OFFLINE DISTRIBUTIONAL WORLD MODEL VALUE ITERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned world models enable scalable planning in reinforcement learning but remain challenging in the offline setting, where planning can exploit erroneous rewards and transitions induced by model approximation error. While prior work has established distributional value iteration under known dynamics, little is known about distributional planning over learned world models in partially observable environments. We address this gap by deriving Wasserstein bounds that characterize how world-model approximation error affects latent return distributions and by introducing a support-dependent reward shift with distributional convergence guarantees under stated assumptions. Building on this analysis, we propose Offline Distributional World model Value Iteration (ODWVI), a practical algorithm for support-constrained distributional planning in learned latent world models. Across three offline planning benchmarks, ODWVI achieves the highest mean return among the evaluated methods in the main benchmark configurations. Support ablations further show that performance degrades as planning is extended to increasingly weakly supported actions, highlighting the importance of behavior-policy support when planning with imperfect learned world models.

1 INTRODUCTION

Offline reinforcement learning learns policies from fixed datasets without further environment interaction (Levine et al., 2020). Under partial observability, world models represent interaction histories as latent states and support planning or policy learning through imagined trajectories (Kaelbling et al., 1998; Hafner et al., 2019a; 2025). However, offline planning can exploit inaccurate predictions for actions poorly supported by the data (Kidambi et al., 2020; Yu et al., 2020). Recent work addresses the mismatch between predictive model training and downstream policy optimization (Chen et al., 2026). We study how model error affects distributional value iteration and how behavior support can constrain its backups.

Distributional planning models the entire discounted-return law, extending value iteration beyond expected returns (Bellemare et al., 2023; Preuett et al., 2026). With learned dynamics, approximation error affects this law, while action selection can favor optimistic predictions. We therefore ask: *How can we approximate distributional value iteration in learned latent models while limiting exploitation of weakly supported predictions?*

We introduce *Offline Distributional World model Value Iteration* (ODWVI), combining distributional planning with behavior-supported backups and support-dependent pessimism. Our contributions are:

- We develop simulation lemmas, value-distribution error bounds, and convergence guarantees for learned latent belief dynamics under stated assumptions.
- We introduce a support-dependent left-shift operator and instantiate ODWVI with training-time behavior-support constraints and critic-greedy execution.
- Across three ViZDoom environments and multiple offline data regimes, we report gains over greedy BC and other baselines and characterize sensitivity to backup support size.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

DistRL Distributional reinforcement learning models the full distribution of discounted returns rather than only their expectation, enabling richer treatment of stochasticity, uncertainty, and risk (Bellemare et al., 2017; Dabney et al., 2018b;a; Bellemare et al., 2023). Most prior work develops distributional Bellman operators for fully observable MDPs. More closely related to our setting, DPBVI extends distributional value iteration to POMDPs by defining return distributions over beliefs and adapting point-based value iteration to the distributional case (Preuett et al., 2026). This line of work provides the distributional planning foundation for ODWVI. However, whereas DPBVI assumes known belief dynamics and reward distributions, ODWVI learns a latent belief world model from offline data and performs distributional value iteration inside the learned model.

Offline RL Offline RL learns policies from fixed datasets without further environment interaction, making it vulnerable to distribution shift and overestimation on out-of-distribution (OOD) actions (Levine et al., 2020). Behavior-constrained methods such as BCQ, BEAR, and BRAC restrict policy improvement toward actions supported by the offline data (Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019). CQL instead learns conservative Q-values by penalizing high values on OOD actions (Kumar et al., 2020), while CODAC extends conservative offline RL to return distributions by penalizing OOD return quantiles (Ma et al., 2021). ODWVI addresses OOD actions during imagined planning: rather than regularizing value learning directly on dataset transitions, it constrains distributional Bellman backups inside a learned latent world model using behavior-policy support.

World Models World-model methods learn predictive dynamics and use them for planning or policy optimization. PlaNet and Dreamer learn latent dynamics from high-dimensional observations and use imagined trajectories for planning or actor-critic learning (Hafner et al., 2019a;b; 2020; 2025). Offline model-based methods address model exploitation through uncertainty penalties, pessimistic absorbing states, conservative value learning, or policy-driven model adaptation (Yu et al., 2020; Kidambi et al., 2020; Yu et al., 2021; Chen et al., 2026). The effectiveness of uncertainty penalties depends on their interaction with other planning choices, including rollout horizon (Lu et al., 2022). More recent approaches include PWM, which extracts policies through gradient-based optimization in world models pretrained on offline data (Georgiev et al., 2025), and JOWA, which jointly pretrains world and action-value models from offline data (Cheng et al., 2025). ODWVI focuses specifically on behavior-supported distributional value iteration: it restricts Bellman target construction and next-action bootstrapping to behavior-supported actions in a learned latent belief model, while retaining critic-greedy execution over the full action space.

3 Setting

3.1 Partially Observable Markov Decision Process

We consider a discrete-time POMDP $\mathcal{P}$, defined as a tuple: $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \Omega, b_0, P_R^{\mathcal{P}}, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, $\mathcal{T}(s, a, s') = P(s_{t+1} = s' \mid s_t = s, a_t = a)$ is the state transition model, $\Omega(o', s', a) = P(o_{t+1} = o' \mid s_{t+1} = s', a_t = a)$ is the sensor model, $b_0 \in \Delta$ is the initial belief, Δ is the belief space, $P_R^{\mathcal{P}}(\cdot \mid s, a)$ is the conditional distribution of the immediate reward r when taking action a in state s with support $[R_{\text{min}}, R_{\text{max}}]$, and $0 \leq \gamma < 1$ is the discount factor.

3.2 Offline POMDP

In the offline setting, the agent is constrained to learning from a fixed dataset of online interaction data collected by one or more behavior policies. Formally, a dataset comprised of n trajectories, collected by $\pi_\beta \in \Pi_\beta$, is defined as $D = \{\tau_i\}_{i=1}^n$, where $\tau_i = o_0, a_0, r_0, \ldots, o_{T_i}$ and T_i is the number of timesteps in trajectory i. Crucially, $\mathcal{S}$, Ω, $P_R^{\mathcal{P}}$, and $\mathcal{T}$ are unknown.

3.3 LATENT BELIEF MDP

Every $\mathcal{P}$ induces an equivalent fully observable belief MDP whose state is the posterior belief over unobservable $\mathcal{S}$. The latent belief MDP considered here is defined as $\hat{\mathcal{P}} = \langle \hat{\Delta}, \mathcal{A}, \mathcal{O}, \hat{\tau}, \hat{P_R^{\mathcal{P}}}, \gamma \rangle$, where $\hat{\Delta}$ is latent belief space, $\hat{\tau}$ is the latent belief dynamics model, $\hat{P_R^{\mathcal{P}}}$ is the learned immediate reward distribution, and $\mathcal{A}$, $\mathcal{O}$, γ are equivalent to those in $\mathcal{P}$. Unlike true beliefs, latent beliefs need not correspond to valid probability distributions over latent states. Instead, $\hat{b} \in \hat{\Delta}$ denotes a learned representation of interaction history that is assumed to be approximately sufficient for predicting future observations, rewards, and latent beliefs. We assume latent beliefs satisfy an approximate *Markov property*, such that future observations, rewards, and latent beliefs depend on interaction history only through the current latent belief.

3.4 DISTRIBUTIONAL REINFORCEMENT LEARNING UNDER PARTIAL OBSERVABILITY

True beliefs condition on the complete observed history, including rewards. Distributional RL models the discounted-return law $Z^\pi(b)$ under randomness in rewards, hidden states, transitions, and observations. The true belief-space evaluation operator is

$$\widetilde{\mathcal{T}}_{PO}^{\pi} Z(b) \overset{D}{=} R + \gamma \tau^\pi Z(b), \tag{1}$$

where $A \sim \pi(\cdot \mid b)$, $\tau^\pi Z(b) \overset{D}{=} Z(B')$, and $B' = \tau(b, A, R, O)$ is the posterior after the observed reward and next observation. Reward and next belief are sampled from their true joint conditional law. The continuation return is sampled at B', independently of R conditional on B'. The optimality operator is similarly defined, selecting an action that maximizes the expected Bellman target. Fixed-policy evaluation is a γ-contraction under the supremum p-Wasserstein metric; optimality convergence additionally requires the action-gap and regularity conditions used below (Preuett et al., 2026; Bellemare et al., 2023).

4 DISTRIBUTIONAL PLANNING UNDER LEARNED LATENT DYNAMICS

We compare true and learned distributional Bellman operators and bound the resulting return-distribution error.

4.1 LATENT DISTRIBUTIONAL BELLMAN OPTIMALITY OPERATORS

Let $\phi : \Delta \to \hat{\Delta}$ be the latent representation and let $Z(\hat{b})$ denote a state-return random variable. Define

$$\widetilde{\mathcal{T}}_{PO,\phi}^{*} Z(b) \overset{D}{=} R + \gamma \widetilde{\tau}^* Z(b),$$
$$\widehat{\mathcal{T}}_{PO}^{*} Z(\hat{b}) \overset{D}{=} \hat{R} + \gamma \hat{\tau}^* Z(\hat{b}),$$

where $\widetilde{\tau}^* Z(b) \overset{D}{=} Z(\phi(B'))$ and $\hat{\tau}^* Z(\hat{b}) \overset{D}{=} Z(\hat{B}')$. The pairs (R, B') and $(\hat{R}, \hat{B}')$ are sampled jointly from P at $(b, a^*(b))$ and $\hat{P}$ at $(\hat{b}, \hat{a}^*(\hat{b}))$, respectively. Each action maximizes its model's expected Bellman target, with fixed tie-breaking. Continuation returns are sampled from Z at the next latent belief, independently of the reward conditional on that belief. The true-model backup is indexed by b, and is compared with the learned backup at $\hat{b} = \phi(b)$; exact sufficiency of ϕ is not assumed.

4.2 SIMULATION LEMMA

Fix $p \in [1, \infty)$, and let $\mathcal{P}_p(\mathbb{R})$ denote probability laws with finite p-th moments. Let $\eta : \hat{\Delta} \to \mathcal{P}_p(\mathbb{R})$ denote the laws of Z; $\mathcal{T}\eta$ denotes the laws induced by $\mathcal{T}Z$. Throughout, $\mathcal{A}$ is finite, functions and kernels are measurable, and rewards and return laws have uniformly bounded p-th moments. Equip $\hat{\Delta}$ with metric $d_{\hat{\Delta}}$ and write

$$\bar{w}_p(\eta, \nu) := \sup_{\hat{b} \in \hat{\Delta}} W_p\big(\eta(\hat{b}), \nu(\hat{b})\big).$$

Assumption 1 (Lipschitz return-distribution function)**.** *There exists* $0 \le L_\eta < \infty$ *such that*

$$W_p\big(\eta(\hat{b}_1), \eta(\hat{b}_2)\big) \le L_\eta d_{\hat{\Delta}}(\hat{b}_1, \hat{b}_2)$$

for all $\hat{b}_1, \hat{b}_2 \in \hat{\Delta}$.

Assumption 2 (Joint model approximation error)**.** *For every* b, a*, the true pair* $(R, \phi(B'))$ *at* (b, a) *and the learned pair* $(\hat{R}, \hat{B}')$ *at* $(\phi(b), a)$ *admit a coupling preserving both joint laws such that*

$$\mathbb{E}|R - \hat{R}|^p \le \epsilon_R^p,$$
$$\mathbb{E}\, d_{\hat{\Delta}}(\phi(B'), \hat{B}')^p \le \epsilon_\tau^p,$$

where $\epsilon_R, \epsilon_\tau \ge 0$ *are finite and uniform in* b, a.

These bounds hold under the same coupling, rather than separately optimized marginal couplings. Write $\delta_\eta := \epsilon_R + \gamma L_\eta \epsilon_\tau$.

For the common continuation law η, write

$$Q_\eta^P(b, a) := \mathbb{E}_P[R + \gamma Z(\phi(B')) \mid b, a],$$
$$\hat{Q}_\eta(\hat{b}, a) := \mathbb{E}_{\hat{P}}[\hat{R} + \gamma Z(\hat{B}') \mid \hat{b}, a],$$

where $Z(\hat{b}) \sim \eta(\hat{b})$. Thus $a^*(b) = \arg\max_a Q_\eta^P(b, a)$ and $\hat{a}^*(\hat{b}) = \arg\max_a \hat{Q}_\eta(\hat{b}, a)$.

Assumption 3 (Positive action gap)**.** *For the continuation law* η *under consideration, there exists* $\Delta_{\min} > 0$ *such that*

$$Q_\eta^P(b, a^*(b)) - Q_\eta^P(b, a) \ge \Delta_{\min}$$

for every $b \in \Delta$ *and* $a \ne a^*(b)$.

Lemma 1. *Optimality Operator Simulation Lemma. Under Assumptions 1–3, if* $\delta_\eta < \Delta_{\min}/2$*, then* $a^*(b) = \hat{a}^*(\phi(b))$ *for every* $b \in \Delta$*, and*

$$\sup_{b \in \Delta} W_p\Big(\widetilde{\mathcal{T}}_{PO,\phi}^* \eta(b), \widehat{\mathcal{T}}_{PO}^* \eta(\phi(b))\Big) \le \delta_\eta.$$

This action agreement concerns a common continuation law η.

4.3 Error Bounds

Theorem 1. *Latent Optimality Operator Error Bound. Let* π^* *and* $\hat{\pi}^*$ *be deterministic stationary policies optimal in expected return in the true belief MDP and learned latent MDP, respectively, using the fixed tie-breaking rule. Let* $\eta_P^*(b)$ *and* $\hat{\eta}^*(\hat{b})$ *denote their discounted-return laws. Under Assumption 2, suppose* $\hat{\eta}^*$ *satisfies Assumption 1 with constant* L_**. Define*

$$e_* := \frac{\epsilon_R + \gamma L_* \epsilon_\tau}{1 - \gamma}.$$

If the true optimal action values Q_P^* *satisfy*

$$Q_P^*(b, \pi^*(b)) - Q_P^*(b, a) \ge \Delta_* > 2e_*$$

for every $b \in \Delta$ *and* $a \ne \pi^*(b)$*, then* $\hat{\pi}^*(\phi(b)) = \pi^*(b)$ *for every* b*, and*

$$\sup_{b \in \Delta} W_p\big(\eta_P^*(b), \hat{\eta}^*(\phi(b))\big) \le e_*.$$

4.4 Convergence

Proposition 1. *Latent Optimality Operator Convergence. Fix* $p \in [1, \infty)$ *and* $0 \le \gamma < 1$*. Let the fixed learned latent MDP have finite action space* $\mathcal{A}$ *and uniformly bounded* p*-th reward moments. Let* $\hat{Q}^*$ *and* $\hat{\pi}^*$ *denote its optimal action values and a deterministic stationary optimal policy. Suppose there exists* $\widehat{\Delta}_* > 0$ *such that*

$$\hat{Q}^*(\hat{b}, \hat{\pi}^*(\hat{b})) - \hat{Q}^*(\hat{b}, a) \ge \widehat{\Delta}_*$$

for every $\hat{b} \in \hat{\Delta}$ and $a \neq \hat{\pi}^*(\hat{b})$. Then, for any measurable initial return-distribution function $\eta_0 : \hat{\Delta} \to \mathcal{P}_p(\mathbb{R})$ with uniformly bounded p-th moments, the iterates

$$\eta_{k+1} = \widehat{\mathcal{T}}^*_{PO}\eta_k$$

converge to the learned-model return law $\hat{\eta}^*$ of $\hat{\pi}^*$:

$$\bar{w}_p(\eta_k, \hat{\eta}^*) \longrightarrow 0.$$

Analogous results can be established for fixed-policy evaluation operators. These results are relevant to offline policy evaluation (OPE) and are deferred to Appendix A.

These guarantees are conditional on the stated model-error, Lipschitz, and action-gap assumptions; we do not estimate their constants or establish that the trained networks satisfy them. The map ϕ is a theoretical representation abstraction, not an assertion that the implemented history encoder exactly factors through the true belief.

5 Support-Constrained World Model Planning

The preceding analysis characterizes how world-model approximation error affects return distributions. To encourage safer planning, we regularize distributional value iteration using behavior-policy support, aiming to reduce exploitation of unreliable predictions for weakly represented actions.

5.1 Support-Constrained Left-Shift Operator

Let $\pi_\beta(a \mid \hat{b})$ be a behavior policy learned from the offline dataset. We introduce the support-dependent shift

$$c(\hat{b}, a) = f\big(\pi_\beta(a \mid \hat{b})\big), \tag{2}$$

where f is decreasing, so weaker support incurs larger penalties. The shift is intended to discourage value iteration from favoring weakly supported actions because of optimistic world-model predictions. Behavior likelihood is a heuristic support estimate, not a calibrated model-error bound. The following results establish contraction and convergence of the idealized shifted operators; their use for safer planning motivates the empirical evaluation.

Using this support estimate, define the left-shift evaluation operator

$$\widehat{\mathcal{T}}^{\pi}_{LS} Z(\hat{b}) \overset{D}{=} \hat{R} - c(\hat{b}, A) + \gamma Z(\hat{B}'),$$

where $A \sim \pi(\cdot \mid \hat{b})$, reward and next latent belief are sampled jointly from the learned model, and continuation returns are sampled as in the preceding Bellman operators. For $Z(\hat{b}) \sim \eta(\hat{b})$, the optimality operator $\widehat{\mathcal{T}}^*_{LS}$ uses the same target with

$$A = a^*_{c,\eta}(\hat{b}) := \arg\max_{a \in \mathcal{A}} \big\{\hat{Q}_\eta(\hat{b}, a) - c(\hat{b}, a)\big\},$$

using fixed tie-breaking. Thus action selection maximizes the expected shifted target, not the unpenalized action value.

Assumption 4 (Support-dependent pessimism)**.** *The shift $c : \hat{\Delta} \times \mathcal{A} \to \mathbb{R}$ is measurable, deterministic, fixed during iteration, and independent of η, with*

$$\sup_{\hat{b}, a} |c(\hat{b}, a)| \leq C_c < \infty.$$

Proposition 2. *Left-Shift Optimality Operator Convergence. Fix $p \in [1, \infty)$ and $0 \leq \gamma < 1$. Let the fixed learned latent MDP have finite action space $\mathcal{A}$ and uniformly bounded p-th reward moments, and assume Assumption 4. Let $\hat{Q}^*_c$ and $\hat{\pi}^*_c$ denote the optimal action values and a deterministic stationary optimal policy in the learned model with modified reward*

$$\hat{R}_c = \hat{R} - c(\hat{b}, a).$$

Suppose there exists $\widehat{\Delta}_c > 0$ *such that*

$$\hat{Q}_c^*(\hat{b}, \hat{\pi}_c^*(\hat{b})) - \hat{Q}_c^*(\hat{b}, a) \geq \widehat{\Delta}_c$$

for every $\hat{b} \in \hat{\Delta}$ *and* $a \neq \hat{\pi}_c^*(\hat{b})$*. Then, for any measurable initial return-distribution function* $\eta_0 : \hat{\Delta} \to \mathcal{P}_p(\mathbb{R})$ *with uniformly bounded* p*-th moments, the iterates*

$$\eta_{k+1} = \widehat{\mathcal{T}}_{LS}^* \eta_k$$

converge to the return law $\hat{\eta}_c^*$ *of* $\hat{\pi}_c^*$ *under the modified reward* $\hat{R}_c$*:*

$$\bar{w}_p(\eta_k, \hat{\eta}_c^*) \longrightarrow 0.$$

6 Offline Distributional World Model Value Iteration

ODWVI first trains a recurrent state-space model (RSSM) and behavior policy π_β on the offline dataset, then freezes both and trains a categorical action-return critic using model-generated transitions (Algorithm 1). The RSSM represents observation-action histories as latent beliefs and predicts successor states, rewards, and continuation. The behavior policy supplies the support sets and penalties.

The preceding guarantees concern idealized operators; they do not establish convergence of neural critic training or identify the critic with the return law of its unrestricted execution policy. Supported targets already incorporate the shift c; execution applies no additional shift.

6.1 Support-Constrained Distributional Value Iteration

Given the learned world model and behavior policy, define

$$\mathcal{S}_k(\hat{b}) = \mathrm{TopK}\big(\pi_\beta(\cdot \mid \hat{b})\big).$$

Only actions in $\mathcal{S}_k(\hat{b})$ receive Bellman targets, and next-action bootstrapping is restricted to $\mathcal{S}_k(\hat{b}')$. These restrictions apply during training, not execution. Let $Q(\hat{b}, a)$ denote the mean of the practical critic's predicted categorical action-return distribution. The evaluated policies select

$$\begin{aligned} \pi_{\mathrm{BC}}(\hat{b}) &= \arg\max_{a \in \mathcal{A}} \pi_\beta(a \mid \hat{b}), \\ \pi_{\mathrm{ODWVI}}(\hat{b}) &= \arg\max_{a \in \mathcal{A}} Q(\hat{b}, a). \end{aligned} \tag{3}$$

Both policies are greedy, but ODWVI may execute outside $\mathcal{S}_k(\hat{b})$. For $k = 1$ with a unique behavior mode, zero OOD penalty and a positive margin imply agreement with greedy BC; soft enforcement permits departures (Appendix E).

For supported actions, ODWVI instantiates the support-constrained left-shift operator using

$$c(\hat{b}, a) = (1 - \gamma)\lambda \left(-\log\left(\pi_\beta(a \mid \hat{b}) + \epsilon \right) \right), \tag{4}$$

where $\lambda \geq 0$ controls the degree of pessimism and $\epsilon > 0$ is a small constant for numerical stability. The factor $(1 - \gamma)$ normalizes the per-step support penalty to the scale of discounted returns. Consequently, actions with lower behavior-policy support receive larger pessimistic shifts during value iteration. For actions in $\mathcal{S}_k(\hat{b})$, ODWVI constructs left-shifted distributional Bellman targets using imagined transitions generated by the learned latent belief dynamics and reward model. We denote the C51 cross-entropy loss on these supported, left-shifted Bellman targets by $\mathcal{L}_{\mathrm{support}}$.

Actions outside $\mathcal{S}_k(\hat{b})$ receive no Bellman targets and are instead regularized by a soft penalty on predicted values above $\bar{Q}_{\mathcal{S}}(\hat{b}) - m$. This penalty is not an execution-time action mask:

$$\begin{aligned} \bar{Q}_{\mathcal{S}}(\hat{b}) &= \frac{1}{|\mathcal{S}_k(\hat{b})|} \sum_{a' \in \mathcal{S}_k(\hat{b})} Q(\hat{b}, a'), \\ \mathcal{L}_{\mathrm{ood}}(\hat{b}) &= \frac{1}{|\mathcal{A} \setminus \mathcal{S}_k(\hat{b})|} \sum_{a \notin \mathcal{S}_k(\hat{b})} \Big[Q(\hat{b}, a) - \bar{Q}_{\mathcal{S}}(\hat{b}) + m \Big]_+^2 . \end{aligned} \tag{5}$$

Algorithm 1 Offline Distributional World model Value Iteration (ODWVI)

Require: Offline dataset $\mathcal{D}$, support size k, critic updates N

Learn latent belief MDP $\hat{\mathcal{P}}$ and behavior policy π_β from $\mathcal{D}$; freeze both
Initialize online critic η_θ and target critic $\eta_{\bar{\theta}}$
for $i = 1, \ldots, N$ **do**
  Sample trajectories from $\mathcal{D}$ and infer a minibatch of latent beliefs
  Define $\mathcal{S}_k(\hat{b}) = \mathrm{TopK}\Big(\pi_\beta(\cdot \mid \hat{b})\Big)$ for each sampled latent belief
  for all sampled $\hat{b}$ and $a \in \mathcal{S}_k(\hat{b})$ **do**
    Predict $\hat{r}$ and $\hat{p}_{\mathrm{cont}}$, and imagine $\hat{b}'$ from $\hat{\mathcal{P}}$ at $(\hat{b}, a)$
    Compute $\mathcal{S}_k(\hat{b}') = \mathrm{TopK}\Big(\pi_\beta(\cdot \mid \hat{b}')\Big)$
    Select $a' = \arg\max_{\tilde{a} \in \mathcal{S}_k(\hat{b}')} Q_\theta(\hat{b}', \tilde{a})$
    Form a distributional Bellman target using $\hat{r}$, $\hat{p}_{\mathrm{cont}}$, and $\eta_{\bar{\theta}}(\hat{b}', a')$
    Subtract the support penalty $c(\hat{b}, a)$ in the Bellman target, then apply C51 projection
  end for
  Update θ using $\mathcal{L}_{\mathrm{support}} + \lambda_{\mathrm{ood}}\mathcal{L}_{\mathrm{ood}}$
  Update target critic parameters $\bar{\theta}$
end for
Define execution policy $\pi_{\mathrm{ODWVI}}(\hat{b}) = \arg\max_{a \in \mathcal{A}} Q_\theta(\hat{b}, a)$
return η_θ and π_{ODWVI}

where $m = \alpha(v_{\max} - v_{\min})$. We set $\mathcal{L}_{\mathrm{ood}}(\hat{b}) = 0$ when $\mathcal{A} \setminus \mathcal{S}_k(\hat{b}) = \varnothing$.

The critic is trained using support-constrained Bellman targets for supported actions together with the unsupported-action regularizer. Consequently, ODWVI learns return distributions through value iteration in the learned latent belief MDP while discouraging exploitation of actions that are weakly supported by the offline dataset.

7 Results and Discussion

7.1 Experimental Setup

We evaluate ODWVI on three VizDoom environments: Deadly Corridor, My Way Home, and Deathmatch. Performance is measured using mean undiscounted return. Offline datasets are collected using DreamerV3 and all methods are trained exclusively from offline experience before being evaluated online. Each trained model is evaluated for 100 episodes using a common held-out evaluation seed across methods.

We compare ODWVI against four representative baselines. Behavior Cloning (BC) represents imitation-based offline learning, Dreamer Offline represents offline latent world-model planning, JOWA is a recent offline world-model planning method, and ODWVI-CQL provides a conservative-value-learning comparison using the same frozen world model (Appendix G). Details regarding environment configuration, dataset collection, dataset construction, hyperparameters, and implementation details are deferred to the appendix.

Within each environment and data regime, all compared methods use the same selected offline dataset described in Appendix D; references to the full dataset mean all transitions in that selected subset, not the unfiltered collection.

7.2 Support-Constrained Distributional Value Iteration

Figure 1 evaluates the effect of behavior support on approximate distributional value iteration in Deadly Corridor. Increasing k relaxes the action restriction in imagined Bellman updates and next-action bootstrapping. The learned world model is held fixed, and all variants execute greedily over the full action space.



Figure 1: Effect of support size k on approximate distributional value iteration in Deadly Corridor. Relaxing the training-time support constraint degrades mean policy return. At $k = 128$, all actions participate in Bellman backups. Execution is greedy over the full action space for every k.

ODWVI achieves its highest mean undiscounted return at $k = 1$. As the support constraint is relaxed, mean return decreases from 442.25 at $k = 1$ to 49.99 for value iteration without the top-k backup restriction ($k = 128$), despite using the same learned world model. This pattern is consistent with value iteration amplifying unreliable model predictions when increasingly weakly supported actions enter the backups. It supports the motivation for safer, support-guided planning, while the ablation evaluates the complete procedure rather than isolating the left shift.

The $k = 1$ variant also achieves higher mean returns than greedy BC on Deadly Corridor, Deathmatch, and the 500–1000 trajectory My Way Home dataset (Table 1). Both methods execute greedily, but use different action scores, and ODWVI is not restricted to the behavior model's top-1 action at execution. These are empirical gains of the complete implementation, not a guarantee that critic-driven departures from BC improve returns. Extended support-size comparisons appear in Appendix Figure 3.

7.3 World-Model Planning Across Data Regimes

Figure 2 compares My Way Home models trained on trajectory-index windows 0–500, 500–1000, and 500–2000 from different phases of the online collector. The first two contain 501 episodes each, while the last contains 1,501; their transition counts and success rates differ (Table 3). These windows vary data composition, coverage, and training-data quantity rather than independently controlling world-model prediction error.

On the early 0–500 window, where successful trajectories are rare, all methods obtain low mean returns and BC has the highest reported mean. On the 500–1000 window, ODWVI attains a mean return of 0.88, compared with 0.58 for BC, 0.74 for Dreamer Offline, and 0.05 for ODWVI-CQL. On the success-rich 500–2000 window, ODWVI and Dreamer Offline both obtain a mean return of 0.97 at the reported precision.

These patterns are consistent with changes in the usefulness of world-model planning, but they do not isolate model fidelity: behavior-policy learning and state-action coverage also change across the datasets.

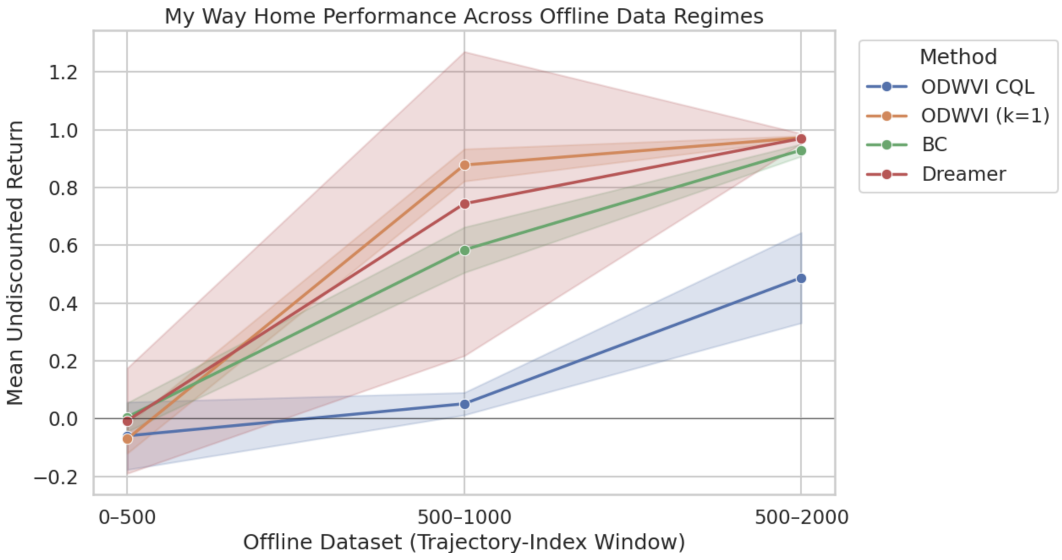

Figure 2: World-model planning across offline data regimes in My Way Home. ODWVI has its largest reported advantage over BC in the intermediate trajectory window, while ODWVI and Dreamer Offline obtain comparable mean returns on later, success-rich data. The windows describe training-data regimes, not independently measured levels of model fidelity.

Table 1: Main benchmark results. Values report the mean and standard deviation across five training runs, each evaluated for 100 episodes using undiscounted returns. My Way Home uses the 500–1000 trajectory window. Higher is better; a dash denotes a result not reported in this study.

Method	Deadly Corridor	My Way Home	Deathmatch
BC	413.23 ± 19.99	0.58 ± 0.08	7.51 ± 0.94
Dreamer	317.57 ± 76.06	0.74 ± 0.53	3.63 ± 1.59
JOWA-G	414.64 ± 33.15	–	–
JOWA-P	374.93 ± 45.04	–	–
ODWVI-CQL	414.06 ± 30.95	0.05 ± 0.04	7.89 ± 0.38
ODWVI ($k = 1$)	$\mathbf{442.25 \pm 37.22}$	$\mathbf{0.88 \pm 0.06}$	$\mathbf{8.07 \pm 0.67}$

7.4 Discussion

ODWVI achieves the highest reported mean among evaluated methods in each main benchmark configuration (Table 1), including improvements over ODWVI-CQL under a shared world model. Across My Way Home data regimes, BC is strongest on the early window, while ODWVI and Dreamer Offline are comparable on the success-rich window (Table 4).

These comparisons evaluate complete procedures. They neither isolate the effects of distributional parameterization, the left shift, and OOD regularization nor directly measure model exploitation or categorical return-law accuracy. The theory concerns idealized operators; the experiments assess deployed-policy returns.

8 Conclusion

We developed distributional planning error bounds and conditional convergence guarantees under learned latent dynamics, and introduced ODWVI with behavior-supported backups and a support-dependent shift. Across three ViZDoom tasks, ODWVI improves returns in several offline regimes with unrestricted critic-greedy execution. Support ablations motivate further study of how behavior coverage affects model exploitation.

AI USE STATEMENT

We used large language models (LLMs) to assist with method implementation, manuscript preparation, and theoretical development. This included generating and revising code, drafting and editing manuscript text, developing proof arguments and generating proof drafts, refining mathematical claims and assumptions, and providing feedback on experimental interpretation and relevant literature.

The initial research questions and core methodological ideas originated with the authors. The authors reviewed and revised the LLM-assisted material and validated the mathematical arguments. All final research decisions remained with the authors, who take full responsibility for the accuracy, originality, and integrity of the paper and accompanying code.

REPRODUCIBILITY STATEMENT

The ODWVI algorithm is specified in Algorithm 1, and complete proofs of the theoretical results are provided in the appendix. Appendices C and D describe the environments, offline data collection, and dataset filtering. Training details and hyperparameters are provided in Appendix G, and the evaluation protocol is described in Appendix E. Source code and experiment configurations will be made publicly available to facilitate reproduction of the reported results.

REFERENCES

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

Jiayu Chen, Le Xu, Aravind Venugopal, and Jeff Schneider. Policy-driven world model adaptation for robust offline model-based reinforcement learning. https://arxiv.org/abs/2505.13709, 2026. Accessed: 2026-07-04.

Jie Cheng, Ruixi Qiao, Yingwei Ma, Binhua Li, Gang Xiong, Qinghai Miao, Yongbin Li, and Yisheng Lv. Scaling offline model-based rl via jointly-optimized world-action model pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=T1OvCSFaum.

Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit quantile networks for distributional reinforcement learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1096–1105. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/dabney18a.html.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018b. ISBN 978-1-57735-800-8.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Ignat Georgiev, Varun Giridhar, Nicklas Hansen, and Animesh Garg. PWM: Policy learning with multi-task world models. In *International Conference on Learning Representations*,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/240ea1741b205ea295721d55184ac43b-Paper-Conference.pdf.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 09–15 Jun 2019a. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. https://arxiv.org/abs/1912.01603, 2019b. Accessed: 2026-07-04.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. https://arxiv.org/abs/2010.02193, 2020. Accessed: 2026-07-04.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, Apr 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(98)00023-X. URL https://www.sciencedirect.com/science/article/pii/S000437029800023X.

Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. ViZDoom: A Doom-based AI research platform for visual reinforcement learning. In *IEEE Conference on Computational Intelligence and Games*, pp. 341–348, Santorini, Greece, Sep 2016. IEEE. URL http://arxiv.org/abs/1605.02097. The best paper award.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f7efa4f864ae9b88d43527f4b14f750f-Paper.pdf.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. https://arxiv.org/abs/1906.00949, 2019. Accessed: 2026-07-04.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

Sergey Levine, Aviral Kumar, G. Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. https://arxiv.org/abs/2005.01643, 2020. Accessed: 2026-07-04.

Cong Lu, Philip J. Ball, Jack Parker-Holder, Michael A. Osborne, and Stephen J. Roberts. Revisiting design choices in offline model-based reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zz9hXVhf40.

Yecheng Jason Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Larry Preuett, Qiuyi Zhang, and Muhammad Aurangzeb Ahmad. Provable distributional value iteration under partial observability. https://arxiv.org/abs/2505.06518, 2026. Accessed: 2026-07-04.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. https://arxiv.org/abs/1911.11361, 2019. Accessed: 2026-07-04.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28954–28967. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f29a179746902e331572c483c45e5086-Paper.pdf.

A Additional Theoretical Results

The main text focuses on optimality operators because ODWVI performs planning through value iteration. However, the same analysis can be carried out for fixed-policy evaluation operators. These results are relevant to offline policy evaluation (OPE) and provide the technical foundation for several optimality results presented in the main paper.

In this section, we introduce latent-space distributional Bellman evaluation operators and establish corresponding simulation lemmas, error bounds, and contraction results. These results are fixed-policy analogues of the optimality results presented in the main text.

A.1 Latent Distributional Bellman Evaluation Operators

For a fixed latent policy π, define

$$\widetilde{\mathcal{T}}^{\pi}_{PO,\phi} Z(b) \overset{D}{=} R + \gamma \widetilde{\tau}^{\pi} Z(b),$$
$$\widehat{\mathcal{T}}^{\pi}_{PO} Z(\hat{b}) \overset{D}{=} \hat{R} + \gamma \hat{\tau}^{\pi} Z(\hat{b}),$$

where $\hat{b} = \phi(b)$, $\widetilde{\tau}^{\pi} Z(b) \overset{D}{=} Z(\phi(B'))$, and $\hat{\tau}^{\pi} Z(\hat{b}) \overset{D}{=} Z(\hat{B}')$. Both backups sample $A \sim \pi(\cdot \mid \hat{b})$. Conditional on A, reward and next belief are sampled jointly from the corresponding model, with continuation sampling as in the optimality operators. The true policy therefore acts on b through $\pi(\cdot \mid \phi(b))$.

A.2 Simulation Lemma

Lemma 2. *Evaluation Operator Simulation Lemma. Under Assumptions 1 and 2, for any fixed latent policy π,*

$$\sup_{b \in \Delta} W_p\Big(\widetilde{\mathcal{T}}^{\pi}_{PO,\phi} \eta(b), \widehat{\mathcal{T}}^{\pi}_{PO} \eta(\phi(b))\Big) \le \delta_{\eta}.$$

A.3 Error Bound

Theorem 2. *Latent Evaluation Operator Error Bound. For a fixed latent policy π, define $\pi^{\uparrow}(a \mid b) := \pi(a \mid \phi(b))$. Let $\eta_P^{\pi^{\uparrow}}(b)$ and $\hat{\eta}^{\pi}(\hat{b})$ denote the discounted-return laws under $\pi^{\uparrow}$ in $\mathcal{P}$ and under π in $\hat{\mathcal{P}}$, respectively. Under Assumption 2, if $\hat{\eta}^{\pi}$ satisfies Assumption 1 with constant L_{π}, then*

$$\sup_{b \in \Delta} W_p\big(\eta_P^{\pi^{\uparrow}}(b), \hat{\eta}^{\pi}(\phi(b))\big) \le \frac{\epsilon_R + \gamma L_{\pi} \epsilon_{\tau}}{1 - \gamma}.$$

A.4 CONTRACTION

Proposition 3. *Latent Evaluation Operator Contraction. Fix $p \in [1, \infty)$ and $0 \leq \gamma < 1$, and assume uniformly bounded p-th learned reward moments. For any fixed measurable policy π, $\widehat{\mathcal{T}}^{\pi}_{PO}$ maps measurable return-distribution functions with uniformly bounded p-th moments into the same space and satisfies*

$$\bar{w}_p\big(\widehat{\mathcal{T}}^{\pi}_{PO}\eta, \widehat{\mathcal{T}}^{\pi}_{PO}\eta'\big) \leq \gamma \bar{w}_p(\eta, \eta').$$

A.5 SUPPORT-CONSTRAINED LEFT SHIFT EVALUATION OPERATOR

Proposition 4. *Left-Shift Evaluation Operator Contraction. Fix $p \in [1, \infty)$ and $0 \leq \gamma < 1$. Assume uniformly bounded p-th learned reward moments and Assumption 4. For any fixed measurable policy π, the operator $\widehat{\mathcal{T}}^{\pi}_{LS}$ maps measurable return-distribution functions with uniformly bounded p-th moments into the same space and satisfies*

$$\bar{w}_p\big(\widehat{\mathcal{T}}^{\pi}_{LS}\eta, \widehat{\mathcal{T}}^{\pi}_{LS}\eta'\big) \leq \gamma \bar{w}_p(\eta, \eta').$$

Its unique fixed point is the return law $\hat{\eta}^{\pi}_{c}$ under π in the learned model with reward $\hat{R}_c = \hat{R} - c(\hat{b}, a)$.

B PROOFS

B.1 PROOF OF LEMMA 2

Lemma 2. *Evaluation Operator Simulation Lemma. Under Assumptions 1 and 2, for any fixed latent policy π,*

$$\sup_{b \in \Delta} W_p\Big(\widetilde{\mathcal{T}}^{\pi}_{PO,\phi}\eta(b), \widehat{\mathcal{T}}^{\pi}_{PO}\eta(\phi(b))\Big) \leq \delta_\eta.$$

Proof. Fix b, a and take the coupling in Assumption 2. Conditional on the coupled reward–transition pairs, use a shared independent uniform variable and inverse CDFs to optimally couple $G \sim \eta(\phi(B'))$ and $\hat{G} \sim \eta(\hat{B}')$. This preserves each model's conditional continuation law and gives

$$\begin{aligned}
\mathbb{E}|G - \hat{G}|^p &= \mathbb{E}\, W_p^p\big(\eta(\phi(B')), \eta(\hat{B}')\big) \\
&\leq L_\eta^p \mathbb{E}\, d_{\hat{\Delta}}(\phi(B'), \hat{B}')^p \\
&\leq L_\eta^p \epsilon_\tau^p.
\end{aligned}$$

For $Y = R + \gamma G$ and $\hat{Y} = \hat{R} + \gamma \hat{G}$, Minkowski's inequality yields

$$\begin{aligned}
W_p\big(\mathcal{L}(Y), \mathcal{L}(\hat{Y})\big) &\leq \|Y - \hat{Y}\|_{L^p} \\
&\leq \|R - \hat{R}\|_{L^p} + \gamma \|G - \hat{G}\|_{L^p} \\
&\leq \epsilon_R + \gamma L_\eta \epsilon_\tau = \delta_\eta.
\end{aligned}$$

For a stochastic policy, sample the same $A \sim \pi(\cdot \mid \phi(b))$ in both backups and use this construction conditional on $A = a$. The resulting targets $Y^\pi, \hat{Y}^\pi$ have the required evaluation laws and satisfy

$$\mathbb{E}|Y^\pi - \hat{Y}^\pi|^p \leq \sum_{a \in \mathcal{A}} \pi(a \mid \phi(b)) \delta_\eta^p = \delta_\eta^p.$$

Taking the p-th root and then the supremum over b proves the result. □

B.2 PROOF OF LEMMA 1

Lemma 1. *Optimality Operator Simulation Lemma. Under Assumptions 1–3, if $\delta_\eta < \Delta_{\min}/2$, then $a^*(b) = \hat{a}^*(\phi(b))$ for every $b \in \Delta$, and*

$$\sup_{b \in \Delta} W_p\Big(\widetilde{\mathcal{T}}^{*}_{PO,\phi}\eta(b), \widehat{\mathcal{T}}^{*}_{PO}\eta(\phi(b))\Big) \leq \delta_\eta.$$

This action agreement concerns a common continuation law η.

Proof. Fix b. Applying Lemma 2 to the policy that always selects a, and using the fact that the difference of means is bounded by $W_1 \le W_p$, gives

$$|Q_\eta^P(b,a) - \hat{Q}_\eta(\phi(b),a)| \le \delta_\eta \qquad \text{for every } a.$$

Let $a^* = a^*(b)$. For every $a \neq a^*$,

$$\begin{aligned}\hat{Q}_\eta(\phi(b),a^*) - \hat{Q}_\eta(\phi(b),a) &\\ \ge Q_\eta^P(b,a^*) - Q_\eta^P(b,a) - 2\delta_\eta &\\ \ge \Delta_{\min} - 2\delta_\eta > 0.&\end{aligned}$$

Thus $\hat{a}^*(\phi(b)) = a^*(b)$. Both optimality backups now use the same action, so the fixed-action case of Lemma 2 bounds their Wasserstein distance by δ_η. Taking the supremum over b completes the proof. □

B.3 Proof of Theorem 2

Theorem 2. *Latent Evaluation Operator Error Bound. For a fixed latent policy π, define $\pi^\uparrow(a \mid b) := \pi(a \mid \phi(b))$. Let $\eta_P^{\pi^\uparrow}(b)$ and $\hat{\eta}^\pi(\hat{b})$ denote the discounted-return laws under $\pi^\uparrow$ in $\mathcal{P}$ and under π in $\hat{\mathcal{P}}$, respectively. Under Assumption 2, if $\hat{\eta}^\pi$ satisfies Assumption 1 with constant L_π, then*

$$\sup_{b\in\Delta} W_p\big(\eta_P^{\pi^\uparrow}(b), \hat{\eta}^\pi(\phi(b))\big) \le \frac{\epsilon_R + \gamma L_\pi \epsilon_\tau}{1-\gamma}.$$

Proof. Set $\mu := \eta_P^{\pi^\uparrow}$, $h := \hat{\eta}^\pi \circ \phi$, and $\delta_\pi := \epsilon_R + \gamma L_\pi \epsilon_\tau$. Let $\widetilde{\mathcal{T}}_{PO}^{\pi^\uparrow}$ denote the true belief-space evaluation operator. The return laws satisfy

$$\mu = \widetilde{\mathcal{T}}_{PO}^{\pi^\uparrow}\mu, \qquad \hat{\eta}^\pi = \widehat{\mathcal{T}}_{PO}^{\pi}\hat{\eta}^\pi.$$

Moreover,

$$\big[\widetilde{\mathcal{T}}_{PO}^{\pi^\uparrow}h\big](b) = \widetilde{\mathcal{T}}_{PO,\phi}^{\pi}\hat{\eta}^\pi(b).$$

Thus Lemma 2, applied to $\hat{\eta}^\pi$, gives

$$\sup_{b\in\Delta} W_p\Big(\big[\widetilde{\mathcal{T}}_{PO}^{\pi^\uparrow}h\big](b), h(b)\Big) \le \delta_\pi.$$

Define $D := \sup_b W_p(\mu(b), h(b))$, which is finite by the standing moment bounds. Coupling the same action, reward, and next true belief in two evaluation backups leaves only the discounted continuation discrepancy. Hence the true belief-space evaluation operator is a γ-contraction, and the triangle inequality yields

$$\begin{aligned}D &\le \sup_b W_p\Big(\big[\widetilde{\mathcal{T}}_{PO}^{\pi^\uparrow}\mu\big](b), \big[\widetilde{\mathcal{T}}_{PO}^{\pi^\uparrow}h\big](b)\Big) + \delta_\pi\\ &\le \gamma D + \delta_\pi.\end{aligned}$$

Rearranging gives $D \le \delta_\pi/(1-\gamma)$. □

B.4 Proof of Theorem 1

Theorem 1. *Latent Optimality Operator Error Bound. Let π^* and $\hat{\pi}^*$ be deterministic stationary policies optimal in expected return in the true belief MDP and learned latent MDP, respectively, using the fixed tie-breaking rule. Let $\eta_P^*(b)$ and $\hat{\eta}^*(\hat{b})$ denote their discounted-return laws. Under Assumption 2, suppose $\hat{\eta}^*$ satisfies Assumption 1 with constant L_*. Define*

$$e_* := \frac{\epsilon_R + \gamma L_* \epsilon_\tau}{1-\gamma}.$$

If the true optimal action values Q_P^* satisfy

$$Q_P^*(b,\pi^*(b)) - Q_P^*(b,a) \ge \Delta_* > 2e_*$$

for every $b \in \Delta$ and $a \neq \pi^*(b)$, then $\hat{\pi}^*(\phi(b)) = \pi^*(b)$ for every b, and

$$\sup_{b\in\Delta} W_p\big(\eta_P^*(b), \hat{\eta}^*(\phi(b))\big) \le e_*.$$

Proof. Let V_P^* and $\hat{V}^*$ denote the optimal expected state values, set $g := \hat{V}^* \circ \phi$, and write $\delta_* := \epsilon_R + \gamma L_* \epsilon_\tau$. Since $\hat{\eta}^*$ is the return law of $\hat{\pi}^*$, its mean is $\hat{V}^*$ and $\hat{Q}^* = \hat{Q}_{\hat{\eta}^*}$. Applying Lemma 2 to each constant-action policy and using $|\mathbb{E}X - \mathbb{E}Y| \le W_p(X, Y)$ gives

$$\sup_{b,a} \left| Q^P_{\hat{\eta}^*}(b, a) - \hat{Q}^*(\phi(b), a) \right| \le \delta_*.$$

Let T_P^* be the scalar Bellman optimality operator on true beliefs. Since

$$T_P^* g(b) = \max_a Q^P_{\hat{\eta}^*}(b, a), \qquad g(b) = \max_a \hat{Q}^*(\phi(b), a),$$

taking maxima gives $\|T_P^* g - g\|_\infty \le \delta_*$. Using $V_P^* = T_P^* V_P^*$ and scalar Bellman contraction,

$$\|V_P^* - g\|_\infty \le \gamma \|V_P^* - g\|_\infty + \delta_*,$$

so $\|V_P^* - g\|_\infty \le e_*$. Also, $Q_P^*(b, a) - Q^P_{\hat{\eta}^*}(b, a) = \gamma \mathbb{E}_P[V_P^*(B') - g(B') \mid b, a]$. Consequently,

$$\begin{aligned} \sup_{b,a} |Q_P^*(b, a) - \hat{Q}^*(\phi(b), a)| &\le \gamma \|V_P^* - g\|_\infty + \delta_* \\ &\le \gamma e_* + \delta_* = e_*. \end{aligned}$$

For every $a \neq \pi^*(b)$, the assumed optimal action gap now implies

$$\hat{Q}^*(\phi(b), \pi^*(b)) - \hat{Q}^*(\phi(b), a) \ge \Delta_* - 2e_* > 0.$$

Thus $\hat{\pi}^*(\phi(b)) = \pi^*(b)$ for all b. Applying Theorem 2 to $\pi = \hat{\pi}^*$, whose lifted policy equals π^*, yields

$$\sup_{b \in \Delta} W_p\big(\eta_P^*(b), \hat{\eta}^*(\phi(b))\big) \le \frac{\delta_*}{1 - \gamma} = e_*.$$

□

B.5 Proof of Proposition 3

Proposition 3. *Latent Evaluation Operator Contraction. Fix $p \in [1, \infty)$ and $0 \le \gamma < 1$, and assume uniformly bounded p-th learned reward moments. For any fixed measurable policy π, $\widehat{\mathcal{T}}^\pi_{PO}$ maps measurable return-distribution functions with uniformly bounded p-th moments into the same space and satisfies*

$$\bar{w}_p\big(\widehat{\mathcal{T}}^\pi_{PO}\eta, \widehat{\mathcal{T}}^\pi_{PO}\eta'\big) \le \gamma \bar{w}_p(\eta, \eta').$$

Proof. Fix $\hat{b}$ and sample the same $A \sim \pi(\cdot \mid \hat{b})$ and joint pair $(\hat{R}, \hat{B}')$ in both backups. Conditional on these samples, use a shared independent uniform variable and inverse CDFs to optimally couple $G \sim \eta(\hat{B}')$ and $G' \sim \eta'(\hat{B}')$. The targets $Y = \hat{R} + \gamma G$ and $Y' = \hat{R} + \gamma G'$ have the required Bellman laws. Thus

$$\begin{aligned} &W_p^p\big(\widehat{\mathcal{T}}^\pi_{PO}\eta(\hat{b}), \widehat{\mathcal{T}}^\pi_{PO}\eta'(\hat{b})\big) \\ &\quad \le \mathbb{E}|Y - Y'|^p = \gamma^p \mathbb{E}\, W_p^p\big(\eta(\hat{B}'), \eta'(\hat{B}')\big) \\ &\quad \le \gamma^p \bar{w}_p(\eta, \eta')^p. \end{aligned}$$

Taking the p-th root and supremum over $\hat{b}$ gives the claim. Minkowski's inequality and the uniform moment bounds ensure that the operator maps the stated function space into itself. □

B.6 Proof of Proposition 1

Proposition 1. *Latent Optimality Operator Convergence. Fix $p \in [1, \infty)$ and $0 \le \gamma < 1$. Let the fixed learned latent MDP have finite action space $\mathcal{A}$ and uniformly bounded p-th reward moments. Let $\hat{Q}^*$ and $\hat{\pi}^*$ denote its optimal action values and a deterministic stationary optimal policy. Suppose there exists $\widehat{\Delta}_* > 0$ such that*

$$\hat{Q}^*(\hat{b}, \hat{\pi}^*(\hat{b})) - \hat{Q}^*(\hat{b}, a) \ge \widehat{\Delta}_*$$

for every $\hat{b} \in \hat{\Delta}$ and $a \neq \hat{\pi}^*(\hat{b})$. Then, for any measurable initial return-distribution function $\eta_0 : \hat{\Delta} \to \mathcal{P}_p(\mathbb{R})$ with uniformly bounded p-th moments, the iterates

$$\eta_{k+1} = \widehat{\mathcal{T}}^*_{PO}\eta_k$$

converge to the learned-model return law $\hat{\eta}^*$ of $\hat{\pi}^*$:

$$\bar{w}_p(\eta_k, \hat{\eta}^*) \longrightarrow 0.$$

Proof. Let $V_k(\hat{b}) := \int z\, \eta_k(\hat{b})(dz)$, and let $\hat{V}^*$ be the learned model's optimal expected value. Since each update maximizes its expected Bellman target,

$$V_{k+1}(\hat{b}) = \max_{a \in \mathcal{A}} \mathbb{E}_{\hat{P}}[\hat{R} + \gamma V_k(\hat{B}') \mid \hat{b}, a].$$

Thus (V_k) follows scalar value iteration, and Bellman contraction gives

$$\|V_k - \hat{V}^*\|_\infty \leq \gamma^k \|V_0 - \hat{V}^*\|_\infty \longrightarrow 0.$$

By the definition of $\hat{Q}_\eta$,

$$\varepsilon_k := \|\hat{Q}_{\eta_k} - \hat{Q}^*\|_\infty \leq \gamma \|V_k - \hat{V}^*\|_\infty \longrightarrow 0.$$

Choose K such that $2\varepsilon_k < \widehat{\Delta}_*$ for all $k \geq K$. For every $\hat{b}$ and $a \neq \hat{\pi}^*(\hat{b})$,

$$\hat{Q}_{\eta_k}(\hat{b}, \hat{\pi}^*(\hat{b})) - \hat{Q}_{\eta_k}(\hat{b}, a) \geq \widehat{\Delta}_* - 2\varepsilon_k > 0.$$

Hence the greedy policy equals $\hat{\pi}^*$ at every latent belief for all $k \geq K$, so

$$\eta_{k+1} = \widehat{\mathcal{T}}^{\hat{\pi}^*}_{PO}\eta_k \qquad (k \geq K).$$

Since $\hat{\eta}^* = \widehat{\mathcal{T}}^{\hat{\pi}^*}_{PO}\hat{\eta}^*$, Proposition 3 yields

$$\bar{w}_p(\eta_k, \hat{\eta}^*) \leq \gamma^{k-K} \bar{w}_p(\eta_K, \hat{\eta}^*) \longrightarrow 0.$$

The distance at K is finite because uniform p-th moment bounds are preserved by the discounted backups and hold for $\hat{\eta}^*$. □

B.7 Proof of Proposition 4

Proposition 4. *Left-Shift Evaluation Operator Contraction. Fix $p \in [1, \infty)$ and $0 \leq \gamma < 1$. Assume uniformly bounded p-th learned reward moments and Assumption 4. For any fixed measurable policy π, the operator $\widehat{\mathcal{T}}^\pi_{LS}$ maps measurable return-distribution functions with uniformly bounded p-th moments into the same space and satisfies*

$$\bar{w}_p\big(\widehat{\mathcal{T}}^\pi_{LS}\eta, \widehat{\mathcal{T}}^\pi_{LS}\eta'\big) \leq \gamma \bar{w}_p(\eta, \eta').$$

Its unique fixed point is the return law $\hat{\eta}^\pi_c$ under π in the learned model with reward $\hat{R}_c = \hat{R} - c(\hat{b}, a)$.

Proof. Fix $\hat{b}$ and sample the same $A \sim \pi(\cdot \mid \hat{b})$ and joint reward–transition pair $(\hat{R}, \hat{B}')$ for both backups. Conditional on these samples, optimally couple $G \sim \eta(\hat{B}')$ and $G' \sim \eta'(\hat{B}')$ using a shared independent uniform variable and inverse CDFs. Define

$$Y = \hat{R} - c(\hat{b}, A) + \gamma G, \qquad Y' = \hat{R} - c(\hat{b}, A) + \gamma G'.$$

Their laws are the two shifted backups. Both the reward and shift cancel pathwise, even when π is stochastic. Hence

$$\begin{aligned} &W_p^p\big(\widehat{\mathcal{T}}^\pi_{LS}\eta(\hat{b}), \widehat{\mathcal{T}}^\pi_{LS}\eta'(\hat{b})\big) \\ &\quad \leq \mathbb{E}|Y - Y'|^p = \gamma^p \mathbb{E}|G - G'|^p \\ &\quad = \gamma^p \mathbb{E}\, W_p^p\big(\eta(\hat{B}'), \eta'(\hat{B}')\big) \leq \gamma^p \bar{w}_p(\eta, \eta')^p. \end{aligned}$$

Taking the p-th root and supremum over $\hat{b}$ proves the contraction. Bounded c and the reward moment bound ensure closure of the stated function space. The discounted-return law for $\hat{R}_c$ has uniformly bounded p-th moments and satisfies this Bellman recursion; contraction makes it the unique fixed point. □

B.8 Proof of Proposition 2

Proposition 2. *Left-Shift Optimality Operator Convergence. Fix $p \in [1, \infty)$ and $0 \leq \gamma < 1$. Let the fixed learned latent MDP have finite action space $\mathcal{A}$ and uniformly bounded p-th reward moments, and assume Assumption 4. Let $\hat{Q}_c^*$ and $\hat{\pi}_c^*$ denote the optimal action values and a deterministic stationary optimal policy in the learned model with modified reward*

$$\hat{R}_c = \hat{R} - c(\hat{b}, a).$$

Suppose there exists $\widehat{\Delta}_c > 0$ such that

$$\hat{Q}_c^*(\hat{b}, \hat{\pi}_c^*(\hat{b})) - \hat{Q}_c^*(\hat{b}, a) \geq \widehat{\Delta}_c$$

for every $\hat{b} \in \hat{\Delta}$ and $a \neq \hat{\pi}_c^*(\hat{b})$. Then, for any measurable initial return-distribution function $\eta_0 : \hat{\Delta} \to \mathcal{P}_p(\mathbb{R})$ with uniformly bounded p-th moments, the iterates

$$\eta_{k+1} = \widehat{\mathcal{T}}_{LS}^* \eta_k$$

converge to the return law $\hat{\eta}_c^*$ of $\hat{\pi}_c^*$ under the modified reward $\hat{R}_c$:

$$\bar{w}_p(\eta_k, \hat{\eta}_c^*) \longrightarrow 0.$$

Proof. Let $\hat{\mathcal{P}}_c$ be the learned latent MDP obtained by replacing each reward sample $\hat{R}$ with $\hat{R}_c = \hat{R} - c(\hat{b}, a)$, leaving the dynamics unchanged. Assumption 4 ensures that this is a fixed model with uniformly bounded p-th reward moments.

By definition, $\widehat{\mathcal{T}}_{LS}^*$ selects an action maximizing $\mathbb{E}[\hat{R}_c + \gamma Z(\hat{B}') \mid \hat{b}, a]$ and returns the law of that target. It is therefore exactly the distributional Bellman optimality operator of $\hat{\mathcal{P}}_c$. The assumed gap is the uniform optimal action gap of this modified model. Applying Proposition 1 to $\hat{\mathcal{P}}_c$ yields

$$\bar{w}_p(\eta_k, \hat{\eta}_c^*) \longrightarrow 0,$$

where $\hat{\eta}_c^*$ is the full discounted-return law under $\hat{\pi}_c^*$ for reward $\hat{R}_c$. □

C Environments

We evaluate ODWVI on three partially observable ViZDoom tasks (Kempka et al., 2016). All use the Gymnasium `MultiBinary-v1` interface with RGB observations rendered without the HUD, resized to 64×64, and stored as an image observation component

$$o_t^{\text{image}} \in \{0, \ldots, 255\}^{3 \times 64 \times 64}.$$

We use no frame stacking. Each selected action is repeated for two environment steps, and the agent receives the summed reward over those steps. The full simulator state includes hidden pose, map state, enemies, items, projectiles, script variables, and the episode clock.

C.1 Deadly Corridor

This is a combat-navigation task using `VizdoomDeadlyCorridor-MultiBinary-v1`. The action space is $\{0, 1\}^7$ over `MOVE_LEFT`, `MOVE_RIGHT`, `ATTACK`, `MOVE_FORWARD`, `MOVE_BACKWARD`, `TURN_LEFT`, and `TURN_RIGHT`, giving 128 joint actions. The observation is $o_t = (o_t^{\text{image}}, o_t^{\text{feat}})$, where o_t^{feat} contains the scalar game variable `HEALTH`. At the environment-step level, the reward is progress-based: if X_t is the agent's corridor coordinate and G_t indicates reaching the armor/goal trigger, then

$$r_t = (X_t - X_{t-1}) + 1000 \cdot \mathbf{1}\{G_t\} - 100 \cdot \mathbf{1}\{\text{death at } t\}.$$

Thus moving backward can produce negative reward when it decreases X_t. Episodes terminate on death or reaching the goal and truncate at 2100 environment steps, approximately 1050 agent decision steps.

C.2 MY WAY HOME

This is a sparse-reward navigation task using `VizdoomMyWayHome-MultiBinary-v1`. The action space is $\{0,1\}^5$ over `TURN_LEFT`, `TURN_RIGHT`, `MOVE_FORWARD`, `MOVE_LEFT`, and `MOVE_RIGHT`, giving 32 joint actions. The observation is only the image component, $o_t = o_t^{\text{image}}$, with no additional game variables. The goal location is fixed across episodes, but the agent's initial position and orientation are randomized among scenario-defined start locations. At the environment-step level, the reward is

$$r_t = -0.0001 + 1.0 \cdot \mathbf{1}\{\text{goal reached at } t\}.$$

Episodes terminate when the goal is reached and truncate at 2100 environment steps, approximately 1050 agent decision steps.

C.3 DEATHMATCH

This is a combat task using `VizdoomDeathmatch-MultiBinary-v1` with a restricted ODWVI action interface. The action space is $\{0,1\}^9$ over `ATTACK`, `MOVE_LEFT`, `MOVE_RIGHT`, `MOVE_FORWARD`, `MOVE_BACKWARD`, `TURN_RIGHT`, `TURN_LEFT`, `SELECT_NEXT_WEAPON`, and `SELECT_PREV_WEAPON`, giving 512 joint actions. The observation is $o_t = (o_t^{\text{image}}, o_t^{\text{feat}})$, where o_t^{feat} contains `KILLCOUNT`, `HEALTH`, `ARMOR`, `SELECTED_WEAPON`, and `SELECTED_WEAPON_AMMO`. Rewards are based on kills and vary by enemy type: `Zombieman`: 1, `ShotgunGuy`: 3, `MarineChainsawVzd`: 3, `ChaingunGuy`: 4, `Demon`: 3, and `HellKnight`: 10. Episodes truncate at 4200 environment steps, approximately 2100 agent decision steps.

D DATASETS

D.1 TRAJECTORY DATASET COLLECTION

We collected offline trajectory datasets by training an online DreamerV3 agent in each ViZDoom environment and recording the real environment interaction stream into Minari. Each transition stores the observation, action, reward, termination flag, and truncation flag; each episode stores metadata including undiscounted return, episode length, training seed, and source run identifier. All collection runs used training seed 360749. If collection ended during an unfinished episode, the open trajectory was finalized and retained as a partial trajectory.

The nominal collection horizons were 1M environment steps for Deathmatch, 800k for MyWayHome, and 1.5M for Deadly Corridor. Each run was extended by roughly 10k additional environment steps, producing approximately 1.01M, 810k, and 1.51M transitions, respectively. The online collector learned substantial goal-reaching behavior in MyWayHome, but did not reliably solve Deathmatch or Deadly Corridor.

For Deathmatch and Deadly Corridor, the offline datasets used by ODWVI retained approximately the top 25% of episodes by undiscounted return. Episodes were sorted by descending return, with local episode index used as a deterministic tie-breaker. For MyWayHome, we used overlapping, inclusive, zero-based trajectory-index windows to represent phases of the online run: 0–500, 500–1000, and 500–2000. Early MyWayHome trajectories mostly contain poor behavior and timeout episodes, while later trajectories increasingly contain successful goal-reaching behavior.

E EVALUATION PROCEDURE

For each method and offline training configuration, we train five models using different training seeds. Each trained policy is evaluated for 100 episodes. We first average undiscounted episodic returns within each training run, then report the mean and standard deviation of these five run-level averages. Evaluation uses a common held-out seed across methods.

For offline DreamerV3, all transitions in the selected dataset for the corresponding environment and data regime are loaded into the replay buffer. Training then follows the DreamerV3 procedure with replay from this fixed dataset replacing environment interaction.

Table 2: Full collected trajectory datasets. Returns are undiscounted episode returns. Bracketed ranges show the minimum and maximum.

Environment	Transitions	Episodes	Success Rate	Return Med. [Min, Max]	Length Med. [Min, Max]
Deadly Corridor	1.51M	39,027	4.0%	340.28 [-116.00, 2,285.99]	35 [2, 1,050]
Deathmatch	1.01M	2,592	0.0%	3.00 [0.00, 44.00]	356 [2, 1,449]
My Way Home	810k	2,004	64.6%	0.97 [-0.21, 1.00]	143 [1, 1,050]

Table 3: Filtered trajectory datasets used for offline training. Return medians appear above bracketed first and third quartiles.

Environment	Filter	Transitions	Episodes	Success Rate	Return Med. [Q1, Q3]
Deadly Corridor	Top 25% return	580k	9,757	15.9%	489.40 [449.29, 646.79]
Deathmatch	Top 25% return	367k	648	0.0%	14.00 [11.00, 20.00]
My Way Home	Index 0–500	489k	501	7.4%	-0.21 [-0.21, -0.21]
My Way Home	Index 500–1000	216k	501	61.3%	0.97 [0.82, 0.98]
My Way Home	Index 500–2000	321k	1,501	83.8%	0.98 [0.96, 0.99]

ODWVI is trained in two phases. In the first phase, we train the world model and the behavior policy from the offline dataset. This same trained world model is then reused for all ODWVI variants, including CQL and behavior-support variants with $k \in \{1, 2, 4, 8, 128\}$. In the second phase, the world model is frozen and only the critic component is trained. The behavior-cloning baseline is evaluated using the behavior policy learned during phase-one training; this is the same behavior policy used to define the behavior-support sets for all k values.

BC is evaluated by selecting the highest-probability action under the learned behavior policy π_β, while ODWVI selects the highest-value action over the full action space (Equation 3). ODWVI's top-k restriction applies only to training targets and bootstrapping.

For $k = 1$ at a belief with a unique behavior mode, write $a_\beta = \pi_{\mathrm{BC}}(\hat{b})$. If $m > 0$, the regularizer satisfies

$$\mathcal{L}_{\mathrm{ood}}(\hat{b}) = 0 \implies Q(\hat{b}, a) \le Q(\hat{b}, a_\beta) - m \quad \text{for all } a \neq a_\beta.$$

Thus zero penalty at that belief forces agreement with greedy BC. The finite-weight soft loss does not impose these inequalities as hard constraints, so critic-greedy execution can depart from BC. Such departures are not evidence that values of actions lacking Bellman targets are accurate; performance is evaluated through true-environment returns.

For JOWA, we use a local implementation adapted to train on the same Minari/VizDoom offline datasets and action vocabulary as the other offline methods, using the 150M JOWA configuration. JOWA is trained from RGB observations with the same valid-action masking used for VizDoom; non-image state features are not used. We evaluate JOWA with both greedy Q-value action selection (`greedy_q`) and beam-search planning with horizon two (`beam_h2`), matching the original JOWA evaluation protocol in which planning is enabled with horizon $H = 2$ (Cheng et al., 2025). To limit training time and computational cost, we evaluated the 150M JOWA configuration only on Deadly Corridor. JOWA results for My Way Home and Deathmatch are therefore not reported.

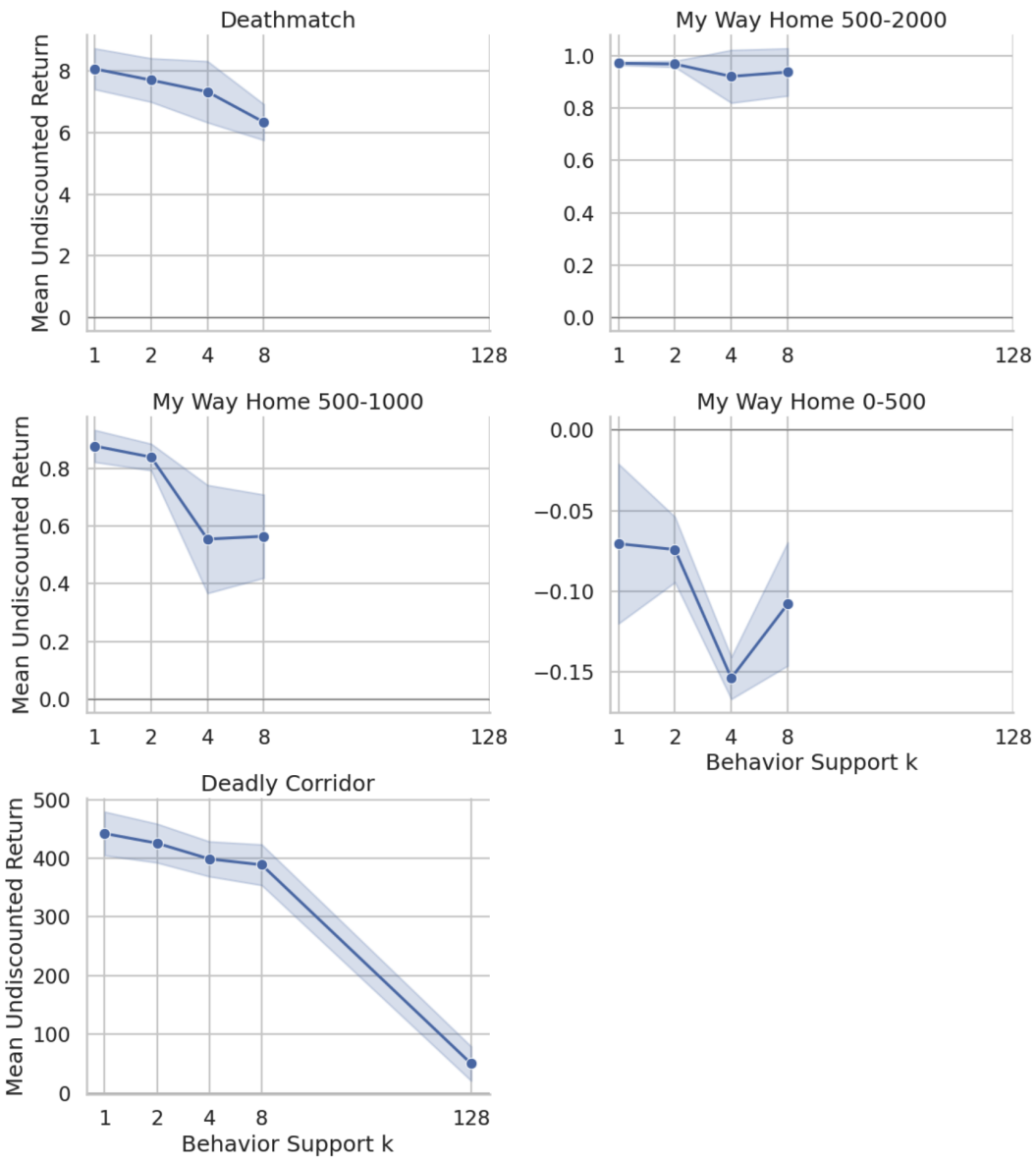

Figure 3: Effect of training support size k on mean undiscounted return across environments and offline data regimes. Smaller backup support sets generally yield higher mean returns, with the strongest trend in Deadly Corridor. All variants execute greedily over the full action space.

F Extended Results

Figure 3 extends the training-support comparison across environments and offline data regimes. Smaller support sets generally obtain higher mean returns, with the clearest decline as k increases in Deadly Corridor. Execution remains critic-greedy over the full action space.

Table 4 gives the complete reported comparisons. Small-support ODWVI variants obtain the highest means on Deadly Corridor, Deathmatch, and the 500–1000 My Way Home window. BC has the highest reported mean on the early window, while ODWVI and Dreamer Offline are comparable on the success-rich window. The comparison with ODWVI-CQL evaluates complete training objectives under a shared latent world model, rather than isolating one component.

Table 4: Detailed benchmark results. Values report the mean and standard deviation across five training runs, each evaluated for 100 episodes using undiscounted returns. Higher is better; a dash denotes a result not reported in this study.

Method	Deadly Corridor	Deathmatch	MWH 0–500	MWH 500–1000	MWH 500–2000
BC	413.23 ± 19.99	7.51 ± 0.94	**0.00 ± 0.05**	0.58 ± 0.08	0.93 ± 0.02
Dreamer	317.57 ± 76.06	3.63 ± 1.59	-0.01 ± 0.18	0.74 ± 0.53	**0.97 ± 0.02**
JOWA-G	414.64 ± 33.15	–	–	–	–
JOWA-P	374.93 ± 45.04	–	–	–	–
ODWVI ($k = 1$)	**442.25 ± 37.22**	**8.07 ± 0.67**	-0.07 ± 0.05	**0.88 ± 0.06**	**0.97 ± 0.01**
ODWVI ($k = 2$)	425.77 ± 33.13	7.71 ± 0.71	-0.07 ± 0.02	0.84 ± 0.05	**0.97 ± 0.01**
ODWVI ($k = 4$)	398.89 ± 29.87	7.33 ± 1.00	-0.15 ± 0.01	0.55 ± 0.19	0.92 ± 0.10
ODWVI ($k = 8$)	388.89 ± 34.69	6.34 ± 0.58	-0.11 ± 0.04	0.56 ± 0.14	0.94 ± 0.09
ODWVI ($k = 128$)	49.99 ± 29.22	–	–	–	–
ODWVI-CQL	414.06 ± 30.95	7.89 ± 0.38	-0.06 ± 0.12	0.05 ± 0.04	0.49 ± 0.16

Table 5: ODWVI hyperparameters. DC, DM, and MWH denote Deadly Corridor, Deathmatch, and My Way Home, respectively. Settings are shared unless otherwise indicated.

Parameter	Value
Discount γ	0.997
Behavior-support size k	$1, 2, 4, 8$; also 128 for DC
Left-shift weight λ	0.1
OOD weight λ_{ood}	5.0
Margin fraction α	0.05
Probability stabilizer ϵ	10^{-6}
Categorical atoms	255
Return bounds $[v_{\min}, v_{\max}]$	DC: $[-100, 100]$ DM: $[-10, 10]$ MWH: $[-2, 2]$
World-model learning rate	4×10^{-5}
Behavior-policy learning rate	5×10^{-4}
Critic learning rate	3×10^{-4}
Batch size / sequence length	16 / 64
Phase-one updates	20,000
Phase-two updates	50,000
Target-update coefficient τ	0.02

G Implementation Details and Hyperparameters

We instantiate the latent world model using an RSSM based on the DreamerV3 `size200m` configuration (Hafner et al., 2025). The world model and behavior policy are trained jointly during phase one and frozen during ODWVI critic training. The behavior policy uses three hidden layers of width 1024, while the categorical critic uses three hidden layers of width 512. Training uses trajectory sequences of length 64 while preserving recurrent state across consecutive chunks and resetting it at episode boundaries. The target critic is initialized from the online critic and updated after every critic update according to $\bar{\theta} \leftarrow (1-\tau)\bar{\theta} + \tau\theta$. The training hyperparameters are reported in Table 5. Dataset construction and evaluation procedures are described in Appendices D and E.

ODWVI-CQL ODWVI-CQL trains a categorical critic on dataset transitions encoded by the shared frozen world model, combining the distributional Bellman loss with the CQL penalty $\log \sum_a \exp Q(\hat{b}, a) - Q(\hat{b}, a_{\text{data}})$. We use a regularization coefficient of 0.1, following JOWA (Cheng et al., 2025). Bootstrap action selection uses the full action space; the behavior-support restriction, left shift, and unsupported-action margin penalty are omitted.