce metacognitive test-time reasoning, a new paradigm that enables vision–language models to self-reflect and learn adaptively during inference.
- We develop a complete MCTR pipeline, including (i) pre-deployment supervised reasoning fine-tuning, and (ii) test-time adaptation through metacognitive reinforcement learning with memory accumulation and self-consistency rewards.
- We demonstrate state-of-the-art zero-shot adaptation to twelve unseen long-horizon Atari games, with ablations confirming the necessity of metacognitive components for human-like flexibility.

## 2. Related Work

**Vision-Language Models and Test Time Adaptation.** Vision-language models show strong reasoning via chain-of-thought prompting [57], tree-based search [23, 61], and process reward models [9, 41, 53, 59], with applications in embodied decision-making [4, 13], games [14, 33], and robotics [2, 27]. Yet behaviour remains largely crystallised, drawing on pre-training patterns rather than adapting to novel structures. Test-time adaptation enables adjustment to new distributions without retraining [51, 54]. Prompt-based methods [12, 34, 49] rely on in-context learning but struggle with unfamiliar structure; test-time training [3, 19, 51] performs gradient updates at inference and is costly; test-time reinforcement learning [50, 62, 64] exploits environmental feedback but requires extensive interaction. Alternatives use entropy minimisation [39, 54] or perplexity reduction [26]. Fluid intelligence, the ability to solve novel problems without prior knowledge, remains challenging [6]. ARC [7, 18] targets rule induction from few examples, with approaches spanning programme synthesis [1], neurosymbolic reasoning [25], and large language models [22]. Unlike methods using statistical proxies or meta-training, our approach explicitly evaluates adaptation quality through metacognitive self-reflection, linking cognitive principles with practical test-time adaptation.

**Cognition-inspired Reasoning Models.** Cognitive theories motivate architectures that integrate human-like processes. Kahneman’s dual process account [31] separates fast, intuitive judgements (System 1) from slow, deliberate reasoning (System 2), a distinction realised in recent

systems [24, 35, 58, 60]. Examples include LLM2, which combines generation with process-based verification [60], and SOFAI, which coordinates heuristic and deliberate routines for adaptability [20]. Hierarchical models extend this across levels of abstraction; HRM links high level planning with low level execution [52]. A complementary line studies metacognition, which monitors and controls one’s own inference [8, 30, 37]. Classical architectures separate object level reasoning from meta level control [8], while recent work explores self reflection in language models [28, 36, 42, 48]. Many approaches require meta training [16, 38] or focus on single turn refinement. We instantiate the dual level architecture of Nelson and Narens [37] for test time reasoning, enabling continuous metacognitive adaptation through structured monitoring and control without meta training.

### 3. Preliminaries

We briefly review two paradigms central to test-time adaptation: Test-Time Training (TTT) and Test-Time Reinforcement Learning (TTRL).

**Test-time Training.** TTT enables parametric models to adapt during inference through dynamic parameter updates [3]. For each test input (or batch)  $x_{\text{test}}$ , TTT constructs a temporary training dataset  $\mathcal{D}_{\text{TTT}}$  and optimizes model parameters to minimize a loss function:

$$\theta_{\text{test}} = \arg \min_{\theta} \sum_{d_{\text{TTT}} \in \mathcal{D}_{\text{TTT}}} \mathcal{L}(\text{LM}(d_{\text{TTT}}; \theta)). \quad (1)$$

The adapted parameters  $\theta_{\text{test}}$  are then used for prediction. Previous work typically constructs  $\mathcal{D}_{\text{TTT}}$  using unsupervised objectives applied to  $x_{\text{test}}$  alone. In few-shot settings, we extend TTT by treating demonstration examples  $\{(x_k, y_k)\}_{k=1}^K$  as supervised training instances.

**Test-time Reinforcement Learning.** To address scenarios where labeled supervision is unavailable during deployment, TTRL [64] reformulates test-time adaptation as a sequential decision problem. The model maintains a stochastic policy  $\pi_{\theta}(y|x)$  that generates responses  $y$  conditioned on input prompts  $x$ , where  $\theta$  represents the adaptable parameters. TTRL constructs self-supervised training signals through output self-consistency. For each test input  $x$ , the model samples a collection of  $N$  candidate responses  $\mathcal{Y} = \{y_1, y_2, \dots, y_N\}$  from its current policy. These candidates are then aggregated—typically through majority voting or answer clustering—to produce a consensus prediction  $\hat{y}$ . The intuition is that this consensus captures the model’s most confident prediction, which can serve as a reliable pseudo-label for parameter refinement. The adaptation objective encourages the policy to assign higher probability to responses that align with the consensus:

$$\mathcal{L}_{\text{TTRL}}(\theta) = -\mathbb{E}_{y \sim \pi_{\theta}(\cdot|x)} [\text{sim}(y, \hat{y}) \cdot \log \pi_{\theta}(y|x)], \quad (2)$$

where  $\text{sim}(y, \hat{y})$  quantifies the similarity between generated output  $y$  and consensus  $\hat{y}$ . Parameters are updated via gradient descent to minimize this loss, effectively steering the model toward more self-consistent predictions. This approach enables unsupervised test-time adaptation by exploiting the model’s internal agreement structure rather than relying on external labels.

### 4. Methodology

Drawing on the hierarchical structure of human metacognition, MCTR endows vision-language models with adaptive test-time reasoning through (1) a meta-reasoning module for reflective knowledge formation and (2) an meta-reasoning module for knowledge-guided decision-making.

As illustrated in Fig. 2, MCTR instantiates this dual-process architecture via two synergistic modules. Governed by an adaptive scheduler, meta-reasoning module performs memory analysis and trajectory reasoning to discover operational experiences, such as task mechanisms and strategies, stored as natural language descriptions. The scheduler prioritizes frequent analysis when knowledge is sparse and gradually reduces invocations as memory matures. Complementarily, meta-reasoning module leverages these experiences both non-parametrically, through prompt-based context injection to guide reasoning, and parametrically, via metacognitive test-time reinforcement learning (MCT-RL) at every fixed timestep interval, to continually refine its policy as the knowledge base evolves. Crucially, MCT-RL enables the meta-reasoning module to optimize its reasoning and decision-making policy while efficiently reusing test-time data under no external supervision.

The remainder of this section is organised as follows: Section 4.1 details the meta-reasoning module; Section 4.2 describes the meta-reasoning module.

#### 4.1. Meta-Reasoning Module

We now describe the meta-reasoning module, as shown at the top of Fig. 2, which operates at the metacognitive level to discover task knowledge from interaction experience. Meta-reasoning module maintains an explicit knowledge memory  $\mathcal{M}_t$  that guides both VLM reasoning and policy optimization (MCT-RL).

Rather than invoking meta-reasoning module at every timestep, an adaptive scheduler dynamically modulates meta-reasoning frequency. Let  $k_t$  denote the interval (in timesteps) between consecutive meta-reasoning module activations. The scheduler amplifies this interval exponentially:

$$k_{t+1} = \text{clip}(k_t / \gamma, k_{\min}, k_{\max}) \quad (3)$$

Where  $\gamma < 1$  controls the adaptation rate, and  $\text{clip}(\cdot, k_{\min}, k_{\max})$  enforces interval bounds. This mechanism naturally prioritizes frequent meta-reasoning during

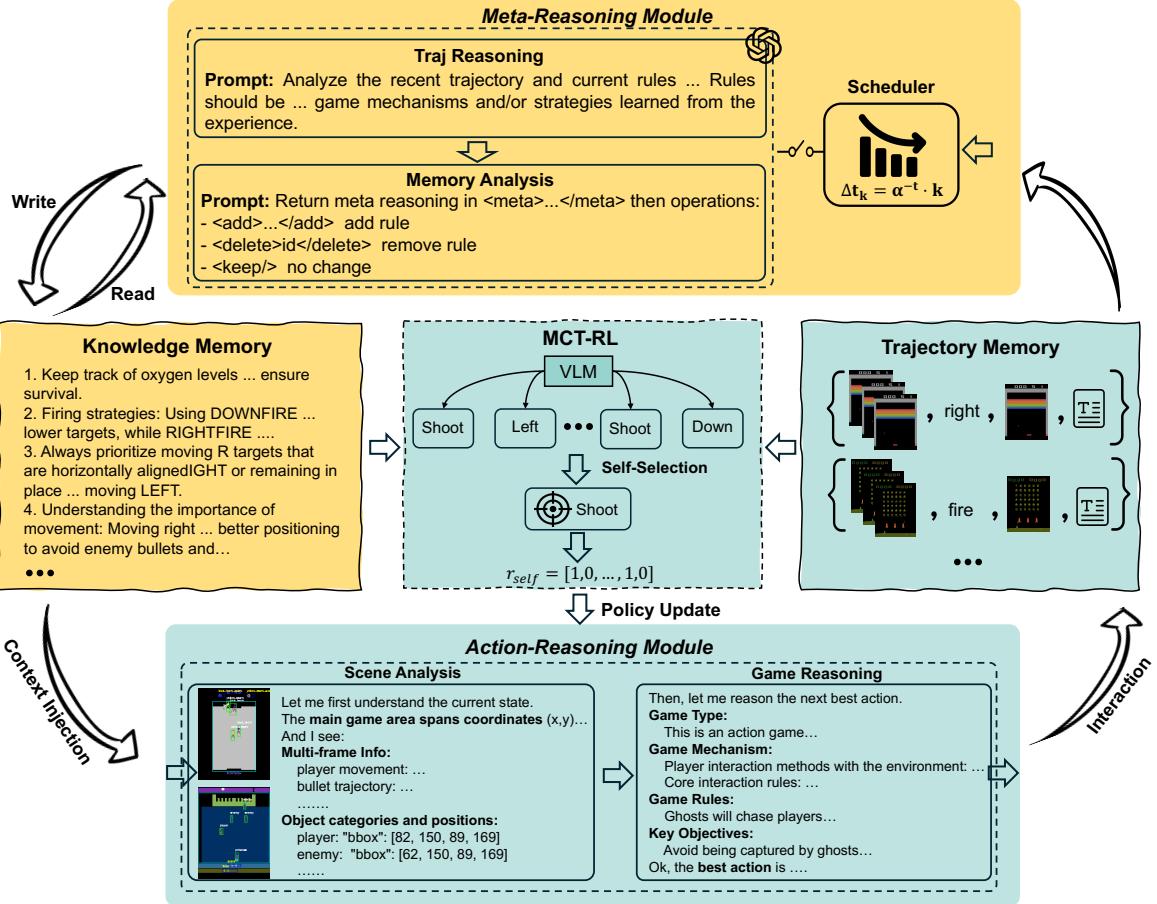


Figure 2. **Overview of Meta-Cognitive Test-Time Reasoning (MCTR) Framework.** The **meta-reasoning module** (yellow, meta-level process in human metacognitive model) performs retrospective analysis on accumulated trajectory memory and generates memory operations ( $<\text{add}>$ ,  $<\text{delete}>$ ,  $<\text{keep}>$ ) to update knowledge memory. A scheduler dynamically adjusts meta-reasoning frequency to prioritize frequent analysis when knowledge is sparse and reduce invocations as memory matures. The **meta-reasoning module** (green, object-level process) handles real-time decision-making by injecting knowledge into context to guide multi-step VLM reasoning. The ARM continuously refines its policy at test time through reinforcement learning using self-supervised signals from action majority voting and real-time knowledge and trajectory memory.

early exploration when  $\mathcal{M}_t$  is sparse and knowledge is rapidly evolving, while reducing computational overhead as the knowledge base stabilizes and matures.

Upon activation at timestep  $t$ , meta-reasoning module retrieves a trajectory segment  $\tau_{[t-k:t]} = \{(s_i, a_i, r_i^{env}, r_i^{self}, s_{i+1})\}_{i=t-k}^t$  spanning the most recent  $k$  timesteps. The VLM  $f_\phi$  then performs sequential generation to analyze this experience and update the knowledge base  $\mathcal{M}_t = \{e_1, e_2, \dots, e_{N_t}\}$ , which stores task understanding as a collection of structured rules. The memory is initialized empty ( $\mathcal{M}_0 = \emptyset$ ) and evolves through iterative meta-reasoning cycles. The generation process factorizes into two stages:

$$\begin{aligned} p_\phi(\Omega_t, \mu_t \mid \tau_{[t-H:t]}, \mathcal{M}_{t-1}) \\ = p_\phi(\mu_t \mid \tau_{[t-k:t]}, \mathcal{M}_{t-1}) \cdot p_\phi(\Omega_t \mid \mu_t, \tau_{[t-k:t]}, \mathcal{M}_{t-1}), \end{aligned} \quad (4)$$

where the process first generates a meta-analysis  $\mu_t \sim p_\phi(\cdot \mid \tau_{[t-k:t]}, \mathcal{M}_{t-1})$  that articulates observed patterns, strategies, and insights from the trajectory, formatted within  $<\text{meta}> \dots </\text{meta}>$  tags. Conditioned on this analysis, the model then generates a sequence of memory operations  $\Omega_t \sim p_\phi(\cdot \mid \mu_t, \tau_{[t-k:t]}, \mathcal{M}_{t-1})$ , where each operation takes the form  $\omega \in \{\langle \text{add} \rangle(e), \langle \text{delete} \rangle(id), \langle \text{keep} \rangle\}$ . These operations modify  $\mathcal{M}_{t-1}$  to produce the updated knowledge base  $\mathcal{M}_t$ , which is subsequently provided to the meta-reasoning module for knowledge-conditioned decision-making and to MCT-RL for policy optimization.

## 4.2. Action-Reasoning Module

As shown at the bottom of Fig. 2, the meta-reasoning module serves as the execution engine of our framework, responsible for real-time environment interaction and continuous policy refinement. It integrates metacognitive in-

sights with visual perception to adapt to unknown environments. We model the test-time adaptation problem as an unknown Markov Decision Process (MDP) defined by the tuple  $\mathcal{E} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \rho_0 \rangle$ , where  $\mathcal{S}$  denotes the pixel-level state space, with each state composed of the three latest frames of the game;  $\mathcal{A}$  is the discrete action space containing legal in-game movements;  $\mathcal{T}$  represents the unknown transition dynamics;  $r$  is the reward function, which may be sparse and difficult for credit assignment;  $\gamma$  is the discount factor; and  $\rho_0$  is the initial state distribution. The objective is to learn a policy  $\pi_\theta : \mathcal{S} \times \mathcal{M} \rightarrow \Delta(\mathcal{A})$ , parameterised by  $\theta$ , which conditions on both the observation history (state  $s_t$ ) and the dynamic knowledge memory  $\mathcal{M}$  to maximise the expected cumulative return:

$$J(\theta, \mathcal{M}) = \mathbb{E}_{\tau \sim \pi_\theta(\cdot | \mathcal{M})} \left[ \sum_{t=0}^T \gamma^t r(s_t, a_t) \right], \quad (5)$$

where  $\tau = (s_0, a_0, r_0, \dots)$  denotes an interaction trajectory.

The policy  $\pi_\theta$  is implemented via a Vision-Language Model (VLM)  $f_\theta$  that executes a two-stage autoregressive generation process. In the first stage, the VLM analyses the visual input  $s_t$  to generate structured semantic tokens  $z_t \sim \pi_\theta(\cdot | s_t)$ . The representation  $z_t = (pos_t, cat_t, dyn_t)$  encodes the game images, containing object coordinates ( $pos_t$ ), object category ( $cat_t$ ), and dynamic attributes ( $dyn_t$ ) such as moving direction. In the second stage, the VLM conditions on the generated scene  $z_t$  and the external knowledge  $\mathcal{M}$  to produce the response. This response emerges sequentially: the model first generates a reasoning chain  $r_t$  that explicates game understanding grounded in  $\mathcal{M}$  (including mechanics and rules), followed by the final executable action  $a_t$ . The decision process is thus modelled as:

$$\pi_\theta(a_t | s_t, \mathcal{M}) = \sum_{z_t \in \mathcal{Z}} \pi_\theta(a_t | z_t, \mathcal{M}) \cdot \pi_\theta(z_t | s_t), \quad (6)$$

where  $\mathcal{Z}$  is the space of structured scene representations. This formulation realises a knowledge-augmented policy where abstract game knowledge directly modulates object-level control as aligned with human metacognitive model in Fig. 1.

To equip the meta-reasoning module with robust initial capabilities, we perform SFT on our specialised reasoning dataset. We augment raw gameplay trajectories with high-quality natural language rationales generated by a teacher model. By training on these traces, comprising a reflective `<think>` analysis and an `<answer>` segment, the meta-reasoning module is initialised with strong priors for scene understanding and logical deduction before entering the test-time phase. Once the model learns the reasoning pattern, we optimise the model at the test-time both non-parametrically by injecting the memory into the

model context, and parametrically by reinforcement learning paradigm MCT-RL.

In MCT-RL, the meta-reasoning module follows a periodic learning process inspired by human metacognition, interleaving active knowledge acquisition with online optimization. This reflection further enhances the action-reasoning module's ability to effectively leverage acquired recent knowledge for problem-solving. We apply MCT-RL at every fixed interval of  $T$  steps to explore promising reasoning paths by reusing past states  $\{s_t\}_{t=|\tau|-T}^{|\tau|}$  stored in the trajectory memory  $\mathcal{D}$ . During test-time interaction, all experiences are continuously recorded in real-time into  $\mathcal{D}$ , forming an accumulated buffer  $\mathcal{D} = (s_t, a_t, r^{env}, r^{self}, s_{t+1})_{t=1}^{|\tau|}$ . This buffer jointly supports knowledge extraction in the meta-reasoning module and policy optimization in MCT-RL, enabling the agent to iteratively refine its reasoning and decision-making based on recent experiences.

To address the challenge of sparse environmental rewards in long-horizon tasks, we replace traditional sparse environmental rewards with an internal, reasoning-based signal derived through majority voting [64]. For each past state  $s_t$  at timestep  $t$ , the model samples  $K$  candidates with explicit reasoning rationales:

$$\{a_i\}_{i=1}^K \sim \prod_{i=1}^K p_\theta(\cdot | s_t, \mathcal{M}), \quad (7)$$

Candidates are evaluated via majority voting to extract the golden action:  $a_t^* = \arg \max_{a \in \mathcal{A}} \sum_{i=1}^K \mathbb{I}(a_i = a)$ . Each candidate is assigned by an indicator reward function comparing candidate with the golden action:

$$r_t(s_t, a) = \mathbb{I}(a = a_t^*) = \begin{cases} 1, & \text{if } a = a_t^*, \\ 0, & \text{otherwise.} \end{cases}$$

The policy is updated to maximise consistency with this consensus using GRPO [46] algorithm, where the final objective can be written as:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{s_t \sim S_{\mathcal{D}}[-T:], \{a_i\}_{i=1}^K \sim \pi_{\theta_{old}}(\cdot | s_t)} \left[ \frac{1}{K} \sum_{i=1}^K \frac{1}{|a_i|} \sum_{t=1}^{|a_i|} \min \left( w_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(w_{i,t}(\theta)) \hat{A}_{i,t} \right) \right], \quad (8)$$

where the importance ratio  $w_{i,t}(\theta)$  is defined as:

$$w_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t} | s_t)}{\pi_{\theta_{old}}(a_{i,t} | s_t)}, \quad (9)$$

and the advantage  $\hat{A}_{i,t}$  for action  $a_{i,t}$  is defined as:

$$\hat{A}_{i,t} = \frac{r_t(s_t, a_{i,t}) - \text{mean}(\{r_t(s_t, a_{i,t})\}_{i=1}^G)}{\text{std}(\{r_t(s_t, a_{i,t})\}_{i=1}^G)}. \quad (10)$$

Table 1. Performance comparison on 33 seen Atari games and 12 unseen games. The upper block shows results on seen games used for the SFT phase, while the lower block presents results on unseen games. The first four columns report pretrained VLMs without task-specific fine-tuning. The last columns show ablations removing: both components (w/o RL&MR), meta-reasoning (w/o MR), and MCT-RL (w/o RL). **Bold** numbers indicate the best performance for each game, while underlined numbers denote the best performance within the ablation studies. “Top-1 Count” rows report the number of games where each method achieves the highest score.

Game	Pretrained Methods				Finetuned Methods			
	Gemini-2.5	GPT-4o	Qwen2.5-VL-7B	DeepSeek-VL2	MCTR w/o RL&MR	MCTR w/o MR	MCTR w/o RL	MCTR
Alien	0	30	0	490	<b>10160</b>	/	/	/
Amidar	0	1	30	16	<b>988</b>	/	/	/
Assault	42	42	189	21	<b>399</b>	/	/	/
Asterix	850	500	550	350	<b>6100</b>	/	/	/
Atlantis	0	0	0	4000	<b>51600</b>	/	/	/
BankHeist	0	0	0	0	<b>30</b>	/	/	/
BeamRider	0	0	176	<b>528</b>	176	/	/	/
Bowling	12	0	12	3	<b>86</b>	/	/	/
Boxing	-14	<b>-12</b>	-19	-29	-882	/	/	/
Breakout	2	11	7	11	<b>50</b>	/	/	/
ChopperCommand	300	0	800	1100	<b>8700</b>	/	/	/
DemonAttack	0	0	60	<b>295</b>	40	/	/	/
DoubleDunk	<b>-2</b>	-12	<b>-2</b>	-12	-6	/	/	/
ElevatorAction	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	/	/	/
FishingDerby	-54	-54	-54	<b>-46</b>	-123	/	/	/
Gopher	0	20	20	0	<b>900</b>	/	/	/
Hero	150	0	0	0	<b>28625</b>	/	/	/
JourneyEscape	-20200	-52700	-78300	<b>-11700</b>	-53800	/	/	/
Kangaroo	0	0	0	0	<b>2000</b>	/	/	/
Krull	40	280	0	0	<b>2170</b>	/	/	/
KungFuMaster	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	/	/	/
Phoenix	220	<b>240</b>	200	80	<b>240</b>	/	/	/
Pong	-27	<b>-23</b>	-27	-26	-67	/	/	/
Qbert	0	0	0	200	<b>6850</b>	/	/	/
Riverraid	120	0	3430	960	<b>8320</b>	/	/	/
RoadRunner	0	0	0	1400	<b>42500</b>	/	/	/
SpaceInvaders	0	0	<b>525</b>	345	100	/	/	/
StarGunner	300	0	0	300	<b>1200</b>	/	/	/
TimePilot	400	400	400	600	<b>2100</b>	/	/	/
Tutankham	0	0	1	0	<b>2</b>	/	/	/
UpNDown	400	580	0	590	<b>7290</b>	/	/	/
WizardOfWor	0	0	1000	200	<b>4100</b>	/	/	/
Zaxxon	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	/	/	/
<b>Top-1 Count (seen)</b>	2/33	3/33	2/33	5/33	<b>23/33</b>	/	/	/
AirRaid	<b>1125</b>	150	0	175	125	200	75	<b>475</b>
BattleZone	2000	0	0	2000	5000	0	6000	<b>12000</b>
Berzerk	550	450	750	350	650	150	1000	<b>1100</b>
Carnival	1280	960	20	740	600	500	360	<b>2660</b>
CrazyClimber	0	0	0	0	1100	1000	1800	<b>5600</b>
Freeway	0	5	0	1	17	<b>23</b>	22	<b>23</b>
Frostbite	0	0	0	140	310	240	430	<b>650</b>
IceHockey	<b>-2</b>	<b>-2</b>	<b>-2</b>	-6	-11	-16	-8	-12
NameThisGame	0	0	590	740	1720	70	1630	<b>2170</b>
Robotank	1	0	2	4	<b>5</b>	<b>5</b>	3	<b>5</b>
Sequest	160	0	0	0	20	40	80	<b>660</b>
Skiing ↓	-6663	-6663	-6663	-6656	-20074	<b>-30057</b>	-29452	-20334
<b>Top-1 Count (unseen)</b>	2/12	1/12	1/12	0/12	1/12	3/12	0/12	<b>9/12</b>

## 5. Experiments

In this section, we conduct comprehensive experiments to evaluate MCTR’s effectiveness in test-time adaptation to novel environments. Our evaluation focuses on the Atari game benchmark, comprising 45 diverse games that challenge decision-making across varying strategic requirements and visual complexity. Specifically, Section 5.1 introduces the experimental setting, Section 5.2 presents the main results, and Section 5.3 provides a case study.

### 5.1. Experimental Setting

The MCTR begins with supervised fine-tuning of Qwen 2.5–7B–VL as the base model. We train on 33

Atari games, collecting 10,000 samples per title using DQN\_modern [21] policies in the NoFrameskip-v4 environment. Data collection operates on four stacked frames with  $\varepsilon$ -greedy or Boltzmann exploration, storing at each timestep a three-frame state representation, the executed action, the received reward, and a binary emulator snapshot. To generate language supervision for these trajectories, we first derive visual grounding through game-specific OpenCV settings that specify play regions, colour masks, and templates. These settings enable automatic extraction of bounding boxes for key entities within each frame. We then prompt Gemini 2.0 Flash with propositions parsed from these visual groundings to obtain stepwise reasoning

**Table 2. Ablation study on adaptive interval scheduling.** Comparison of different initial intervals ( $k$ ) and growth rates ( $\lambda$ ) across a subset of Atari games. Bold values indicate best performance per game.

$k$	$\lambda$	IceHockey	BattleZone	AirRaid	Frostbite	Carnival
3	1.00	-13	2000	300	310	540
9	1.00	-16	4000	100	370	360
15	1.00	<b>-12</b>	2000	0	270	420
20	1.00	<b>-12</b>	<b>6000</b>	<b>325</b>	<b>440</b>	<b>440</b>
15	1.18	<b>-8</b>	6000	0	370	300
3	0.85	<b>-12</b>	<b>12000</b>	<b>475</b>	<b>650</b>	<b>2660</b>

traces. Each trace comprises a reflective `<think>` analysis followed by an `<answer>` that restates the action, ensuring alignment with the competent control demonstrated by the open-source DQN policies. The resulting SFT samples integrate three frames, the DQN action, extracted bounding boxes, and Gemini-generated reasoning. Training proceeds for six epochs across all games. Further details on dataset construction are provided in supplementary material.

During test-time adaptation, the meta-reasoning interval  $k$  is initialised to 3 and dynamically adjusted according to  $k \leftarrow k/0.85$  after each meta-reasoning cycle, bounded within [2, 15] to balance computational cost with adaptation responsiveness. The memory bank maintains a maximum capacity of 20 entries to store recent experiences. Following this initialization, the model undergoes MCT-RL every 100 interaction steps and each MCT-RL stage runs for five epochs with a group size of 8.

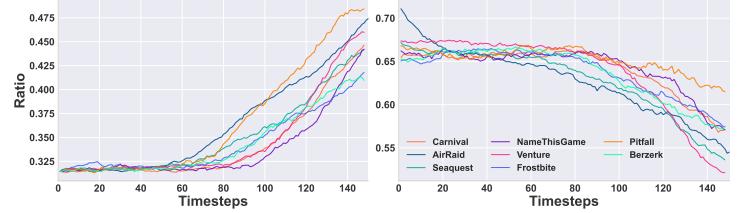
To enable efficient parameter updates during this adaptation phase, we employ Low-Rank Adaptation (LoRA) with rank-64 adapters and scaling factor  $\alpha = 32$  applied to all linear layers except the vision modules, which remain frozen to preserve visual understanding. This configuration maintains the base model’s reasoning and perception capabilities whilst enabling lightweight fine-tuning with substantially fewer trainable parameters than full model updates. The prompting templates used for reasoning and reflection are provided in the supplementary material.

## 5.2. Main Results

**Performance Comparison in Atari Games.** Table 1 presents comprehensive evaluation results across 45 Atari games, divided into 33 seen games used during supervised fine-tuning (upper block, gray) and 12 unseen games for generalisation testing (lower block, pink). We compare pre-trained vision-language models, our supervised fine-tuning baseline, and ablation variants of our MCTR approach.

The SFT (MCTR w/o RL&MR) baseline achieves the highest overall performance on seen tasks (23 out of 33 top-1 scores), demonstrating that supervised reasoning-augmented fine-tuning effectively aligns the model with expert policies, while general pre-trained VLMs show limited competence with near-zero or negative rewards. However,

**Figure 3. MCT-RL dynamics analysis.** **Left:** Majority voting ratio increases as the agent progressively favors higher-quality, self-consistent actions. **Right:** Agreement ratio with historical trajectory memory declines, indicating active policy revision with newly acquired environmental knowledge.



this strong seen performance does not transfer to novel environments, the SFT baseline achieves only 1 out of 12 top-1 scores on unseen games. In contrast, our proposed MCTR achieves the best generalization on unseen games, securing 9 out of 12 top-1 results. MCTR interleaves reflection-based meta-reasoning with MCT-RL, enabling both efficient adaptation and exploration in new environments. The performance gains are particularly pronounced in complex games: on BattleZone, MCTR achieves 12000 compared to 5000 for the SFT baseline; on CrazyClimber, it reaches 5600 versus 1100; and on Carnival, it attains 2660 versus 600. Notably, pretrained VLMs continue to struggle on unseen games (0 top-1 scores for the best baseline DeepSeek-VL2), highlighting the critical importance of test-time adaptation.

**Ablation Study.** As shown in Table 1, the last four columns reveal the contribution of each component. Removing both test-time RL and meta-reasoning (MCTR w/o RL&MR) yields strong seen performance (23/33) but poor unseen generalization (1/12), indicating supervised fine-tuning alone cannot transfer to novel environments. Adding test-time RL alone (MCTR w/o MR) improves unseen performance to 3/12 top-1 scores, while adding meta-reasoning alone (MCTR w/o RL) achieves similar results but excels in different games, as MR favours strategic adjustment (e.g., BattleZone: 6000) while MCT-RL enables policy optimisation (e.g., Freeway: 23). The complete MCTR system achieves 9/12 top-1 scores, demonstrating strong synergy: meta-cognitive reflection guides exploration during test-time RL, as evidenced by substantial improvements on BattleZone (12000 vs. 6000) and CrazyClimber (5600 vs. 1800). This integration improves average unseen performance by 275% over the SFT baseline.

**Effect of Adaptive Meta-Reasoning Scheduling.** To verify the effect of different adaptive interval scheduling strategies, we conducted ablation studies across a subset of Atari games, as shown in Table 2. The adaptive exponential schedule ( $k=3$ ,  $\lambda=0.85$ ) consistently outperforms both fixed-frequency baselines ( $\lambda=1.0$ ) and slower growth variants ( $\lambda=1.18$ ) across all tested environments, achieving particularly strong results on BattleZone (12000) and Carnival (2660). Excessively frequent meta-reasoning intervals (e.g.,

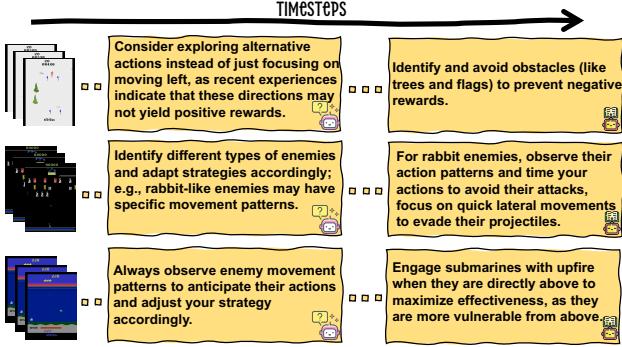


Figure 4. Case study illustrating the temporal evolution of meta-reasoning knowledge across three unseen games. Early reflections exhibit exploratory metacognition focused on hypothesis generation, while later reasoning crystallises into goal-directed, procedural strategies grounded in learnt environmental dynamics.

$k=3$ ,  $\lambda=1.0$ ) demonstrate degraded performance as insufficient experience accumulates between reflections, introducing instability without generating new strategic knowledge. Conversely, large fixed intervals (e.g.,  $k=20$ ) provide inadequate reflective guidance during early adaptation when strategic reassessment is most critical. The exponential schedule addresses both limitations: it enables dense reflection during the exploration phase when information gain is maximal, then progressively reduces frequency as the policy converges, yielding efficient test-time adaptation.

**MCT-RL Learning Dynamics Analysis.** Fig. 3 shows how MCT-RL enhances adaptive decision-making capabilities during test-time adaptation. The left panel illustrates the evolution of majority voting reward, showing that the framework progressively favors reasoning traces that produce higher-quality actions with stronger internal agreement. Conversely, the right panel reveals declining agreement between action regeneration in the MCT-RL phase-informed by newly acquired knowledge and historically selected actions stored in trajectory memory that were generated using outdated environmental understanding. This decreasing alignment signals active policy revision as the agent continuously refines its decision-making strategy based on accumulated experience. Together, these dynamics show that MCT-RL achieves dual objectives: it maintains reasoning stability through multi-path consensus while simultaneously enabling genuine policy adaptation in response to evolving environmental knowledge.

### 5.3. Case Study

Fig. 4 presents a temporal analysis of meta-reasoning knowledge evolution across three unseen games, revealing a systematic progression from metacognitive exploration to situated expertise. The cases capture a typical transition in reasoning abstraction that parallels human skill acquisition.

Early-stage reasoning (left column) exhibits character-

istic features of metacognitive monitoring under uncertainty. The module generates high-level epistemic directives, “identify different types of enemies” and “consider exploring alternative actions”, that function as learning objectives rather than action policies. This phase demonstrates active hypothesis generation, where the agent constructs a problem space by deliberately probing environmental contingencies. Notably, these early reflections emphasise what to learn rather than how to act, indicating metacognitive awareness of knowledge gaps.

Mid-to-late stage reasoning (right column) reveals a qualitative shift toward procedural expertise and situated decision-making. Abstract principles crystallize into concrete, context-specific strategies: precise timing patterns for rabbit enemies, spatial positioning tactics for submarine engagement, and obstacle avoidance heuristics. This progression reflects the compilation of declarative knowledge into procedural schemas, as knowledge becomes increasingly grounded in sensorimotor contingencies and reward structures specific to each game’s dynamics.

This epistemic trajectory (from exploratory metacognition to goal-directed proceduralization) mirrors the cognitive transition observed in human skill acquisition, where initial controlled processing gives way to automated expertise. The meta-reasoning module thus exhibits genuine adaptive intelligence: it not only learns what patterns exist but dynamically regulates how to learn them, embodying a test-time instantiation of metacognitive model.

## 6. Conclusion

We presented Meta-Cognitive Test-Time Reasoning, a framework that equips vision-language models with human-like adaptive capabilities through structured meta-reasoning and experience-driven learning during inference. Drawing on human metacognition, MCTR implements a two-level architecture where a meta-reasoning module discovers and consolidates strategic knowledge from experience trajectories, while an action-reasoning module applies this knowledge for task execution and refines its policy online through our proposed metacognitive test-time reinforcement learning. MCTR exhibits strong test-time generalization on the Atari benchmark, securing 9 out of 12 top-1 scores on unseen games. Ablation studies confirm that meta-reasoning and test-time learning provide complementary adaptation mechanisms: the former extracts transferable strategic patterns while the latter enables policy refinement through environmental feedback. Learning dynamics analysis demonstrates that MCT-RL balances reasoning stability through multi-path consensus with genuine policy adaptation, as evidenced by increasing internal agreement among current reasoning traces and declining alignment with historical actions. Qualitative analysis of reasoning traces reveals that strategies evolve from exploratory trial-and-error to structured, goal-directed patterns that mirror human adaptation to unfamiliar tasks.

## References

- [1] Samuel Acquaviva et al. Communicating natural programs to humans and machines. *NeurIPS*, 2022. 2
- [2] Michael Ahn et al. Do as i can, not as i say: Grounding language in robotic affordances. *CoRL*, 2022. 2
- [3] Ekin Akyurek et al. Surprising effectiveness of test-time training for abstract reasoning. *NeurIPS*, 2024. 1, 2, 3
- [4] Anthony Brohan et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *CoRL*, 2023. 2
- [5] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-v>, 2025. Accessed: 2025-02-02. 1
- [6] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 1, 2
- [7] François Chollet and Marcus Hutter. Arc-agı: A benchmark for artificial general intelligence. *arXiv preprint*, 2024. 2
- [8] Michael T Cox and Anita Raja. Metacognition in computation: A selected research review. *Artificial Intelligence*, 169(2):104–141, 2005. 3
- [9] DeepSeek-AI. Deepseek-r1: Towards deep reinforcement of reasoning in large language models. *arXiv preprint arXiv:2501.01717*, 2025. 1, 2
- [10] DeepSeek-AI, Aixin Liu, Peitian Bei, Bing Hu, Bingxuan Liu, Bochao Wu, Chengda Cao, Chenggang Cheng, Chengqi Chen, Chong Dai, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1
- [11] Daniel L Dinsmore et al. Metacognition in education: Translating theory into practice. *Educational Psychology Review*, 2023. 2
- [12] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128, 2024. 1, 2
- [13] Danny Driess et al. Palm-e: An embodied multimodal language model. *ICML*, 2023. 2
- [14] Yuqing Du et al. Guiding pretraining in reinforcement learning with large language models. *ICML*, 2023. 2
- [15] Anastasia Efklides. Metacognition and affect: The role of feelings in self-regulated learning. *Educational Psychologist*, 58(3):202–219, 2023. 2
- [16] Chelsea Finn et al. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 3
- [17] Damien S. Fleur, Bert Bredeweg, and Wouter van den Bos. Metacognition: ideas and insights from neuro- and educational sciences. *npj Science of Learning*, 6(1):13, 2021. 1, 2
- [18] ARC Prize Foundation. Arc-agı benchmark. <https://arcprize.org/arc-agı>, 2025. 1, 2
- [19] Yossi Gandelsman et al. Test-time training with masked autoencoders. *NeurIPS*, 2022. 2
- [20] Fausto Giunchiglia et al. Fast, slow, and metacognitive thinking in ai. *npj Artificial Intelligence*, 2025. 3
- [21] Florin Gogianu, Tudor Berariu, Lucian Buşoniu, and Elena Burceanu. Atari agents, 2022. 6
- [22] Ryan Greenblatt et al. Language models can solve the abstraction and reasoning corpus. *arXiv preprint*, 2024. 2
- [23] Shibo Hao et al. Reasoning with language model is planning with world model. *EMNLP*, 2023. 2
- [24] Tao He et al. Leveraging dual process theory in language agent framework for real-time simultaneous human-ai collaboration. *arXiv preprint arXiv:2502.11882*, 2025. 3
- [25] Mikel Hodel et al. Solving the abstraction and reasoning corpus with neural program synthesis. *arXiv preprint*, 2023. 2
- [26] Edward Hu et al. Test-time language model adaptation via perplexity minimization. *arXiv preprint*, 2025. 2
- [27] Wenlong Huang et al. Voxposer: Composable 3d value maps for robotic manipulation with language models. *CoRL*, 2023. 2
- [28] Li Ji-An et al. Language models are capable of metacognitive monitoring and control of their internal activations. *arXiv preprint arXiv:2505.13763*, 2025. 3
- [29] Zhengbao Jiang, Weijia Yuan, et al. Role-play with large language models. *arXiv preprint arXiv:2310.01798*, 2023. 1
- [30] Ben Johnson. Metacognition for artificial intelligence system safety: An approach to safe and desired behavior. *Safety Science*, 151:105743, 2022. 3
- [31] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011. 2
- [32] Brenden M Lake et al. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017. 2
- [33] Lisa Lee et al. Vla: Vision-language-action models for robotic manipulation. *arXiv preprint*, 2023. 2
- [34] Jiachang Liu et al. What makes good in-context examples for gpt-3? *ACL Workshop*, 2023. 1, 2
- [35] Jinghan Liu et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025. 3
- [36] Aman Madaan et al. Self-refine: Iterative refinement with self-feedback. *NeurIPS*, 2023. 3
- [37] Thomas O Nelson and Louis Narens. Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26:125–173, 1990. 1, 2, 3
- [38] Alex Nichol et al. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 3
- [39] Shuaicheng Niu et al. Efficient test-time model adaptation without forgetting. *ICML*, 2022. 2
- [40] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. 1
- [41] OpenAI. Openai o1 system card. *Technical Report*, 2024. 1, 2
- [42] Joon Sung Park et al. Metacognition is all you need? using introspection in generative agents to improve goal-directed behavior. *arXiv preprint arXiv:2401.10910*, 2024. 3

- [43] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1
- [44] Michelle Rivers. Metacognition about practice testing: a review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review*, 2021. 1, 2
- [45] Claudia M. Roebers. Executive function and metacognition: Towards a unifying framework. *Developmental Review*, 2017. 2
- [46] 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
- [47] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024. 4
- [48] Noah Shinn et al. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*, 2023. 3
- [49] Manli Shu, Weili Nie, De-An Huang, et al. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022. 1, 2
- [50] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 1, 2
- [51] Yu Sun, Xiaolong Wang, Zhuang Liu, et al. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 1, 2
- [52] Yuhao Sun et al. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025. 3
- [53] Qwen Team. Qwen2.5-reasoner: Enhancing llm reasoning via verbal correction and process supervision. *arXiv preprint arXiv:2501.00235*, 2025. 1, 2
- [54] Dequan Wang et al. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 2
- [55] 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. 1
- [56] Taylor Webb, Keith Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 2023. 1
- [57] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 1, 2
- [58] Xiang Wei et al. Reasoning on a spectrum: Aligning llms to system 1 and system 2 thinking. *arXiv preprint arXiv:2502.12470*, 2025. 3
- [59] Fengli Xu, Qianyue Hao, Zefang Zong, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025. 1, 2
- [60] Cheng Yang et al. Llm2: Let large language models harness system 2 reasoning. *arXiv preprint arXiv:2412.20372*, 2024. 3
- [61] Shunyu Yao, Dian Cui, Shunyan Zhao, et al. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. 1, 2
- [62] Zixiang Zhang et al. Rest: Reinforced self-training for language modeling. *arXiv preprint*, 2024. 2
- [63] Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, and Jun Wang. Memento: Fine-tuning llm agents without fine-tuning llms, 2025. 1
- [64] Yuxin Zuo, Kaiyang Zhang, Shang Qu, et al. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025. 1, 2, 3, 5

# Adapting Like Humans: A Metacognitive Agent with Test-time Reasoning

## Supplementary Material

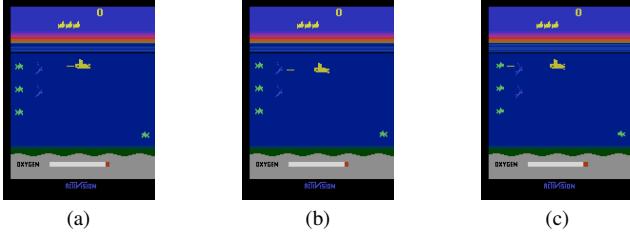


Figure 5. Example of an observation which contains three images, from Atari game Seaquest.

## 7. Supervised Fine-tuning

We build our supervised fine-tuning pipeline on top of the R1-V framework [5], with training details specified in the following sections.

### 7.1. Reasoning Dataset Creation

**Data Collection.** Existing Atari corpora rarely align step-level vision, text, and action or preserve restorable context, so we build a dataset that does. It provides images, parser summaries, and action-affirming rationales for studies of perception-to-decision reasoning and test-time adaptation. We release a stable schema, fixed splits, loaders, and documentation, and supply hashes where snapshots are restricted, enabling reproducible training and fair validation.

We collect trajectories in Atari environments configured as NoFrameskip-v4. A DQN<sub>modern</sub> policy operates on stacks of three consecutive frames (Figure 5) and selects discrete actions under either  $\varepsilon$ -greedy or Boltzmann exploration with a small default exploration rate. At the beginning of each episode the environment is reset. At every step we render the RGB frame, update a buffer containing the three most recent frames, and stack observations as a channel-first tensor for policy input.

We log the action identifier and its human-readable label, the immediate reward, and the running return. To support exact restoration during later analysis we archive a binary snapshot of the emulator runtime state. We maintain per-episode identifiers, step counters, and throughput statistics, and write summary files reporting average episode length and return. After the interaction budget is exhausted, we assemble the per-step records into Hugging Face DatasetDict objects with explicit train/test splits, cast image columns to the native image type, and save them to disk. The same rollout procedure builds both the super-

vised fine-tuning corpus and the reinforcement learning corpus; only the choice of games, interaction budget, and split assignment differ. Each game is accompanied by a version-controlled OpenCV configuration specifying the legal play region, color masks, object templates, thresholds, morphological kernels, and fixed UI counters for downstream parsing.

**Language Reasoning with Gemini.** For each recorded step we load the game-specific OpenCV configuration and parse the most recent frame within the legal region using template matching. Detected elements are stored and converted into concise natural language propositions describing class, location, and value when applicable. A short-term buffer retains the three most recent descriptions to provide immediate temporal context.

Using these descriptions together with the action label from DQN, we compose a prompt with fixed system and user instructions and query Gemini-2.0-Flash to obtain a step-level reasoning trace. The model returns a structured output comprising a reflective `<think>` segment and an `<answer>` segment that restates the chosen action, thereby pairing perception, interpretation, and decision as a single supervision signal. The complete prompt for Gemini is as below:

```
You are shown three consecutive video
frames from an Atari-style game.
An oracle has already chosen its next move:
{action}.

Your task is to adopt the oracle's
perspective and articulate the reasoning
behind this decision. **Do not** mention or
describe the action itself in your output.

Begin by analyzing the visual understanding
of each frame. For each timestep, you will
be given a description of detected objects
and their bounding boxes in the format:
(x_top_left, y_top_left, x_bottom_right,
y_bottom_right). All coordinates use a top-
left origin, with the frame size being 210
pixels in height and 160 pixels in width.

- Frame 1 (timestep -2):
{understanding_buffer[0]}

- Frame 2 (timestep -1):
{understanding_buffer[1]}
```

```
- Frame 3 (current timestep):  
{understanding_buffer[2]}
```

Next, write a concise thought process-no more than 10 sentences-that logically explains why the oracle made this decision. Your reasoning should be natural and grounded in the visual context, and you could reason it from following aspects but not limited to:

- **Game Genre Inference**: Based solely on visual cues, infer whether the game is action, shooting, strategy, or another genre.
- **Gameplay Mechanics**: Describe how the agent appears to interact with the environment (e.g., movement, avoidance, targeting) and how feedback or rewards might be structured.
- **Key Observations**: Point out any significant objects, obstacles, goals, or patterns of movement across the frames that likely influenced the decision.
- **Goal Analysis**: Explain what the agent seems to be trying to achieve immediately and over the long term.
- **Decision Justification**: Summarize why the inferred action is appropriate given the current game context-without stating or implying the action.

Avoid referencing the action name, any game titles, franchises, or platforms. Focus entirely on visual reasoning and inferred gameplay logic.

Please directly output your reasoning without any additional text or formatting.

We append these annotations to the rollout records and package them as Hugging Face datasets with train/test splits. Each sample contains the image sequence, actions, rewards, episode metadata, the restorable emulator state, the textual state descriptions, and the associated reasoning. The same parsing, prompting, and storage settings are applied in both the supervised fine-tuning and reinforcement learning stages so that data semantics and annotation protocols remain consistent.

## 7.2. Example Data

System prompt:

You are a general-purpose game agent focused on scene understanding, key object recognition, and intelligent action decision-making based on continuous game frames. Please analyze and reason based on the following dimensions:

**Scene Analysis**: First, Extract dynamic information from the current frame and several previous frames to analyze the following:

- **Object Categories**: Identify entities such as the agent/player, enemies, obstacles, bullets, items, etc.
- **Spatial Positions**: For each object, use the bounding box format (`x_top_left`, `y_top_left`, `x_bottom_right`, `y_bottom_right`). The coordinate origin (0,0) is at the top-left of the frame.

**Game Reasoning**: Next, Based on the overall game context and the current scene, infer the most reasonable next atomic action. Your reasoning should integrate both static game rules and dynamic scene information, and potentially cover the following aspects :

- **Game Genre Inference**: Based solely on visual cues, infer whether the game is action, shooting, strategy, or another genre.
- **Gameplay Mechanics**: Describe how the agent appears to interact with the environment (e.g., movement, avoidance, targeting) and how feedback or rewards might be structured.
- **Key Observations**: Point out any significant objects, obstacles, goals, or patterns of movement across the frames that likely influenced the decision.
- **Goal Analysis**: Explain what the agent seems to be trying to achieve immediately and over the long term.
- **Decision Justification**: Summarize why the inferred action is appropriate given the current game context-without stating or implying the action.

**Action Decision**: Based on the scene understanding and inferred intent, choose the most reasonable next atomic action from the following valid action set:

NOOP, FIRE, UP, RIGHT, LEFT, DOWN, UPRIGHT, UPLEFT, DOWNRIGHT, DOWNLEFT, UPFIRE, RIGHTFIRE, LEFTFIRE, DOWNFIRE, UPRIGHTFIRE, UPLEFTFIRE, DOWNRIGHTFIRE, DOWNLEFTFIRE.

Instructions:

- Base your reasoning on multiple frames to capture temporal patterns and object interactions.
- Ensure your decision aligns with inferred goals and basic game logic.
- Wrap your reasoning in `<think>...</think>` tags.
- Wrap your final action decision in `<answer>...</answer>` tags, containing only the action name.

Problem prompt:

You are given a sequence of consecutive video game frames (observations).

Your task is to analyze the visual scene and reason about the game's dynamics to infer the most appropriate next action.

- Provide your reasoning based Scene Analysis and Game Reasoning dimensions, inside '`<think>...</think>`' tags.
- Then, output the \*\*final Action Decision \*\* inside '`<answer>...</answer>`' tags.

Choose the action from the following valid set:  
 'NOOP', 'FIRE', 'UP', 'RIGHT', 'LEFT', 'DOWN', 'UPRIGHT', 'UPLEFT', 'DOWNRIGHT', 'DOWNLEFT', 'UPFIRE', 'RIGHTFIRE', 'LEFTFIRE', 'DNDFIRE', 'UPRIGHTFIRE', 'UPLEFTFIRE', 'DOWNRIGHTFIRE', 'DOWNLEFTFIRE'.

Example answer:

```
<think>Let me go with understanding the current state. The main game area spans from coordinates (0, 104) to (159, 171). I see a player at (20, 137, 28, 174). a ball at (27, 147, 31, 157). a target at (133, 119, 135, 123). a target at (129, 125, 131, 129). a target at (125, 131, 127, 135). a target at (133, 131, 135, 135). a target at (121, 137, 123, 141). a target at (129, 137, 131, 141). a target at (125, 143, 127, 147). a target at (133, 143, 135, 147). a target at (129, 149, 131, 153). a target at (133, 155, 135, 159). a score at (38, 5, 125, 18). a time at (29, 16, 61, 35).
```

Then, Let me consider what the optimal next action would be.

The game appears to be about targeting or hitting specific objects. The player is positioned on the left, with a collection of objects on the right side of the screen.

The targets remain in the same position. There is a score and a timer. It seems the goal is to hit those targets within the given time to increase the score. The player is close to the baseline, and the relative positions of the player and the target have not changed. A change in the movement might not be optimal at this time  
`.</think>`  
`<answer>UPFIRE</answer>`

### 7.3. Hyperparameters

Hyperparameter	Value
Learning rate	$2 \times 10^{-5}$
Training epochs	10
Max sequence length	4096
Per-device batch size	16
Gradient accumulation steps	4
Precision	b16
Packing	Enabled
Gradient checkpointing	Enabled
Distributed training	ZeRO-3 (via Accelerate)

Table 3. Core hyperparameters used for supervised fine-tuning.

## 8. Meta-reasoning

### 8.1. Hyperparameters

Hyperparameter	Value
Memory capacity	20
Max reasoning tokens	1024
Initial interval $k_{\text{init}}$	3
Minimum interval $k_{\text{min}}$	3
Maximum interval $k_{\text{max}}$	15
Meta-reasoning model	gpt-4o-mini

Table 4. Key hyperparameters for the Meta-Reasoning Module.

### 8.2. Prompt

```
You are managing a memory system of rule (current: {current_count}/{max_capacity} rule).
```

```

Analyze the recent trajectory and current
rules to decide memory operations.
Return meta reasoning process in <meta>
reason about the memory operations you are
going to make</meta>
followed by arbitrary number of operations
using these tags:
<add> Some knowledge about the game
mechanics, or good strategies summarised
from the experience with reasoning </add> -
to add a new rule (if space available)
<delete>rule_id</delete> - to remove rule
by ID (0-based index)
<keep/> - to make no changes

rules should be the game mechanisms and/or
strategies learned from the experience.
Avoid duplicates and contradictions.
To update a rule, use <delete>old_id</
delete> then <add>new rule</add>.
If no changes needed, use <keep/>.

```

## 9. Metacognitive Test-time Reinforcement Learning

### 9.1. Hyperparameters

We build our test-time reinforcement learning pipeline on top of the VERL framework [47], leveraging its GRPO training pipeline while integrating our self-implemented TTRL reward mechanism.

Hyperparameter	Value
Reward manager	TTRL
Algorithm	GRPO
LoRA rank	64
LoRA $\alpha$	32
LoRA target modules	all linear layers (non-visual)
Max prompt length	2048
Max response length	2048
Train batch size	8
PPO mini-batch size	8
Micro-batch size / GPU	4
Learning rate	$3 \times 10^{-6}$
Rollout group size $n$	8
Samples per prompt (reward)	2
Entropy coefficient	0
KL loss in actor	disabled
KL in reward	disabled
Total epochs	5

Table 5. Key hyperparameters used for test-time reinforcement learning with MCT-RL.