

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Constructive Memory for Decision Making: Offline Reinforcement Learning with Value-Guided Action Reconstruction

Permalink

<https://escholarship.org/uc/item/13n7h7wj>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Ying, Yuancong

Liu, Quan

Wu, Lan

Publication Date

2026-09-01

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Constructive Memory for Decision Making: Offline Reinforcement Learning with Value-Guided Action Reconstruction

Yuancong Ying^{†1}, Quan Liu* (quanliu@suda.edu.cn)¹ & Lan Wu^{‡1}

¹School of Computer Science and Technology, Soochow University

Abstract

Biological agents refine behavioral strategies through offline consolidation without real time environmental interaction. Cognitive science suggests this capability stems from constructive memory, where the brain actively reconstructs the past guided by current goals rather than mechanically replaying it. Drawing on this, we propose the Value Guided Action Reconstruction model to operationalize this mechanism for offline reinforcement learning. Specifically, we model policy learning as a value guided generative process. By injecting value signals into a variational autoencoder latent space, we simulate mental rehearsal prioritizing high value behaviors. To mitigate memory distortion risks, we introduce a metacognitive confidence monitoring mechanism. This module dynamically regulates value guidance intensity by evaluating the discrepancy between reconstructed actions and original traces, thereby balancing optimization with fidelity. Empirical evaluations on MuJoCo and Maze2d tasks demonstrate that our model enhances robustness, highlighting the computational advantages of constructive memory in facilitating offline generalization.

Keywords: Constructive Memory; Offline Reinforcement Learning; Generative Models; Metacognitive Confidence Monitoring

Introduction

Learning occurs not only during real-time interaction with the environment but also during intervals of rest and sleep. This phenomenon known as offline consolidation constitutes a critical phase for biological agents to optimize strategies. Biological evidence suggests that hippocampal replay during rest is not a mechanical video-like playback of past experiences but rather a constructive simulation process aimed at modeling future events (Daou & Lee, 2025; Schacter & Adis, 2007; Schacter & Thakral, 2024). During this process, memory retrieval is strongly biased by value signals. Recent neuroscience research confirms that dopamine signals prioritizing high-value replays are crucial for future decision-making (Diekmann & Cheng, 2023; Roscow et al., 2025), which is computationally equivalent to prioritized experience replay (Mattar & Daw, 2018). Through this internal mental rehearsal, agents can distill optimized decision strategies from limited experiences without external feedback to achieve rapid adaptation and generalization (Whittington et al., 2020).

Offline Reinforcement Learning provides a computational framework for this biological capability by attempting to learn optimal policies solely from fixed historical episodic memories. However, existing computational models face a fundamental dilemma between fidelity to memory and policy optimization. On one end of the spectrum, whether it is sequence modeling approaches represented by Decision Transformer (L. Chen et al., 2021), Elastic Decision Transformer (Wu et al., 2023) and Decision ConvFormer (J. Kim et al., 2023) or correction-based value iteration algorithms like SCAS (Mao et al., 2024), there is a distinct tendency towards rigid adherence to historical data distributions. While the former improves predictive capability by optimizing context representation and the latter ensures safety by suppressing out-of-distribution actions, their core paradigms confine the agent within the safe zone of the support set which effectively avoids extrapolation errors. However, this excessive conservatism restricts the exploration of potential optimal solutions and results in a lack of constructive capacity to reorganize suboptimal memory fragments into optimal strategies. On the other end, emerging generative methods such as Diffusion-QL (Wang et al., 2022) and IDQL (Hansen-Estruch et al., 2023) attempt to leverage the powerful expressivity of diffusion models to construct complex policies guided by value functions. Although these methods break free from rigid imitation, they lack adaptive assessment of sample reliability. Consequently, they are prone to succumbing to cognitive illusions under aggressive guidance from the critic which leads to the generation of high-risk actions.

Drawing inspiration from constructive memory theories, we propose the Value-Guided Action Reconstruction (VGAR) model. Unlike approaches that view policy learning as simple function fitting, VGAR formalizes it as a value-guided generative process (Zheng et al., 2025). First, we formalize the latent reconstructive nature of memory using a Conditional Variational Autoencoder (CVAE) where we impose top-down value guidance during the mental rehearsal of the agent. This mechanism compels the generative model to prioritize the construction of high-value behavioral schemas rather than mechanically replicating suboptimal historical traces. Second, to mitigate the risk of memory distortion, we introduce a metacognitive confidence monitoring module. By measuring the discrepancy between generated actions and original memory traces, this mechanism dynamically downregulates value signals when deviation is excessive effectively reverting

[†]First author.

[‡]Second author.

*Corresponding author.

to a conservative strategy. Empirical results on MuJoCo and Maze2d tasks demonstrate that VGAR maintains competitive baseline performance while exhibiting distinct advantages in handling noisy memories. These findings provide computational support for the view that constructive reconstruction serves not as a systemic flaw but as an adaptive mechanism facilitating offline generalization and policy stitching (S. Kim et al., 2024).

Preliminaries

Problem Formulation: Offline RL as Memory Replay

We first formalize the environmental interaction of the agent within a standard Markov Decision Process (MDP) framework, represented as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. Here, $s \in \mathcal{S}$ represents the perceptual state, $a \in \mathcal{A}$ denotes the action, $\mathcal{P}(s'|s, a)$ describes the environmental dynamics, $\mathcal{R}(s, a)$ is the reward function, and $\gamma \in [0, 1)$ serves as the discount factor. The objective of the agent is to maximize the expected cumulative discounted reward:

$$J(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right]. \quad (1)$$

In the offline setting, the agent cannot interact with the environment $\mathcal{P}$ in real time and is restricted to a static dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$ generated by a mixture behavior policy π_β (Levine et al., 2020).

Within this framework, we conceptualize the offline dataset $\mathcal{D}$ as an episodic memory buffer storing traces of past perception-action interactions. Consequently, offline policy optimization functions essentially as a constructive replay mechanism, where the agent retrieves and recombines memory traces in a state detached from external inputs. This process aims to distill behavioral schemas capable of maximizing future adaptability, specifically the optimal policy π^* , from suboptimal historical experiences (Ghosh & Gilboa, 2014). Such policy improvement requires the agent to overcome memory incompleteness to achieve effective behavioral generalization while maintaining fidelity to the valid underlying dynamics.

Value Estimation via Asymmetric Error Encoding

To implement the aforementioned memory consolidation process, we adopt an Actor-Critic architecture where the critic assesses value and the actor executes selection. We utilize parameterized neural networks $V_\phi(s)$ and $Q_\theta(s, a)$ to constitute the internal evaluation model, assessing the long-term value of states and actions.

Consistent with the asymmetric encoding characteristics of reward prediction errors observed in dopamine neurons (Dabney et al., 2020), we leverage expectile regression (Kostrikov et al., 2022) to update the state value network V_ϕ :

$$\mathcal{L}_\phi = \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[L_2^\tau (Q_\theta(s, a) - V_\phi(s)) \right], \quad (2)$$

where $L_2^\tau(u) = |\tau - \mathbb{I}[u < 0]| \cdot u^2$ represents the asymmetric squared loss function. The expectile parameter $\tau \in (0, 1)$ modulates the relative weight the system assigns to positive versus negative prediction errors, corresponding to signals that are better or worse than expected. Simultaneously, the action value network Q_θ is updated by minimizing the temporal difference error:

$$\mathcal{L}_\theta = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[(Q_\theta(s, a) - (r + \gamma V_\phi(s')))^2 \right]. \quad (3)$$

The actor is parameterized as a policy network $\pi(a | s)$, acting as the executive entity. Unlike rote memorization, its goal is to extract generalized behavioral patterns from the memory. This implies that the agent not only retains past trajectories but also masters generalizable action rules, enabling adaptive responses even when encountering novel states structurally similar to past experiences.

Constructive Memory Framework

Inspired by the biological mechanism of constructive memory consolidation, this paper proposes the VGAR framework as illustrated in Figure 1. The core philosophy of VGAR lies in redefining offline data from static constraints that must be strictly adhered to into raw materials for reconstructing latent high-value behaviors. Unlike methods that strictly constrain policies within the original behavioral distribution to avoid risks, VGAR posits that offline reinforcement learning is an active constructive process. In this view, the model utilizes generative models to extract elements from past experiences and recombines them into superior policies based on value signals. To prevent this reconstruction from deviating from original experiences and producing hallucinations, we further introduce a metacognitive confidence monitoring module designed to dynamically regulate the confidence level of internal simulations.

Value-Guided Reconstruction

Even during the resting state, biological memory systems function not as static storage mechanisms but rather as future-oriented generative models. We utilize a conditional variational autoencoder to model this latent memory space. While standard CVAEs aim to minimize reconstruction error to faithfully restore data, memory retrieval during policy optimization should serve future adaptive goals. Consequently, we introduce value guidance into the generative objective. We inject the Q-value function as a value-guided feedback signal into the decoding process of the latent space. This compels the generative model not merely to recover the original action a but to prioritize reconstructing behaviors $\tilde{a}$ that yield higher expected returns in the current state s . We formalize

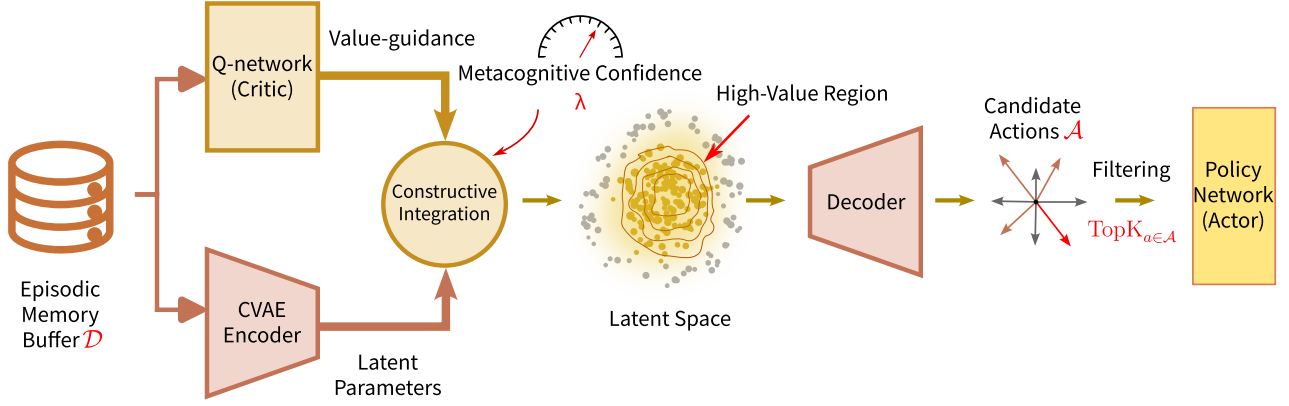


Figure 1: The framework of VGAR, which formalizes the principle of constructive memory for offline planning. The architecture integrates raw memory traces with value assessment, where the CVAE Encoder processes sensory inputs while the Critic provides value-guided signals. These streams merge in the Constructive Integration module to facilitate memory reorganization. Crucially, the Metacognitive Confidence Monitoring mechanism serves as a regulatory gate for the Latent Space, ensuring that the reconstruction prioritizes the High-Value Region associated with optimal outcomes. Finally, the Decoder translates these refined latent plans into Candidate Actions, which undergo filtering to guide the policy optimization.

this mechanism as the joint optimization objective given by

$$\mathcal{L}_{\text{Total}} = \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[\underbrace{(a - \tilde{a})^2}_{\text{Memory Fidelity}} - \lambda \cdot \underbrace{Q(s, \tilde{a})}_{\text{Value Guidance}} \right] + \frac{1}{2} D_{KL}(\mathcal{N}(\mu, \sigma) \parallel \mathcal{N}(0, 1)). \quad (4)$$

In this equation, the first term represents memory fidelity which ensures that reconstruction does not deviate from the original experience. The second term represents value guidance where the metacognitive confidence λ drives behavior optimization towards high-value regions. Finally, the third term D_{KL} serves as a latent space regularization term that constrains the posterior distribution to approximate a standard normal distribution to prevent overfitting to limited experiences.

Metacognitive Confidence Monitoring

In this constructive reconstruction process, the core risk lies in the emergence of memory distortion. To maximize expected value signals, the reconstructed actions may severely deviate from the distribution of original experiences. Since value networks are often unreliable in out-of-distribution regions, these distorted behaviors typically correspond to value overestimation and high risks. To address this issue, we implement a metacognitive monitoring mechanism within the model that simulates the process by which the brain evaluates confidence in recalled content (Fleming & Daw, 2017). This mechanism monitors the difference between the reconstructed action $\tilde{a}$ and the original experience a in real time, and transforms this discrepancy into a metacognitive confidence λ defined as

$$\lambda(s, a, \tilde{a}) = c \cdot e^{-\|a - \tilde{a}\|_2}, \quad (5)$$

where c represents the baseline confidence coefficient determining the maximum intensity of value guidance permitted under the limit of zero distortion. This mechanism establishes a dynamic negative feedback loop. As the reconstructed action severely deviates from the original memory, measured by $\|a - \tilde{a}\|_2$, the model identifies a heightened risk of hallucination, causing a corresponding decrease in metacognitive confidence. This automatically suppresses the injection of value guidance signals and forces the generation process to revert to a more conservative mode of memory fidelity. Conversely, when the reconstruction is highly consistent with original experience, the model deems the current memory credible and permits the intervention of stronger value signals to optimize the policy.

Policy Consolidation Process

The final policy learning process simulates the consolidation loop moving from episodic mental rehearsal to behavioral schemas. As illustrated in Algorithm 1, this process comprises three synergistic cognitive steps.

(i) The agent first updates its internal evaluation model using asymmetric loss based on reward signals from the episodic memory buffer $\mathcal{D}$ to establish a robust value baseline.

(ii) Subsequently, the agent executes mental rehearsal. It utilizes the CVAE to generate a series of candidate actions $\mathcal{A} = \{\tilde{a}_i\}_{i=1}^n$ within the latent space and calculates the metacognitive confidence λ as defined in Equation 5 in real time to regulate the intensity of value guidance. Then, it filters out the k actions with the highest expected returns from the generation pool and calculates their average Q-value given by

$$\mathcal{A}_{\text{top-}k} = \text{TopK}_{a \in \mathcal{A}} Q(s, a), \quad (6)$$

Algorithm 1 VGAR: Value-Guided Action Reconstruction

Input: Episodic Memory Buffer $\mathcal{D}$, asymmetric loss coefficient τ , simulation samples n , rehearsal size k .

- 1: Initialize CVAE $\mathcal{G} = \{E_\psi, D_\xi\}$, Critic Network V_ϕ, Q_θ , and Policy Network π_ϕ .
 - 2: **for** each consolidation step **do**
 - 3: Sample batch of memory traces (s, a, r, s') from $\mathcal{D}$.
 - 4: **// Phase I**
 - 5: Update Value function V_ϕ according to Eq. (2).
 - 6: Update Q-function Q_θ according to Eq. (3).
 - 7: Encode inputs: $\mu, \sigma = E_\psi(s, a)$.
 - 8: Sample latent variable: $z \sim \mathcal{N}(\mu, \sigma)$.
 - 9: Reconstruct action: $\tilde{a}_{\text{recon}} = D_\xi(s, z)$.
 - 10: Compute metacognitive confidence $\lambda(s, a, \tilde{a}_{\text{recon}})$ via Eq. (5).
 - 11: Update CVAE parameters ψ, ξ by minimizing $\mathcal{L}_{\text{Total}}$ in Eq (4).
 - 12: **// Phase II**
 - 13: Generate n candidate actions:
 $\mathcal{A} = \{\tilde{a}_i \sim D_\xi(s, z_i)\}_{i=1}^n$.
 - 14: Select top- k high-value candidates according to Eq. (6).
 - 15: Compute $\bar{Q}_{\text{top-}k}$ according to Eq. (7).
 - 16: **// Phase III**
 - 17: Sample target action: $\tilde{a} \sim \text{Uniform}(\mathcal{A}_{\text{top-}k})$.
 - 18: Update Policy π_ϕ (Eq. (8)).
 - 19: **end for**
-

$$\bar{Q}_{\text{top-}k} = \frac{1}{k} \sum_{a \in \mathcal{A}_{\text{top-}k}(s)} Q(s, a). \quad (7)$$

(iii) Finally, the agent integrates these preferred actions via the parameterized policy network π_ϕ . Specifically, we randomly sample a representative action $\tilde{a}$ from the high-value set $\mathcal{A}_{\text{top-}k}$ which effectively enhances robustness while preserving the exploration potential of the policy. Based on this sampled action, we utilize advantage-weighted regression (Peng et al., 2019) to update the policy network parameters defined as

$$\mathcal{L}_\phi = -e^{\alpha(\bar{Q}_{\text{top-}k} - V(s))} \log \pi_\phi(\tilde{a} | s), \quad (8)$$

where α serves as the temperature coefficient. By minimizing the objective in Equation 8, the policy network π learns those reconstructed behaviors possessing high value.

Experiments

Experimental Setup

We evaluated VGAR on the D4RL benchmark suite, focusing on two task domains with distinct cognitive demands.

The Gym-MuJoCo domain comprises three environments, namely *Hopper*, *Walker2d*, and *HalfCheetah*. These tasks simulate the fine motor control of biological organisms that require the agent to coordinate multiple joints under complex dynamic constraints to achieve stable locomotion. We treat the datasets of varying quality provided by D4RL, such

as *random*, *medium*, *medium-replay*, and *medium-expert*, as episodic memory buffers characterized by different signal-to-noise ratios and completeness levels. These datasets serve to test the robustness of the model regarding local behavioral optimization.

The Maze2d domain consists of maze navigation tasks of three difficulty levels, specifically *umaze*, *medium*, and *large*. The core challenge of this task lies in testing the capacity for trajectory stitching. The datasets comprise numerous sub-optimal trajectories with diverse start and end points and lack globally optimal demonstrations. Consequently, the agent cannot solve the problem through simple imitation but must rely on a constructive reconstruction mechanism. By recombining fragmented memory segments within the latent space, the agent builds a globally connected cognitive map, thereby discovering shortcuts that did not exist in the original data.

We benchmark VGAR against representative offline RL algorithms, including standard value-based methods (TD3+BC (Fujimoto & Gu, 2021), IQL (Kostrikov et al., 2022)) and recent generative or constraint-based approaches (SQL (Xu et al., 2023), DOGE (Li et al., 2022), LAPO (X. Chen et al., 2022), A2PO (Qing et al., 2024)).

Experimental Results

Table 1 presents the comparison between VGAR and representative offline RL baselines. In **MuJoCo** motor control tasks, VGAR demonstrates robustness in handling complex distributions. Specifically, in *Walker2d* and *HalfCheetah* tasks, VGAR significantly outperforms traditional conservative methods on memory buffers with lower data quality (*medium*, *medium-replay*), while maintaining comparable performance on *medium-expert*. This evidence suggests that its constructive reconstruction effectively balances generation fidelity with value maximization, allowing for the extraction of optimal behaviors from sub-optimal memories without degrading expert performance. In the more challenging **Maze2d** spatial navigation tasks, VGAR’s advantage becomes increasingly prominent as maze complexity rises. While not achieving the best performance on *maze2d-umaze*, VGAR successfully overcomes the bottleneck of long-trajectory stitching in *maze2d-medium* and *maze2d-large*. This capability is attributed to VGAR’s **constructive memory mechanism**, which actively establishes cognitive bridges between fragmented memory traces, thereby achieving global optimality in long-horizon planning tasks.

Validating Cognitive Mechanisms

To validate the necessity of VGAR’s core cognitive components, we conducted an ablation study comparing the full model with its variants, as shown in Figure 2.

First, removing value guidance from the generative objective (i.e., setting $\lambda = 0$ in Eq. 4) resulted in a performance drop across most tasks. Although subsequent selection mechanisms can still filter for relatively superior actions, a generative module lacking top-down value guidance limits **mental**

Table 1: Normalized performance comparison on D4RL Gym-MuJoCo and Maze2d datasets. The highest scores are highlighted in **bold**. Results for VGAR are reported as the mean $\pm$ standard deviation over 4 random seeds. Baseline scores for SQL (Xu et al., 2023) and DOGE (Li et al., 2022) are reported from their original papers, while results for other methods are sourced from A2PO (Qing et al., 2024).

Tasks	TD3+BC	IQL	CQL	SQL	DOGE	LAPO	A2PO	VGAR (Ours)
halfcheetah-medium	48.3	47.4	44.0	48.3	45.3	46.0	47.1	59.59 $\pm$ 0.98
hopper-medium	59.3	66.3	58.5	75.5	98.6	51.6	80.3	99.54 $\pm$ 2.33
walker2d-medium	83.7	78.3	72.5	84.2	86.8	80.8	84.9	91.11 $\pm$ 1.23
halfcheetah-medium-replay	44.6	44.2	45.5	44.8	42.8	41.9	44.8	54.44 $\pm$ 1.42
hopper-medium-replay	60.9	94.7	95.0	99.7	76.2	50.1	101.6	90.35 $\pm$ 2.84
walker2d-medium-replay	81.8	73.9	81.6	81.2	87.3	60.6	82.8	94.99 $\pm$ 0.79
halfcheetah-medium-expert	90.7	86.7	91.6	94.0	78.7	94.2	95.6	92.82 $\pm$ 0.31
hopper-medium-expert	98.0	91.5	105.4	111.8	102.7	111.0	113.4	105.60 $\pm$ 2.08
walker2d-medium-expert	110.1	109.6	108.8	110.0	110.4	110.9	112.1	109.55 $\pm$ 0.95
total	677.4	692.6	702.9	749.5	728.8	647.1	762.6	797.99
halfcheetah-random	11.3	14.5	23.4	-	17.8	27.0	25.5	33.06 $\pm$ 0.99
hopper-random	9.2	8.2	3.0	-	21.1	16.1	18.4	10.96 $\pm$ 1.69
walker2d-random	1.1	6.5	2.1	-	0.9	1.9	3.6	13.13 $\pm$ 7.09
maze2d-umaze	24.2	56.2	17.0	-	-	78.0	133.3	85.59 $\pm$ 1.77
maze2d-medium	33.5	25.7	22.5	-	-	43.2	84.0	130.28 $\pm$ 2.38
maze2d-large	128.5	45.7	2.5	-	-	69.7	127.6	208.97 $\pm$ 1.43

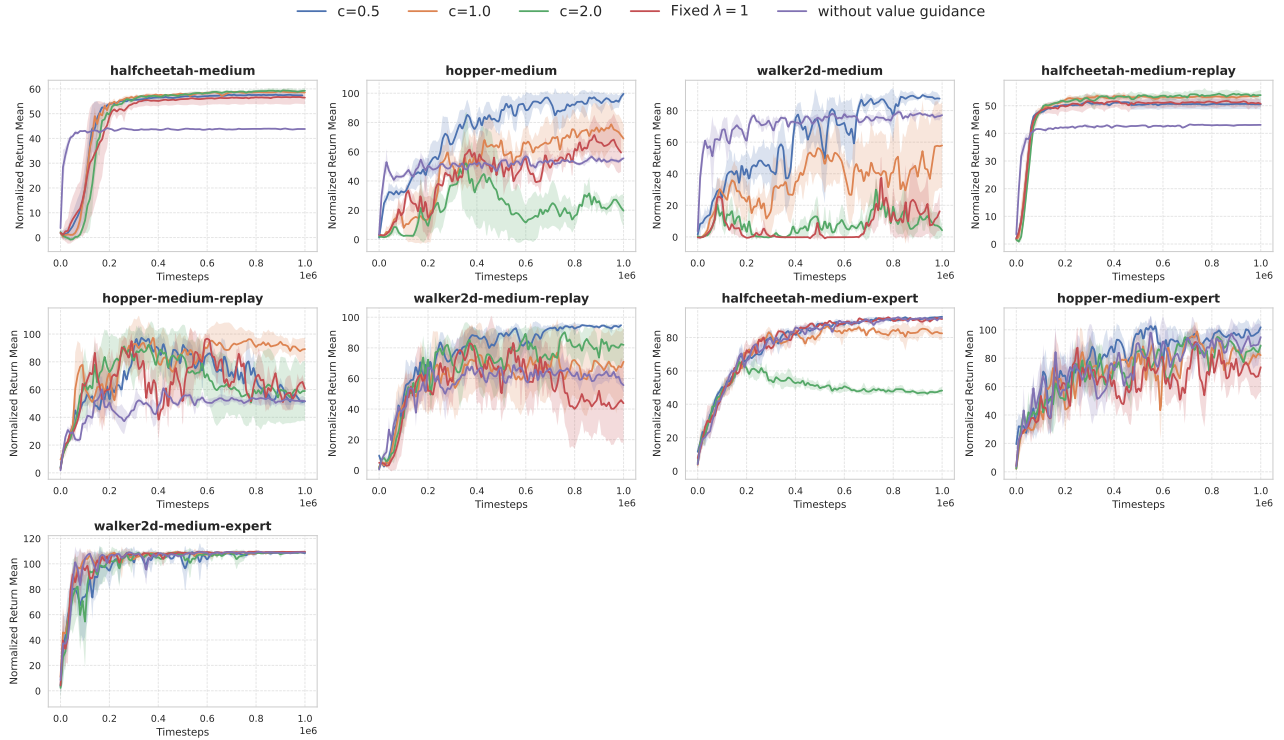


Figure 2: Ablation study results on Gym-MuJoCo benchmarks. The solid lines represent the mean normalized return across 4 seeds, while the shaded regions indicate the standard deviation.

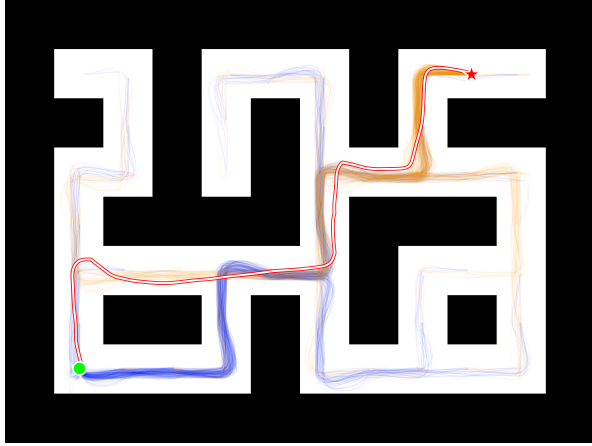


Figure 3: Trajectory visualization in maze2d-large. VGAR generates an efficient path (red curve) connecting the start (green circle) and goal (red star) by integrating suboptimal data.

rehearsal to the mere replay of historical data support. This implies that the agent loses the capability to actively construct **high-value** behaviors that transcend original experiences, preventing **constructive memory** from surpassing the quality limitations of the suboptimal history.

Second, the analysis of the metacognitive mechanism indicates that variants employing a fixed confidence coefficient generally perform suboptimally. The absence of the dynamic **metacognitive feedback loop** (Eq. 5) makes it difficult for the model to detect the accumulation of **reconstruction distortions**. Without this real-time check, the model cannot dynamically suppress hallucinated actions when deviation from original memory becomes excessive, leading to premature convergence or performance volatility.

In contrast, VGAR’s metacognitive confidence monitoring demonstrates adaptability by dynamically modulating effective confidence levels based on the **baseline confidence coefficient** c . Empirical results show that a lower baseline provides stronger constraints in environments like *Hopper* and *Walker2d*, effectively suppressing high-risk hallucinations by enforcing memory fidelity. Conversely, a higher baseline allows the model to employ aggressive value guidance in *HalfCheetah* to accelerate policy extraction when reconstruction error is low.

Visualizing Constructive Memory

Figure 3 presents a visualization analysis of planning trajectories within the *maze2d-large* environment. As illustrated, blue traces originate from the starting position, while orange traces converge toward the goal marked by a red star. Although the maximum trajectory length in the original memory buffer reaches 600 time steps, the dataset exhibits a scarcity of complete paths directly connecting the start to the goal. In

contrast, the distinct red trajectory generated by VGAR visually demonstrates the advantage of Constructive Memory. By reconstructing and stitching fragmented memories in the latent space, the agent successfully discovers a novel shortcut requiring only 398 time steps. This evidence confirms that VGAR transcends the mechanical replay of historical data and achieves policy optimization through the constructive reconstruction mechanism.

Conclusion

Inspired by the cognitive theory of constructive memory, this paper proposes the VGAR framework, which formalizes the mechanism of memory reconsolidation via a value-guided generative process. This approach enables the agent to actively filter suboptimal noise within the latent space and integrate fragmented experiences into effective and robust policies. Empirical evaluations on MuJoCo and Maze2d benchmarks strongly validate the robustness of VGAR against out-of-distribution risks as well as its capacity for constructive planning. While the generative nature of VGAR implies higher computational overhead during inference, this work effectively translates cognitive insights into a computational framework for data-driven reinforcement learning. Future work will explore how to consolidate the constructive planning process into efficient habitual policies and further refine the adaptive regulation mechanism of metacognitive parameters.

Acknowledgments

Drafting assistance was provided by LLMs.

References

- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34, 15084–15097.
- Chen, X., Ghadirzadeh, A., Yu, T., Wang, J., Gao, Y., Li, W., Bin, L., Finn, C., & Zhang, C. (2022). Lapo: Latent-variable advantage-weighted policy optimization for offline reinforcement learning. *Advances in Neural Information Processing Systems*.
- Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792), 671–675.
- Daou, A., & Lee, H. (2025). Editorial: Hippocampal function and reinforcement learning. *Frontiers in Computational Neuroscience*, 19, 1595369.
- Diekmann, N., & Cheng, S. (2023). A model of hippocampal replay driven by experience and environmental structure facilitates spatial learning. *Elife*, 12, e82301.

- Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1), 91.
- Fujimoto, S., & Gu, S. (2021). A minimalist approach to offline reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 20132–20145, Vol. 34). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/a8166da05c5a094f7dc03724b41886e5-Paper.pdf
- Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? a historical perspective on current neuroscience literature. *Neuropsychologia*, 53, 104–114.
- Hansen-Estruch, P., Kostrikov, I., Janner, M., Kuba, J. G., & Levine, S. (2023). Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*.
- Kim, J., Lee, S., Kim, W., & Sung, Y. (2023). Decision conformer: Local filtering in metaformer is sufficient for decision making. *arXiv preprint arXiv:2310.03022*.
- Kim, S., Choi, Y., Matsunaga, D. E., & Kim, K.-E. (2024). Stitching sub-trajectories with conditional diffusion model for goal-conditioned offline rl. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12), 13160–13167.
- Kostrikov, I., Nair, A., & Levine, S. (2022). Offline reinforcement learning with implicit q-learning. *International Conference on Learning Representations*. <https://openreview.net/forum?id=68n2s9ZJWF8>
- Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.
- Li, J., Zhan, X., Xu, H., Zhu, X., Liu, J., & Zhang, Y.-Q. (2022). Distance-sensitive offline reinforcement learning. *Deep Reinforcement Learning Workshop NeurIPS 2022*.
- Mao, Y., Wang, Q., Chen, C., Qu, Y., & Ji, X. (2024). Offline reinforcement learning with ood state correction and ood action suppression. *Advances in Neural Information Processing Systems*, 37, 93568–93601.
- Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, 21(11), 1609–1617.
- Peng, X. B., Kumar, A., Zhang, G., & Levine, S. (2019). Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*.
- Qing, Y., Liu, S., Cong, J., Chen, K., Zhou, Y., & Song, M. (2024). A2po: Towards effective offline reinforcement learning from an advantage-aware perspective. *Advances in Neural Information Processing Systems*, 37, 29064–29090.
- Roscow, E. L., Howe, T., Lepora, N. F., & Jones, M. W. (2025). Post-learning replay of hippocampal-striatal activity is biased by reward-prediction signals. *Nature Communications*, 16(1), 10394.
- Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481), 773–786.
- Schacter, D. L., & Thakral, P. P. (2024). Constructive memory and conscious experience. *Journal of Cognitive Neuroscience*, 36(8), 1567–1577. https://doi.org/10.1162/jocn_a_02201
- Wang, Z., Hunt, J. J., & Zhou, M. (2022). Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*.
- Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The tolmachenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5), 1249–1263.
- Wu, Y.-H., Wang, X., & Hamaya, M. (2023). Elastic decision transformer. *Advances in neural information processing systems*, 36, 18532–18550.
- Xu, H., Jiang, L., Li, J., Yang, Z., Wang, Z., Chan, W. K. V., & Zhan, X. (2023). Offline rl with no ood actions: In-sample learning via implicit value regularization. *ICLR*.
- Zheng, H., Shen, L., Luo, Y., Ye, D., Xu, S., Du, B., Shen, J., & Tao, D. (2025). Value-guided decision transformer: A unified reinforcement learning framework for online and offline settings. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.