SEMI-IMPLICIT BELIEF APPROXIMATION WITH LEARNED TRANSPORT AND REWEIGHTING FOR POMDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Belief estimation in POMDPs requires approximating the posterior over latent states given a history of actions and observations. In many domains, however, the observations carry no information about the state for many consecutive timesteps, and the belief's errors accumulate. Keeping these errors small requires consistency in carrying the belief forward until the next informative observation and expressiveness in correcting it in a single update when that observation arrives. Existing belief representations assume that informative observations arrive regularly: they either rely on ground-truth states during training, or lose the hypothesis that the next informative observation supports. We propose Generative Belief Variational Inference (GBVI), built on the insight that any change in a belief decomposes into transport, which moves probability mass, and reweighting, which adjusts it in place. Semi-implicit variational inference splits the belief into components that form a density without a closed form, a memory-conditioned Hamiltonian flow transports them across timesteps, and a gated recurrent memory reweights them, with both operations learned from observations alone. On Symmetric Light-Dark, occluded Atari, and Area Search, GBVI outperforms recurrent, variational, particle-based, and state-supervised baselines, and on Area Search its belief error against the exact posterior stays low as the period without informative observations grows while the baselines' error rises.

1 INTRODUCTION

Partially Observable Markov Decision Processes (POMDPs) (Åström, 1965) model sequential decision-making under uncertainty about the underlying state. The central object is the belief, the probability distribution over latent states at each timestep conditioned on the history of past actions and observations up to that point (Kaelbling et al., 1998). The Bayes update (Kaelbling et al., 1998) propagates it through the transition dynamics and corrects it with each observation. In many domains, however, observations keep arriving but contain no features relevant to the latent state for many consecutive timesteps (Ma et al., 2020). Examples include a tracked object that stays behind an occluder (Makansi et al., 2019) and a search target that remains outside the sensor's range (Zhang et al., 2024). Over such periods, an approximate belief relies on prediction, with little correction from the observations, and its errors compound from one timestep to the next (Cao et al., 2023; Ma et al., 2020; Sinopoli et al., 2004).

Keeping these errors small is hard because it places two requirements on the belief representation. The first is belief consistency. Until the next informative observation arrives, what is known about the state comes from earlier observations, so the representation has to carry it forward from one timestep to the next, whereas the current observations carry no relevant features and must not overwrite it (De et al., 2024; Ma et al., 2020; Shchendrigin et al., 2026). The second is expressiveness: when an informative observation finally arrives, the belief must be corrected at once, which is possible only if the state that this observation supports is still among its hypotheses (Olson et al., 2023; Ma et al., 2019).

Existing belief representations assume that informative observations arrive regularly or are missing only briefly, and have no mechanism for keeping the belief accurate when this assumption fails. The

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.

first group learns the belief update from ground-truth states during training (Wang et al., 2023; Avalos et al., 2024), but such states are rarely available outside simulation (Pinto et al., 2017; Ebi et al., 2026), and supervising the update requires labelled trajectories that cover the state space (Wang et al., 2023; Levine et al., 2020). The second group learns from observations alone but cannot correct the belief in a single update when an informative observation arrives, because the hypothesis it supports is no longer in the belief. Recurrent encoders compress the history into a single vector (Hausknecht & Stone, 2017; Kapturowski et al., 2018; Ni et al., 2023; Lu et al., 2023), which holds one hypothesis at a time. Variational methods fit an explicit distribution with a per-step evidence lower bound (ELBO) (Krishnan et al., 2015; Hafner et al., 2020; Chen et al., 2022; Hafner et al., 2024), which requires a tractable density (Mohamed & Lakshminarayanan, 2017; Huszár, 2017) and so bounds the shapes the belief can take, however flexible the family. Particle and set-membership methods need no density (Gordon et al., 1993; Igl et al., 2018; Ma et al., 2020; 2019), but either resample and drop a low-weight hypothesis (Arulampalam et al., 2002) before the observation that supports it arrives, or keep every hypothesis at equal weight (Zhang et al., 2025; Wei et al., 2023) and cannot shift weight to the one the observation supports.

Both requirements concern how the belief changes over time: consistency before an informative observation arrives, expressiveness when it does. The Wasserstein–Fisher–Rao (WFR) geometry (Chizat et al., 2010) describes such change: any change in a distribution decomposes into transport, which moves mass, and reweighting, which adds or removes mass in place. We therefore develop a belief representation that can be moved and reweighted. Applying this decomposition to meet the two requirements is, however, nontrivial. Neither operation is computable directly on a continuous distribution, so previous works on this geometry represent the distribution by a finite set of weighted components and derive the rules for moving and reweighting them from a target given in advance (Yan et al., 2023; Lambert et al., 2022). However, in many POMDP domains the transition and observation models are unknown, so the rules have to be learned from observations, and a belief produced by learned rules has no closed-form density. We therefore adopt semi-implicit variational inference (SIVI) (Yin & Zhou, 2018), which represents the belief by components that can be moved and reweighted on their own and trains it without a closed-form density.

Built on these insights, we propose Generative Belief Variational Inference (GBVI), whose belief at each step is the marginal of a generative process over K components. Across timesteps, a memory-conditioned Hamiltonian flow (Caterini et al., 2018) transports these components, including those with low weight, without letting two collapse, and an RG-LRU memory (De et al., 2024) reweights them, holding the weights through the gap and shifting them when an informative observation arrives. Both operations are learned from observations alone, and together they carry one set of components along the trajectory instead of redrawing it at every step. We compare GBVI against recurrent, variational, particle-based, and state-supervised baselines on Symmetric Light-Dark, on Atari under a fixed and a flickering curtain, and on Area Search, where a moving vehicle leaves the sensor's view and stays hidden for many steps. GBVI achieves the highest return across these domains. On Area Search, where the exact posterior is available, we measure belief error directly and split it into a transport part and a reweighting part: the baselines' error grows with the length of the gap, whereas GBVI's stays low on both parts.

Our main contributions are as follows:

- We identify maintaining a belief through a long period without informative observations as a problem existing belief representations are not robust to, and show that the belief update over such a period decomposes into transport and reweighting, a lens for designing and diagnosing belief representations beyond GBVI.
- We propose GBVI, a belief representation that realizes this decomposition: semi-implicit variational inference splits the belief into a set of components that form a density, a memory-conditioned Hamiltonian flow transports them, and an RG-LRU memory reweights them, with both operations learned from observations alone and no access to the true state.
- GBVI is a modular belief representation that integrates with existing reinforcement learning frameworks. Integrated with PPO (Schulman et al., 2017), it outperforms recurrent, variational, particle-based, and state-supervised baselines on Symmetric Light-Dark, occluded Atari, and Area Search, where the exact posterior lets us measure belief error directly and confirm that GBVI keeps both the transport and the reweighting error low as the gap grows.

2 Related Work

Observations that keep arriving but carry no information about the latent state over a long period have been noted in many domains. In multi-object tracking, a target hidden behind an occluder for many frames is treated as a new object when it reappears (Cao et al., 2023; Tokmakov et al., 2021). In UAV search and rescue, a target that leaves the sensor's range keeps moving while the search continues around where it was last seen (Zhang et al., 2024). In object-centric world models, an object that leaves the view is dropped and not expected back (Singh et al., 2021). In memory tasks for reinforcement learning such as T-Maze (Ni et al., 2023; Morad et al., 2023), the agent sees a cue, then walks a corridor whose observations say nothing about it, and the longer the corridor, the more often it takes the wrong branch at the end. In networked control, a Kalman filter's estimation error diverges when observations arrive too rarely (Sinopoli et al., 2004). In these settings, how far the belief can be carried through the gap, and whether it can be corrected when an observation returns, depends on how it is represented.

One line of work learns the belief update from ground-truth states during training. Believer (Wang et al., 2023) learns a state encoder from state-labelled offline data, then models a belief over that encoding separately from the policy. The Wasserstein Believer (Avalos et al., 2024) learns a latent transition model against the true states and derives the belief update from it. Asymmetric actor-critic methods (Pinto et al., 2017; Ebi et al., 2026) give the critic the state or another privileged signal, so the actor, which sees only observations, trains on a value estimate computed from the state. In tracking, PermaTrack (Tokmakov et al., 2021) learns to follow occluded objects from synthetic videos that label them while hidden. These methods need the state at training time, which simulators provide and deployment domains rarely do, and supervising the update needs labelled trajectories that cover the state space (Levine et al., 2020); GBVI is trained in a separate stage too, but from actions and observations alone.

A second line learns from observations alone but loses the hypothesis that the next informative observation will support. Recurrent encoders compress the history into a single vector, which holds one hypothesis at a time (Hausknecht & Stone, 2017; Kapturowski et al., 2018; Lu et al., 2023; Ni et al., 2023), and a Gaussian in a state-space layer holds a single mode (Luis et al., 2025). Variational methods fit an explicit belief with a per-step ELBO (Krishnan et al., 2015; Han et al., 2019; Hafner et al., 2024; Chen et al., 2022), which needs the belief's density and so fixes its family before training; learned Kalman filters (Fraccaro et al., 2017; Becker et al., 2019) carry a Gaussian through long gaps under the same constraint. Particle methods (Ma et al., 2019; 2020; Igl et al., 2018; Singh et al., 2021) hold several hypotheses but drop those with low weight at resampling (Arulampalam et al., 2002), and set-membership beliefs (Wei et al., 2023) keep every consistent hypothesis at equal weight and so cannot favour the one an observation supports. In each case the belief's form is fixed before training, whereas a correction learned from observations alone produces a belief with no closed-form density. Semi-implicit variational inference (Yin & Zhou, 2018) removes this restriction: the variational distribution is the marginal of an explicit conditional under a mixing distribution that only needs to be sampleable, so its density need not have a closed form (Huszár, 2017; Mohamed & Lakshminarayanan, 2017). We build GBVI on this construction.

A third line holds several hypotheses at each step but does not carry them from one step to the next. Per-step variational posteriors with flexible families, such as the Hamiltonian (Caterini et al., 2018) and semi-implicit (Yin & Zhou, 2018) variational autoencoders, generate a new set of components from a summary of the history at every step, as do Believer (Wang et al., 2023), which samples its belief afresh, and deep belief Markov models (Arcieri et al., 2026), which output a new belief from an explicit family, so a hypothesis that loses weight during the gap is not carried to the observation that would support it. ESCORT (Zhang et al., 2025) keeps its particles across steps but ties them only with a soft consistency regulariser that does not keep two from collapsing onto one another. Table 1 groups the baselines by this distinction; carrying components across steps without collapse is what the transport in Section 3 is built for.

The Wasserstein–Fisher–Rao geometry describes how a distribution changes in two ways: mass moves from one place to another, or mass is created and removed where it is (Chizat et al., 2010; Liero et al., 2017). We call the two transport and reweighting: on weighted components, transport moves them and reweighting changes their weights (Appendix D). They match the two requirements: during the gap the belief changes by transport alone, and when an observation returns the correction

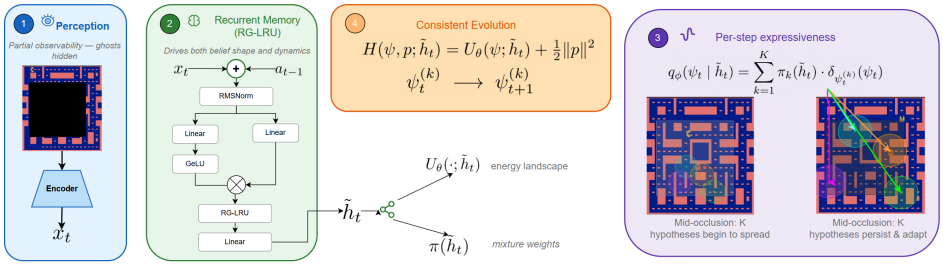


Figure 1: Overall framework of GBVI. Any change in the belief decomposes into transport and reweighting, and semi-implicit variational inference splits the belief into K components that can be moved and reweighted on their own. An encoder and an RG-LRU block (De et al., 2024) summarize the history into a state $\tilde{h}_t$, which is read out twice: as a potential U_θ for the Hamiltonian flow that transports each component without letting two collapse, and as the weights $\pi(\tilde{h}_t)$ that reweight them. Their weighted marginal is the per-step belief b_t that PPO consumes.

is a reweighting. Neither suffices alone (Yan et al., 2023): transport is slow to move mass to a distant place, and reweighting cannot create mass where there is none, the two failures of the previous paragraphs. Neither is computable on a continuous distribution, so prior work on this geometry uses such components, particles updated toward a maximum-likelihood target (Yan et al., 2023) or Gaussians moved toward a fixed posterior (Lambert et al., 2022), with the rules derived from that target. A POMDP gives no such target, since its transition and observation models are unknown, so the rules have to be learned from observations, and a belief produced by learned rules has no closed-form density. We therefore bring the semi-implicit construction above into this geometry: its components can be moved and reweighted by learned rules, and the belief they form needs no closed-form density.

Building on these insights, we realize transport with a Hamiltonian flow and reweighting with a gated recurrent memory. Recurrent networks and gradient descent on a learned potential both pull every component toward what the memory currently favours, so over a long gap the components collapse and the hypotheses they held cannot be recovered; Hamiltonian dynamics on a memory-conditioned potential (Neal, 2012; Caterini et al., 2018) preserve volume, so the components follow the history without collapsing over a gap of any length. A volume-preserving map moves no mass between components, so the weights come from a gated recurrent memory (De et al., 2024) that holds its state while observations are uninformative and rewrites it when an informative one arrives.

3 Approach

We present *G*enerative *B*elief *V*ariational *I*nference (GBVI), a variational framework for approximating beliefs in POMDPs (Figure 1; notation in Appendix B, per-timestep update in Algorithm 1, assumptions in Appendix E.1). It rests on two insights: any change in the belief decomposes into transport and reweighting, and semi-implicit variational inference gives components that can be moved and reweighted on their own while forming a belief with no closed-form density. GBVI extends SIVI from a single posterior to a sequence of beliefs by letting one set of components persist across timesteps. A memory-conditioned Hamiltonian flow (Neal, 2012) transports the components through periods without informative observations, which keeps the belief consistent, and a gated recurrent memory (De et al., 2024) reweights them in a single update when an observation arrives, which keeps it expressive. Both maps are learned from observations alone.

3.1 Belief as a Set of Components

Components. In a POMDP the belief $b_t(s) = p(s_t = s \mid h_t)$ is the posterior over the hidden state given the history h_t, which the Bayes update predicts through the transition and corrects with each observation. GBVI approximates it over a latent state $z_t \in \mathcal{Z} \subseteq \mathbb{R}^d$, with an encoder mapping each observation o_t to a feature x_t. Following Section 2, the belief is represented by K weighted Gaussian components,

$$\psi_t = (\mu_t, \log \sigma_t) \in \mathbb{R}^{2d}, \qquad q(z_t \mid \psi_t) = \mathcal{N}\big(z_t \mid \mu_t, \mathrm{diag}(\sigma_t)^2\big)\,. \tag{1}$$

Each component is one hypothesis about the latent state, and the belief is $\sum_{k=1}^{K} \pi_k \, q(z_t \mid \psi_t^{(k)})$ with $\pi \in \Delta^{K-1}$: transport moves the components and reweighting changes π.

This decomposition is the structure of the Bayes update itself.

Proposition 1. *One step of the Bayes update is a transport that carries each component through the transition to its prediction $\bar{q}_k$ with the weights unchanged, then a reweighting that scales mass in place by the likelihood $O(o_{t+1} \mid \cdot, a_t)$: it maps $\pi_k \mapsto \pi_k \, \ell_k / \sum_j \pi_j \ell_j$ by the evidence ℓ_k of o_{t+1} under $\bar{q}_k$, and reshapes component k in place to $\tilde{q}_k \propto O(o_{t+1} \mid \cdot, a_t) \, \bar{q}_k$. If o_{t+1} carries no information about the state, $O(o_{t+1} \mid \cdot, a_t)$ is constant, all ℓ_k are equal, and the reweighting is the identity.*

Corollary 2. *A representation that drops components over L steps without informative observations has, at the next one, a WFR error of at least c times the dropped mass, and this bound is nondecreasing in L.*

Proofs are in Appendix E. The proposition holds for any transition and observation model, so it fixes what has to be learned when they are unknown: a map that moves components and a map that changes their weights, holding them while observations are uninformative. A POMDP gives no target to derive either map from (Section 2), so both have to be learned, and a belief produced by learned maps has no closed-form density on which the standard ELBO could be evaluated. Semi-implicit variational inference (Yin & Zhou, 2018) needs a closed-form density only of the conditional $q(z_t \mid \psi_t)$ in Eq. (1) and lets the distribution over ψ_t be implicit. Its bound survives components carried and reweighted by learned maps (Theorem 3).

Mixing distribution. Standard SIVI redraws ψ from noise at every step (Yin & Zhou, 2018); GBVI instead moves each component from its previous position, as the proposition requires, along a flow Φ on a potential $U_\theta(\,\cdot\,; \tilde{h}_t)$ read from a memory summary $\tilde{h}_t$ of the history (Section 3.3), with a fresh momentum,

$$\psi_t^{(k)} = \Phi\Big(\psi_{t-1}^{(k)}, \, p_t^{(k)}; \, U_\theta(\,\cdot\,; \tilde{h}_t)\Big), \qquad p_t^{(k)} \sim \mathcal{N}(0, I). \tag{2}$$

The mixing distribution of component k, $q_\phi(\psi \mid \psi_{t-1}^{(k)}, \tilde{h}_t)$, is the law of Eq. (2) over the momentum: it is implicit, and the history reaches the belief by moving the components.

Marginal belief. Marginalizing the components gives the per-step belief

$$h_\phi(z_t \mid \psi_{t-1}, \tilde{h}_t) = \sum_{k=1}^{K} \pi_k(\tilde{h}_t) \int q(z_t \mid \psi) \, q_\phi(\psi \mid \psi_{t-1}^{(k)}, \tilde{h}_t) \, d\psi, \tag{3}$$

which has no closed-form density; its shape is set by where the components are and how they are weighted, not by a family chosen in advance, and with $K = 1$ it collapses to a single Gaussian. Sections 3.2, 3.3 and 3.4 give the flow, the weights, and the training.

3.2 Transport: Memory-Conditioned Hamiltonian Flow

The flow Φ of Eq. (2) realizes the transport of Proposition 1: it carries each component from its previous position to the next, leaving the weights alone, through a period of any length. The observation-only methods of Section 2 fail here by dropping a component or letting two collapse, which Corollary 2 says cannot be undone, so transport has to follow the history without letting two components collapse.

The memory state $\tilde{h}_t$ (Section 3.3) parameterizes a learnable potential $U_\theta(\psi; \tilde{h}_t)$ over the component space, with θ shared across timesteps, whose low-energy regions are the configurations the history supports. While observations are uninformative, the memory holds $\tilde{h}_t$ and the components keep moving on the same energy surface. What was known before the gap carries them, and nothing overwrites it, which is the consistency requirement. Pulling the components down this potential would draw them together, so GBVI gives each a momentum $p \in \mathbb{R}^{2d}$ and moves it under the Hamiltonian

$$H(\psi, p; \tilde{h}_t) = U_\theta(\psi; \tilde{h}_t) + \frac{1}{2}\|p\|^2. \tag{4}$$

A fresh $p_t^{(k)} \sim \mathcal{N}(0, I)$ per component and timestep gives a random push that the gradient of the potential bends, so a component can reach configurations the history newly supports rather than only descend. The components move from $\psi_t^{(k)}$ to $\psi_{t+1}^{(k)}$ by L Leapfrog steps of size ϵ (Eq. (15), Algorithm 1). The Leapfrog map is symplectic and preserves volume exactly at any step size (Neal, 2012; Leimkuhler & Reich, 2005), so two distinct components are never mapped to the same position and the K hypotheses stay distinct over a period of any length, keeping the mass of Corollary 2 in the belief. Discretization changes where the components go, not this property: with the potential C^1 and ϵ bounded, the distance between two components shrinks by at most a bounded factor per step. Since each component moves a bounded distance per step, GBVI assumes that the latent state evolves continuously between timesteps; the proof, the scope, and the background on Hamiltonian dynamics are in Appendices E and D.

3.3 Reweighting: Gated Recurrent Memory

The reweighting of Proposition 1 has two parts, and the memory realizes the one that moves mass between components: in the exact update it scales each weight by the evidence of the observation under that component, and does nothing when the observation is uninformative. The evidence cannot be computed without the observation model, so GBVI learns a map that holds the weights through uninformative observations and shifts them when an informative one arrives. The flow cannot be this map, since a volume-preserving map moves no mass between components. GBVI uses the RG-LRU (De et al., 2024), a gated linear recurrence (Appendix D) suited to partial observability (Lu et al., 2024). At each timestep the memory takes the encoder feature x_t and the previous action a_{t-1}, concatenated as $\hat{x}_t = [x_t; a_{t-1}]$, and updates its state by

$$\tilde{h}_t = \alpha_t \odot \tilde{h}_{t-1} + \sqrt{1 - \alpha_t^2} \odot (i_t \odot \hat{x}_t), \tag{5}$$

with a learned initial state $\tilde{h}_0$, an input gate i_t, and a decay α_t set by a recurrence gate r_t, all computed from $\hat{x}_t$ (Eq. (13)). The decay interpolates between keeping the previous state and overwriting it with the gated input; when the input carries no information about the state, the gates keep α_t close to one and the state is held, which is the identity reweighting of the proposition.

The state $\tilde{h}_t$ is read out twice. A softmax-normalized linear projection gives the weights $\pi(\tilde{h}_t) \in \Delta^{K-1}$, one per component, and a second linear projection gives the conditioning signal of the potential $U_\theta(\,\cdot\,; \tilde{h}_t)$ in Section 3.2. Both readouts come from the same state, so a component that the potential moves toward a configuration the history supports gains weight at the same time. The weights are read from $\tilde{h}_t$ afresh at each step, not multiplied onto the previous π as in a Fisher–Rao gradient flow (Yan et al., 2023). When an informative observation arrives, the softmax can therefore move mass onto the components it supports in a single update, however low their weight had become. The other components lose weight but stay in the belief and keep being transported.

3.4 Training Objective and Policy Optimization

The belief of Eq. (3) has no closed-form density, so the standard ELBO cannot be evaluated on it, but given the momenta $p_t^{(1)}, \ldots, p_t^{(K)}$ of Eq. (2) it is the explicit mixture $m_t(z_t) = \sum_{k=1}^K \pi_k(\tilde{h}_t)\, q(z_t \mid \psi_t^{(k)})$, and the implicitness comes from the momenta alone. We therefore draw $k \sim \pi(\tilde{h}_t)$, $z_t \sim q(z_t \mid \psi_t^{(k)})$, and form

$$\mathcal{L}^{(t)} \;=\; \mathbb{E}\left[\log \frac{p(o_t, z_t)}{\sum_{k=1}^K \pi_k(\tilde{h}_t)\, q(z_t \mid \psi_t^{(k)})}\right], \qquad \mathcal{L} \;=\; \sum_{t=1}^T \mathcal{L}^{(t)}, \tag{6}$$

with the expectation over the momenta and z_t.

Theorem 3. *For any distribution of the momenta $p_t^{(1)}, \ldots, p_t^{(K)}$ in Eq. (2), and hence for components carried by the flow and weighted by $\pi(\tilde{h}_t)$, $\mathcal{L}^{(t)} \le \log p(o_t \mid \psi_{t-1}, \tilde{h}_t)$.*

The proof is in Appendix E; it uses the convexity of the KL divergence and does not need the components to be independent across timesteps. The objective evaluates densities only through $q(z_t \mid \psi_t)$ and needs only $p(o_t, z_t)$, so the belief is learned from actions and observations with no

Table 1: Success rate on Light-Dark and return on Atari under the fixed (FI) and flickering (FL) curtain, mean$_{\pm\text{standard error}}$ over 100 seeds, grouped by failure mode (Section 4.1); † state-supervised; bold is the best per row.

		Single hypothesis		*Not carried through the gap*			*State-supervised*	
Environment	**GBVI**	**RG-LRU**	**LRU**	**HVAE**	**SIVAE**	**ESCORT**	**WBU†**	**Believer†**
LightDark	$\mathbf{0.872}_{\pm 0.011}$	$0.807_{\pm 0.013}$	$0.778_{\pm 0.018}$	$0.819_{\pm 0.013}$	$0.797_{\pm 0.016}$	$0.826_{\pm 0.023}$	$0.821_{\pm 0.021}$	$0.832_{\pm 0.014}$
Pong (FI)	$\mathbf{17.52}_{\pm 0.14}$	$16.50_{\pm 2.36}$	$11.94_{\pm 2.68}$	$10.75_{\pm 3.62}$	$12.09_{\pm 2.65}$	$14.66_{\pm 2.69}$	$15.02_{\pm 1.14}$	$11.74_{\pm 2.94}$
Pong (FL)	$\mathbf{17.08}_{\pm 0.12}$	$15.26_{\pm 2.44}$	$12.27_{\pm 3.22}$	$12.00_{\pm 3.05}$	$11.92_{\pm 2.92}$	$15.89_{\pm 1.13}$	$15.94_{\pm 2.11}$	$14.93_{\pm 1.97}$
MsPacman (FI)	$\mathbf{1566.4}_{\pm 49.9}$	$1331.2_{\pm 43.7}$	$969.5_{\pm 220.5}$	$939.2_{\pm 63.4}$	$1321.4_{\pm 49.0}$	$1367.2_{\pm 37.4}$	$1297.7_{\pm 79.6}$	$1041.5_{\pm 57.6}$
MsPacman (FL)	$1511.7_{\pm 39.2}$	$\mathbf{1583.7}_{\pm 50.3}$	$1016.0_{\pm 76.1}$	$882.3_{\pm 37.0}$	$1371.6_{\pm 62.0}$	$1410.4_{\pm 44.2}$	$1437.6_{\pm 39.7}$	$1426.6_{\pm 47.9}$
Bowling (FI)	$\mathbf{33.37}_{\pm 0.47}$	$30.02_{\pm 0.10}$	$29.48_{\pm 0.24}$	$28.98_{\pm 0.36}$	$27.61_{\pm 0.24}$	$28.79_{\pm 0.36}$	$29.28_{\pm 0.52}$	$29.33_{\pm 0.25}$
Bowling (FL)	$\mathbf{34.33}_{\pm 0.75}$	$32.14_{\pm 0.36}$	$32.84_{\pm 0.49}$	$30.34_{\pm 0.53}$	$31.10_{\pm 0.37}$	$29.93_{\pm 0.57}$	$32.66_{\pm 0.35}$	$31.79_{\pm 0.33}$
BeamRider (FI)	$\mathbf{1257.4}_{\pm 43.7}$	$1053.3_{\pm 37.9}$	$985.8_{\pm 38.4}$	$720.5_{\pm 48.8}$	$850.2_{\pm 49.9}$	$902.7_{\pm 64.7}$	$997.4_{\pm 47.8}$	$896.5_{\pm 53.3}$
BeamRider (FL)	$916.4_{\pm 27.5}$	$887.9_{\pm 42.9}$	$726.1_{\pm 16.2}$	$741.6_{\pm 11.8}$	$677.6_{\pm 21.1}$	$\mathbf{944.2}_{\pm 97.4}$	$936.9_{\pm 23.2}$	$904.2_{\pm 86.9}$
Asteroids (FI)	$\mathbf{1487.0}_{\pm 31.6}$	$1383.2_{\pm 40.6}$	$1210.8_{\pm 50.2}$	$1176.7_{\pm 31.5}$	$1075.4_{\pm 23.8}$	$1147.3_{\pm 37.9}$	$1319.8_{\pm 41.7}$	$1228.4_{\pm 46.1}$
Asteroids (FL)	$1495.2_{\pm 41.7}$	$1419.5_{\pm 48.6}$	$1288.0_{\pm 44.8}$	$1221.4_{\pm 36.7}$	$1334.1_{\pm 65.7}$	$\mathbf{1504.0}_{\pm 57.2}$	$1428.3_{\pm 49.3}$	$1459.6_{\pm 36.2}$

access to the state, the only difference in training between GBVI and the state-supervised baselines of Section 4.1.

The belief network is trained with this objective and then frozen, so the policy trains on a fixed belief. We integrate GBVI with PPO (Schulman et al., 2017): the K components $\{(\mu_t^{(k)}, \sigma_t^{(k)})\}_{k=1}^{K}$ and their weights $\pi(\tilde{h}_t)$ are flattened into a belief summary $b_t \in \mathbb{R}^{K(2d+1)}$ and concatenated with x_t as input to the actor and critic (Appendix G.1).

4 Experiments

We evaluate GBVI against seven baselines on Symmetric Light-Dark (Platt et al., 2010), five Atari games (Bellemare et al., 2013) under fixed and flickering curtains, and Area Search, where the exact posterior gives the belief error, and ablate its design choices on Bowling.

4.1 Baselines

We implement seven baselines that cover the three failure modes of Section 2. Believer (Wang et al., 2023) and WBU (Avalos et al., 2024) learn the belief update from ground-truth states, marked with a dagger in Table 1. RG-LRU (De et al., 2024) and LRU (Orvieto et al., 2023) compress the history into a single recurrent state, which holds one hypothesis at a time. HVAE (Caterini et al., 2018) and SIVAE (Yin & Zhou, 2018) re-instantiate their components at every step, as does Believer, and ESCORT (Zhang et al., 2025) carries its particles across steps but ties them only with a soft consistency regulariser that does not keep them from collapsing. All baselines share the observation encoder and the PPO stage with GBVI and differ only in the belief passed to the actor and critic.

4.2 Return on Light-Dark and Occluded Atari

Symmetric Light-Dark (Platt et al., 2010) is a continuous-control task in which the agent has to reach one of two symmetric goals, and the observations tell them apart only inside a lit region away from both. Outside it the belief has to hold both hypotheses and carry them through the agent's own motion; a single observation inside decides between them, and a belief that does not correct immediately sends the agent to the wrong goal.

The five Atari games (Bellemare et al., 2013), Pong, MsPacman, Bowling, BeamRider, and Asteroids, are made partially observable by a rectangular curtain over a task-relevant region of the frame. Under the **fixed curtain** (FI) the region stays blanked for the whole episode, so whatever moves behind it is observed again only when it leaves, and until then its belief is carried by prediction alone. Under the **flickering curtain** (FL) the region is blanked or revealed in random intervals of a few timesteps (Appendix G.2), so each reveal has to correct the belief in a single update before the curtain closes again. The two modes thus differ in the length of the gap, and we compare each method's return across them to see how it holds up as the gap grows.

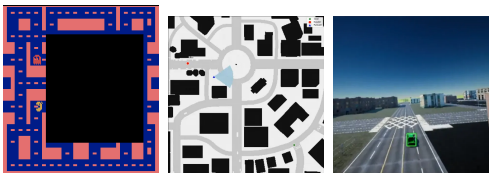


Figure 2: Evaluation domains. Left: MsPacman under the fixed curtain, which blanks the region where the ghosts move. Middle: Area Search, the AirSim city on a 2D grid with buildings as obstacles, the UAV's sensor region, and a vehicle out of view. Right: the UAV's view in AirSim.

Table 2: Belief error on Area Search: WFR^2 to the exact posterior, its transport and reweighting parts, and the mass on the vehicle's cell when it returns to view. †State-supervised.

Method	WFR^2	Transport	Reweighting	Mass at return
GBVI	$\mathbf{0.11}_{\pm 0.02}$	$\mathbf{0.06}_{\pm 0.01}$	$\mathbf{0.05}_{\pm 0.01}$	$\mathbf{0.23}_{\pm 0.02}$
RG-LRU	$0.46_{\pm 0.14}$	$0.10_{\pm 0.03}$	$0.36_{\pm 0.13}$	$0.09_{\pm 0.03}$
WBU†	$0.30_{\pm 0.05}$	$0.08_{\pm 0.01}$	$0.22_{\pm 0.04}$	$0.13_{\pm 0.03}$
ESCORT	$0.42_{\pm 0.14}$	$0.22_{\pm 0.07}$	$0.20_{\pm 0.09}$	$0.08_{\pm 0.04}$
HVAE	$0.75_{\pm 0.18}$	$0.30_{\pm 0.08}$	$0.45_{\pm 0.12}$	$0.05_{\pm 0.02}$
Believer†	$0.61_{\pm 0.13}$	$0.26_{\pm 0.09}$	$0.35_{\pm 0.09}$	$0.08_{\pm 0.03}$

All methods share the encoder, the PPO implementation (Raffin et al., 2021), and all hyperparameters, train for 10^7 environment steps, and are evaluated over 100 episodes on shared seeds (Appendices G and H). Table 1 reports the mean and standard error, success rate on Light-Dark and return on the ten Atari configurations, and supports four conclusions.

GBVI attains the highest return in every configuration or ties for it, by a significant margin on Light-Dark and on seven of the ten Atari configurations and within error of the best baseline on the other three, all under the flickering curtain, where the gap is short, and none under the fixed curtain. *GBVI is trained on actions and observations alone and still outperforms the baselines trained with ground-truth states:* Believer and WBU are below GBVI on Light-Dark and on every fixed-curtain configuration, and tie with it only on BeamRider under the flickering curtain. State supervision does not substitute for carrying the hypotheses through the gap: Believer, which samples its belief afresh at each step, stays in the range of HVAE and SIVAE despite it.

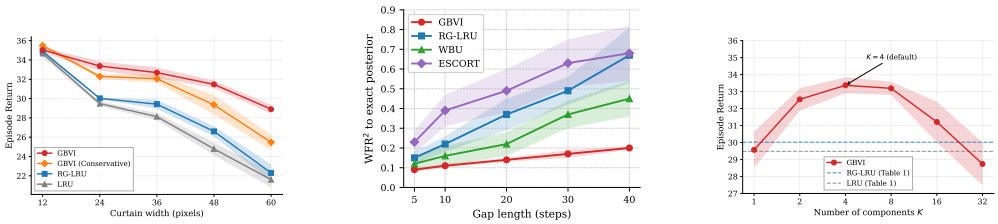
The advantage of GBVI grows with the length of the gap: almost every baseline trails GBVI by more under the fixed curtain than under the flickering one, and the methods that re-instantiate their components at every step trail by the most; the one clear exception is RG-LRU on Pong, where the ball's motion is close to deterministic, so one hypothesis suffices and a recurrent vector loses little. *When the informative observation arrives, GBVI corrects the belief in a single update:* on Light-Dark the first observation in the lit region has to move the belief onto the correct goal at once, and GBVI reaches 0.872, ahead of every baseline, while the single-vector encoders RG-LRU and LRU are lowest, since a vector that holds one hypothesis has nothing to resolve.

4.3 Belief Error on Area Search

To measure belief error directly, we also compare GBVI with the baselines on Area Search, a UAV search-and-rescue domain following Zhang et al. (2024): a UAV searches the AirSim city (Shah et al., 2017) for two moving vehicles with a limited-range, line-of-sight sensor (Figure 2). Each vehicle follows the roads and produces a noisy position reading only while inside the sensor region; once it turns a corner or leaves the range, the belief has to be carried by the motion model along every road it could have taken until it comes back into view. How long it stays out of view is set by the scenario and is the gap length we vary.

The vehicles' motion and sensor models are known, so the exact posterior is a grid filter per vehicle. All methods are the belief networks of Section 4.1, trained on Area Search trajectories as in Section 4.2, evaluated on the same fixed UAV route and readings within a scenario, and read out to a distribution over cells by a probe of the same form for every method (Wang et al., 2023), fit on training episodes and frozen before evaluation.

We measure the Wasserstein–Fisher–Rao distance (Chizat et al., 2010; Liero et al., 2017) between the probed belief and the exact posterior. The ground cost is the shortest road path, and the cutoff is one block: beyond it, mass is removed and re-created rather than moved. The optimal coupling splits the squared distance into a transport part, mass in the wrong place, and a reweighting part, mass missing or in excess (Appendix G.3). Table 2 reports both parts averaged over steps and episodes, together with the mass on the vehicle's cell on the last step before it returns to view, and supports the following.



(a) (b) (c)

Figure 3: (a) Return on Bowling against curtain width for GBVI, its transport-only variant, RG-LRU and LRU. (b) Belief error on Area Search against gap length, for GBVI and one baseline per group of Table 1. (c) Return on Bowling against the number of components K at width 24.

GBVI has the lowest belief error, and the split locates where each baseline fails. Its WFR^2 of 0.11 is a third of the next best, WBU at 0.30, with both parts below 0.07; RG-LRU and WBU lose almost entirely on reweighting, mass on the wrong road, while ESCORT, HVAE and Believer lose on both parts. On the last step before the vehicle returns to view, GBVI holds 0.23 of its mass on the vehicle's cell against at most 0.13 for any baseline, so the hypothesis the returning observation supports is still in its belief.

The baselines' error grows with the gap, and GBVI's rises far more slowly. Against the gap length (Figure 3b), RG-LRU's error rises from 0.15 at 5 steps to 0.67 at 40, WBU's to 0.45 and ESCORT's to 0.68, whereas GBVI's rises from 0.09 to 0.20; at 5 steps WBU is within 0.03 of GBVI, so the advantage is in carrying the hypotheses through the gap (Appendix F.5).

4.4 Ablation Study

Figure 3 reports two ablations on Bowling under the fixed curtain (5M steps, 100 episodes per configuration). *The advantage of GBVI grows with the gap, and transport alone carries part of it:* widening the curtain from 12 to 60 pixels lengthens the time the ball spends behind it, GBVI's return falls the least (width 60: 28.9 against RG-LRU 22.3 and LRU 21.6; Figure 3a), and the transport-only variant, which keeps the Hamiltonian flow but fixes the weights, sits between them, since carrying the components through the gap helps but without reweighting the belief cannot shift mass onto the component the reveal supports (Appendix F.2). Replacing the Hamiltonian flow by a Langevin SDE over the K components or a continuous Langevin chain reaches comparable returns; Leapfrog is kept for its volume preservation and narrowest bands, a choice of stability rather than a structural requirement (Appendix F.3). *The belief needs more than one component:* sweeping $K \in \{1, 2, 4, 8, 16, 32\}$ at width 24 gives an inverted U (Figure 3c), with $K{=}1$ collapsing GBVI to a single closed-form Gaussian at 29.6, level with RG-LRU (30.0) and LRU (29.5), the peak at $K{=}4$ (33.4) and $K{=}8$ (33.2), and $K{=}32$ falling to 28.7 as the weight spread over thirty-two components is too thin for the policy to read; the rise from $K{=}1$ is the gain from letting the belief hold more than one hypothesis, which is what the semi-implicit construction of Section 3 exists for (Appendix F.4).

5 Conclusion

We address belief approximation in POMDPs over long periods in which observations arrive but carry no information about the latent state. Over such a period the belief has to be carried forward consistently, and when an informative observation returns it has to be corrected in a single update. Any change in a belief decomposes into transport, which moves mass, and reweighting, which adjusts it in place. Generative Belief Variational Inference (GBVI) realizes this decomposition on a set of semi-implicit components: a memory-conditioned Hamiltonian flow transports them without letting two collapse, and a gated recurrent memory reweights them, both learned from observations alone. GBVI attains the highest return or ties for it on Symmetric Light-Dark and on Atari under fixed and flickering curtains. On Area Search its error against the exact posterior is lowest on both parts and grows far more slowly with the gap than every baseline's. The decomposition also locates where each baseline fails, a lens for designing belief representations beyond GBVI. Due to space constraints, we defer the discussion of limitations and future work to Appendix A.

ETHICS STATEMENT

This work studies belief approximation in simulated partially observable environments. All experiments use publicly available Atari games and two simulated tasks; no human subjects, personal data, or sensitive information are involved. The method is a general-purpose belief representation for reinforcement learning, and we do not foresee harmful applications specific to this work. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. The main paper and appendix describe the proposed method, experimental setup, environments, evaluation procedures, and implementation details needed to understand and reproduce our experiments: the per-timestep update is given in Algorithm 1, the assumptions and proofs of the theoretical results in Appendices E.1 and E, the architectures, training procedure, hyperparameters, curtain placements, preprocessing, and evaluation protocol with seeds in Appendix G, and the Area Search environment, exact posterior, probe, and belief-error metric in Appendix G.3. In addition, we provide the source code as anonymous supplementary material, including the code needed to train and evaluate our method and the baselines and to reproduce the main results reported in the paper.

AI USE STATEMENT

In this work, we have not used generative AI tools to generate synthetic datasets, to develop the conceptual framework or the research questions, to propose hypotheses, to clean or reformat datasets, or to run experiments; assisting with translation is not applicable to this work. We used generative AI tools for the following tasks. For writing, we used them to improve the readability of the paper, and to condense the text to the page limit. For retrieval and discovery, we used them to summarize existing literature and to locate related work, and we manually verified every reference. For research execution, we used them to assist in implementing the baseline methods and the controlled-gap evaluation script of Appendix F.5, and to produce the code for the tables and figures; all AI-assisted code was reviewed and tested by the authors. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Giacomo Arcieri, Konstantinos G. Papakonstantinou, Daniel Straub, and Eleni Chatzi. Deep belief markov models for pomdp inference. *Neural Netw.*, 196(C), June 2026. ISSN 0893-6080. doi: 10.1016/j.neunet.2025.108386. URL https://doi.org/10.1016/j.neunet.2025.108386.

M. Sanjeev Arulampalam, Simon Maskell, Neil J. Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. Signal Process.*, 50: 174–188, 2002. URL https://api.semanticscholar.org/CorpusID:55577025.

Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965. URL https://api.semanticscholar.org/CorpusID:121222106.

Raphaël Avalos, Florent Delgrange, Ann Nowe, Guillermo Perez, and Diederik M Roijers. The wasserstein believer: Learning belief updates for partially observable environments through reliable latent space models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KrtGfTGaGe.

Philipp Becker, Harit Pandya, Gregor Gebhardt, Cheng Zhao, C. James Taylor, and Gerhard Neumann. Recurrent kalman networks: Factorized inference in high-dimensional deep feature spaces. 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. 544–552. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/becker19a.html.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, June 2013. ISSN 1076-9757. doi: 10.1613/jair.3912. URL http://dx.doi.org/10.1613/jair.3912.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9686–9696, 2023.

Anthony L. Caterini, Arnaud Doucet, and Dino Sejdinovic. Hamiltonian variational auto-encoder, 2018. URL https://arxiv.org/abs/1805.11328.

Xiaoyu Chen, Yao Mu, Ping Luo, Shengbo Li, and Jianyu Chen. Flow-based recurrent belief state learning for pomdps, 2022. URL https://arxiv.org/abs/2205.11051.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18:1 – 44, 2010. URL https://api.semanticscholar.org/CorpusID:46880773.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. URL https://arxiv.org/abs/1406.1078.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing gated linear recurrences with local attention for efficient language models, 2024. URL https://arxiv.org/abs/2402.19427.

Daniel Ebi, Gaspard Lambrechts, Damien Ernst, and Klemens Böhm. Informed asymmetric actor-critic: Leveraging privileged signals beyond full-state access, 2026. URL https://arxiv.org/abs/2509.26000.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3604–3613, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Neil J. Gordon, David Salmond, and Adrian F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113, 1993. URL https://api.semanticscholar.org/CorpusID:12644877.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination, 2020. URL https://arxiv.org/abs/1912.01603.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. URL https://arxiv.org/abs/2301.04104.

Dongqi Han, Kenji Doya, and Jun Tani. Variational recurrent models for solving partially observable control tasks, 2019. URL https://arxiv.org/abs/1912.10703.

Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps, 2017. URL https://arxiv.org/abs/1507.06527.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735.

Ferenc Huszár. Variational inference using implicit distributions, 2017. URL https://arxiv.org/abs/1702.08235.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. 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. 2117–2126. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/igl18a.html.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101:99–134, 1998. URL https://api.semanticscholar.org/CorpusID:5613003.

Steven Kapturowski, Georg Ostrovski, John Quan, Rémi Munos, and Will Dabney. Recurrent experience replay in distributed reinforcement learning. In *International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:59345798.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters, 2015. URL https://arxiv.org/abs/1511.05121.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 14434–14447. Curran Associates, Inc., 2022. doi: 10.52202/068431-1049. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5d087955ee13fe9a7402eedec879b9c3-Paper-Conference.pdf.

Benedict Leimkuhler and Sebastian Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020. URL https://arxiv.org/abs/2005.01643.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, December 2017. ISSN 1432-1297. doi: 10.1007/s00222-017-0759-8. URL http://dx.doi.org/10.1007/s00222-017-0759-8.

Chenhao Lu, Ruizhe Shi, Yuyao Liu, Kaizhe Hu, Simon S. Du, and Huazhe Xu. Rethinking transformers in solving pomdps, 2024. URL https://arxiv.org/abs/2405.17358.

Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Carlos E. Luis, Alessandro G. Bottero, Julia Vinogradska, Felix Berkenkamp, and Jan Peters. Uncertainty representations in state-space layers for deep reinforcement learning under partial observability, 2025. URL https://arxiv.org/abs/2409.16824.

Xiao Ma, Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter recurrent neural networks, 2019. URL https://arxiv.org/abs/1905.12885.

Xiao Ma, Peter Karkus, David Hsu, Wee Sun Lee, and Nan Ye. Discriminative particle filter reinforcement learning for complex partial observations, 2020. URL https://arxiv.org/abs/2002.09884.

Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7137–7146. IEEE, June 2019. doi: 10.1109/cvpr.2019.00731. URL http://dx.doi.org/10.1109/CVPR.2019.00731.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models, 2017. URL https://arxiv.org/abs/1610.03483.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POPGym: Benchmarking partially observable reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=chDrutUTs0K.

Radford Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 06 2012. doi: 10.1201/b10905-6.

Tianwei Ni, Michel Ma, Benjamin Eysenbach, and Pierre-Luc Bacon. When do transformers shine in rl? decoupling memory from credit assignment, 2023. URL https://arxiv.org/abs/2307.03864.

Elizabeth A. Olson, Jana Pavlasek, Jasmine A. Berry, and Odest Chadwicke Jenkins. Counterhypothetical particle filters for single object pose tracking. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3853–3859, 2023. doi: 10.1109/ICRA48891.2023.10160625.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences, 2023. URL https://arxiv.org/abs/2303.06349.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning, 2017. URL https://arxiv.org/abs/1710.06542.

Robert Platt, Russ Tedrake, Leslie Pack Kaelbling, and Tomas Lozano-Perez. Belief space planning assuming maximum likelihood observations. In *Robotics: Science and Systems*, 2010. URL https://api.semanticscholar.org/CorpusID:2693863.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

S. Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *International Symposium on Field and Service Robotics*, 2017. URL https://api.semanticscholar.org/CorpusID:20999239.

Oleg Shchendrigin, Egor Cherepanov, Alexey K. Kovalev, and Aleksandr I. Panov. Memory retention is not enough to master memory tasks in reinforcement learning. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '26, pp. 3344–3346, Richland, SC, 2026. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400723179. doi: 10.65109/SOSN3643. URL https://doi.org/10.65109/SOSN3643.

Gautam Singh, Skand Peri, Junghyun Kim, Hyunseok Kim, and Sungjin Ahn. Structured world belief for reinforcement learning in pomdp. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9744–9755. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/singh21a.html.

Bruno Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, Michael Jordan, and Shankar Sastry. Kalman filtering with intermittent observations. *IEEE Transactions on Automatic Control*, 49: 1453–1464, 10 2004. doi: 10.1109/TAC.2004.834121.

Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10840–10849, 2021. URL https://api.semanticscholar.org/CorpusID:232380249.

Sebastian Tschiatschek, Kai Arulkumaran, Jan Stühmer, and Katja Hofmann. Variational inference for data-efficient model learning in pomdps, 2018. URL https://arxiv.org/abs/1805.09281.

Andrew Wang, Andrew C. Li, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. Learning belief representations for partially observable deep rl. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Wei Wei, Lijun Zhang, Lin Li, Huizhong Song, and Jiye Liang. Set-membership belief state-based reinforcement learning for POMDPs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36856–36867. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wei23d.html.

Yuling Yan, Kaizheng Wang, and Philippe Rigollet. Learning gaussian mixtures using the wasserstein-fisher-rao gradient flow. *ArXiv*, abs/2301.01766, 2023. URL https://api.semanticscholar.org/CorpusID:255416050.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. 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. 5660–5669. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/yin18b.html.

Yunuo Zhang, Baiting Luo, Ayan Mukhopadhyay, Daniel Stojcsics, Daniel Elenius, Anirban Roy, Susmit Jha, Miklos Maroti, Xenofon Koutsoukos, Gabor Karsai, and Abhishek Dubey. Shrinking pomcp: A framework for real-time uav search and rescue, 2024. URL https://arxiv.org/abs/2411.12967.

Yunuo Zhang, Baiting Luo, Ayan Mukhopadhyay, Gabor Karsai, and Abhishek Dubey. Escort: Efficient stein-variational and sliced consistency-optimized temporal belief representation for pomdps, 2025. URL https://arxiv.org/abs/2510.21107.

A Limitations and Future Work

GBVI represents the belief with a fixed number K of diagonal Gaussian components. A hypothesis can only be represented where a component already sits. K therefore has to match the ambiguity of the domain (Figure 3c). Letting components be created and removed is a natural extension. GBVI also assumes that the latent state evolves continuously (Assumption 12). When the state jumps, the belief falls behind. It can still recover once observations return, because the components never collapse (Appendix F.1). Finally, the belief is trained before the policy and then frozen. This isolates the effect of the representation but keeps the belief from adapting to the states the policy later visits. Fine-tuning the belief on on-policy data once the policy has stabilized is the next step.

B Notation and Abbreviations

This section provides reference tables for the mathematical notation and abbreviations used throughout the paper.

B.1 Notation

Table 3 summarizes the key symbols, organized by their role in POMDPs, semi-implicit variational inference, recurrent encoding and Hamiltonian dynamics, and the GBVI architecture.

Table 3: Summary of Notation. Symbols introduced in the SIVI and Hamiltonian dynamics background (e.g., z, ψ, $U(\psi)$, $H(\psi, p)$) are listed below in their sequential, history-conditioned form used in GBVI (e.g., z_t, $\psi_t^{(k)}$, $U_\theta(\psi; \tilde{h}_t)$, $H(\psi, p; \tilde{h}_t)$).

Symbol	Description
POMDP	
$\mathcal{S}, \mathcal{A}, \Omega$	State, action, and observation spaces
s_t, a_t, o_t	Latent state, action, and observation at timestep t
$T(s' \mid s, a)$	Transition function
$R(s, a)$	Reward function
$O(o \mid s', a)$	Observation function
γ	Discount factor, $\gamma \in [0, 1)$
h_t	Action-observation history up to timestep t
$b_t(s)$	Belief at timestep t, $b_t(s) = p(s_t = s \mid h_t)$
η	Normalization constant in Bayes' belief update
Variational Inference	
ϕ	Variational parameters
$\mathcal{L}(\phi)$	Evidence lower bound (ELBO)
Recurrent Encoding	
$\tilde{h}_t$	Recurrent hidden state at timestep t ($\tilde{h}_0$ learned)
f_θ	Parameterized recurrent update function with parameters θ
GBVI	
K	Number of mixture components
$\mathcal{Z} \subseteq \mathbb{R}^d$	Learned latent space of dimension d
z_t	Latent representation of underlying state at timestep t
$x_t, \hat{x}_t$	Encoder feature of o_t; RG-LRU input $\hat{x}_t = [x_t; a_{t-1}]$
$\psi_t^{(k)}, \psi_t$	k-th persistent component in $\mathbb{R}^{2d}$; full set $\psi_t = \{\psi_t^{(k)}\}_{k=1}^K$
$\mu_t^{(k)}, \sigma_t^{(k)}$	Mean and standard deviation of k-th Gaussian component
$\mathcal{N}(\cdot \mid \mu, \Sigma)$	Gaussian density with mean μ and covariance Σ
$q(z_t \mid \psi_t^{(k)})$	Per-component Gaussian conditional
$q_\phi(\psi_t \mid \tilde{h}_t)$	History-conditioned mixing distribution
$\pi_k(\tilde{h}_t), \pi(\tilde{h}_t)$	Mixture weight of component k; full vector in Δ^{K-1}
δ	Dirac delta function
$h_\phi(z_t \mid \tilde{h}_t)$	Marginal variational belief at timestep t
r_t, i_t, α_t	Recurrence gate, input gate, and decay factor in RG-LRU
$\mathrm{sig}(\cdot), \odot$	Sigmoid function; element-wise multiplication
$W_r, W_i, \beta_r, \beta_i, \Lambda, c$	Learnable parameters in RG-LRU
$U_\theta(\psi; \tilde{h}_t)$	Memory-conditioned potential energy
$H(\psi, p; \tilde{h}_t)$	Memory-conditioned Hamiltonian, $U_\theta + \frac{1}{2}\|p\|^2$
$p_t^{(k)}, I$	Momentum for component k, sampled from $\mathcal{N}(0, I)$
ϵ, L	Leapfrog step size and number of steps per timestep
$\nabla_\psi U_\theta$	Gradient of U_θ with respect to ψ
b_t	Flattened belief summary input to PPO, $b_t \in \mathbb{R}^{K(2d+1)}$

B.2 List of Abbreviations

Table 4 lists the abbreviations used throughout the paper, organized by their role in POMDPs, semi-implicit variational inference, recurrent encoding and Hamiltonian dynamics, and the GBVI architecture and baselines.

Table 4: List of Abbreviations.

Abbreviation	Full Form
POMDP	
POMDP	Partially Observable Markov Decision Process
Variational Inference and Belief Geometry	
VI	Variational Inference
ELBO	Evidence Lower Bound
SIVI	Semi-Implicit Variational Inference
WFR	Wasserstein–Fisher–Rao
Recurrent Encoding and Hamiltonian Dynamics	
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
LRU	Linear Recurrent Unit
RG-LRU	Real-Gated Linear Recurrent Unit
ODE	Ordinary Differential Equation
GBVI Architecture and Baselines	
GBVI	Generative Belief Variational Inference
PPO	Proximal Policy Optimization

Algorithm 1 GBVI Algorithm

Require: Previous memory $\tilde{h}_{t-1}$, components $\{\psi_{t-1}^{(k)}\}_{k=1}^{K}$, previous action a_{t-1}, observation o_t
Require: Encoder E_ϕ, RG-LRU f_θ, potential U_θ, weight head W_π, actor π_ζ, critic V_ζ
Require: Leapfrog step size ϵ, steps L

$x_t \leftarrow E_\phi(o_t)$ ▷ encode observation
$\hat{x}_t \leftarrow [x_t;\, a_{t-1}]$ ▷ memory input
$\tilde{h}_t \leftarrow f_\theta(\tilde{h}_{t-1}, \hat{x}_t)$ ▷ RG-LRU recurrence, Eq. 5
$\pi(\tilde{h}_t) \leftarrow \mathrm{softmax}(W_\pi \tilde{h}_t)$ ▷ mixture weights
for $k = 1, \ldots, K$ **do**
  $p_t^{(k)} \sim \mathcal{N}(0, I_{2d})$ ▷ fresh momentum
  $\psi_t^{(k)} \leftarrow \textsc{Leapfrog}(\psi_{t-1}^{(k)}, p_t^{(k)};\, \tilde{h}_t)$ ▷ Hamiltonian flow, Eq. 15
end for
$b_t \leftarrow \mathrm{flatten}\big(\{(\mu_t^{(k)}, \sigma_t^{(k)})\}_{k=1}^{K},\ \pi(\tilde{h}_t)\big) \in \mathbb{R}^{K(2d+1)}$ ▷ belief summary
$p(a \mid b_t) \leftarrow \pi_\zeta([x_t;\, b_t]), \quad V(b_t) \leftarrow V_\zeta([x_t;\, b_t])$ ▷ policy & value

Ensure: $\tilde{h}_t$, $\{\psi_t^{(k)}\}_{k=1}^{K}$, b_t, $p(a \mid b_t)$, $V(b_t)$

function $\textsc{Leapfrog}(\psi, p;\, \tilde{h})$
  $p \leftarrow p - \frac{\epsilon}{2} \nabla_\psi U_\theta(\psi;\, \tilde{h})$ ▷ initial half-kick
  for $\ell = 1, \ldots, L-1$ **do**
    $\psi \leftarrow \psi + \epsilon\, p$
    $p \leftarrow p - \epsilon \nabla_\psi U_\theta(\psi;\, \tilde{h})$ ▷ drift, kick
  end for
  $\psi \leftarrow \psi + \epsilon\, p$ ▷ final drift
  $p \leftarrow p - \frac{\epsilon}{2} \nabla_\psi U_\theta(\psi;\, \tilde{h})$ ▷ half-kick
  return ψ
end function

C GBVI Algorithm

Algorithm 1 summarizes the per-timestep belief update of GBVI: the components of Section 3.1 are transported by the Hamiltonian flow of Section 3.2 and reweighted by the gated recurrent memory of Section 3.3.

D BACKGROUND

D.1 PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

A partially observable Markov decision process (POMDP) (Åström, 1965) is formalized as a tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \Omega, O, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $T(s' \mid s, a)$ is the transition function giving the probability of reaching s' from s via action a, $R(s, a)$ is the reward function, Ω is the observation space, $O(o \mid s', a)$ is the observation function giving the probability of observing o after reaching s' via action a, and $\gamma \in [0, 1)$ is the discount factor. At each discrete timestep $t \in \{0, 1, 2, \ldots\}$, the environment occupies a latent state $s_t \in \mathcal{S}$, the agent selects an action $a_t \in \mathcal{A}$, and the environment transitions to $s_{t+1} \sim T(\cdot \mid s_t, a_t)$ and emits an observation $o_{t+1} \sim O(\cdot \mid s_{t+1}, a_t)$. Unlike a Markov decision process, the agent does not directly observe s_t.

Because the state is hidden, the agent reasons about it through the history of past actions and observations up to timestep t, denoted $h_t = (a_0, o_1, a_1, o_2, \ldots, a_{t-1}, o_t)$. The sufficient statistic for this history is the belief $b_t(s) = p(s_t = s \mid h_t)$, the posterior distribution over the latent state at timestep t conditioned on the history (Kaelbling et al., 1998). After taking action a_t and observing o_{t+1}, the belief evolves via Bayes' rule:

$$b_{t+1}(s') = \eta \, O(o_{t+1} \mid s', a_t) \sum_{s \in \mathcal{S}} T(s' \mid s, a_t) \, b_t(s), \tag{7}$$

where η is a normalization constant ensuring $\sum_{s' \in \mathcal{S}} b_{t+1}(s') = 1$. The sum is the prediction step and the product with O the correction step; Proposition 1 writes this update on a set of weighted components. The optimal policy in a POMDP can be expressed as a mapping from beliefs to actions (Kaelbling et al., 1998), reducing sequential decision-making to maintaining an accurate belief representation along the trajectory.

The shape of b_t depends on how informative h_t is about s_t: when multiple latent states are consistent with h_t, b_t places mass on each of them and takes on a complex, potentially multi-modal shape. Representing such belief shapes tractably along the trajectory is the central challenge addressed in this work.

D.2 WASSERSTEIN–FISHER–RAO GEOMETRY

The Bayes update of Eq. (7) changes the belief in two ways: the prediction step moves probability mass through the transition, and the correction step scales it in place by the likelihood of the observation. The Wasserstein–Fisher–Rao (WFR) geometry (Chizat et al., 2010; Liero et al., 2017) makes this distinction formal for any change of a distribution. A path ρ_t from one distribution to another is driven by a velocity field v_t, which transports mass, and a growth rate g_t, which creates or removes it where it is, through the continuity equation with a source term, and the squared WFR distance is the least cost of such a path,

$$\mathrm{WFR}^2_\delta(\rho_0, \rho_1) = \inf_{(\rho_t, v_t, g_t)} \int_0^1 \int \left(\|v_t\|^2 + \delta^2 g_t^2 \right) d\rho_t \, dt \quad \text{s.t.} \quad \partial_t \rho_t + \nabla \cdot (\rho_t v_t) = \rho_t g_t. \tag{8}$$

The first term is the Wasserstein (transport) cost and the second the Fisher–Rao (reweighting) cost; δ sets the price of creating a unit of mass relative to moving it, and beyond a distance of $\pi\delta$ it is cheaper to remove mass and re-create it than to move it, so $\pi\delta$ acts as a cutoff on how far transport reaches. Section 2 uses this decomposition to name the two requirements on a belief representation, Proposition 1 writes the Bayes update in its terms, and Appendix G.3 uses the distance, with its split into a transport and a reweighting part, as the belief-error metric on Area Search.

Neither operation can be computed directly on a continuous distribution, so work in this geometry represents the distribution by a finite set of weighted components, $\rho = \sum_{k=1}^{K} \pi_k \, q(\cdot \mid \psi_k)$, on which the two operations become concrete: transport moves the component parameters ψ_k and reweighting changes the weights π_k. Given a target functional, such as the negative log-likelihood of data (Yan et al., 2023) or the KL divergence to a known posterior (Lambert et al., 2022), its WFR gradient flow prescribes both updates, and the Fisher–Rao part is multiplicative, so a component with zero weight never regains mass and transport alone is slow to reach a distant region (Yan et al., 2023). A POMDP with unknown transition and observation models offers no such functional; Section 3 therefore keeps the representation, a set of components that are moved and reweighted, but learns the two updates from observations

D.3 Semi-Implicit Variational Inference

Variational inference (VI) (Kingma & Welling, 2022) approximates an intractable posterior $p(z \mid o)$ — the distribution over a latent variable z given an observation o — with a tractable distribution $q_\phi(z)$ parameterized by ϕ, learned by maximizing the evidence lower bound (ELBO)

$$\mathcal{L}(\phi) = \mathbb{E}_{z\sim q_\phi(z)}[\log p(o, z) - \log q_\phi(z)]. \tag{9}$$

The $\log q_\phi(z)$ term requires q_ϕ to admit a closed-form density, restricting the variational family to analytic distributions such as Gaussians, Gaussian mixtures, and normalizing flows (Rezende & Mohamed, 2015).

Semi-implicit variational inference (SIVI) (SIVI; Yin & Zhou, 2018) lifts this restriction through a hierarchical construction that defines the variational distribution as the marginal of a generative process rather than a member of a parametric family. Rather than parameterizing $q_\phi(z)$ directly, SIVI introduces an auxiliary variable ψ, which plays the role of a stochastic parameter for an analytic conditional distribution $q(z \mid \psi)$. The variable ψ is itself drawn from a mixing distribution $q_\phi(\psi)$, giving a two-stage hierarchy

$$z \sim q(z \mid \psi), \qquad \psi \sim q_\phi(\psi). \tag{10}$$

Marginalizing ψ gives the variational distribution

$$h_\phi(z) = \int q(z \mid \psi)\, q_\phi(\psi)\, d\psi. \tag{11}$$

The conditional $q(z \mid \psi)$ is required to be explicit, but the mixing distribution $q_\phi(\psi)$ can be implicit — for example, the output of a neural network transforming random noise. SIVI's surrogate ELBO requires only the density of $q(z \mid \psi)$, not of the marginal $h_\phi(z)$, so $h_\phi(z)$ inherits the flexibility of the implicit $q_\phi(\psi)$ while remaining tractable to optimize, capturing skewness, multimodality, and other features beyond analytic families. The variational distribution is thus determined by the generative process that produces it, not by a closed-form density family chosen in advance.

Although SIVI lifts the closed-form density constraint, applying its generative-process construction to POMDPs requires adaptation. SIVI is formulated for approximating a single posterior $p(z \mid o)$ over a latent variable z given an observation o, with no notion of time. In the POMDP setting of Section D.1, the relevant posterior is instead a sequence of distributions over a latent z_t, one per timestep t, with the conditioning input being the action-observation history h_t rather than a single observation. Adapting SIVI to this setting requires conditioning the hierarchy on the history at each timestep, giving $q(z_t \mid \psi_t, h_t)$ and $q_\phi(\psi_t \mid h_t)$, where ψ_t denotes the auxiliary variable at timestep t. Since h_t grows with t while a neural network conditioning input must have fixed dimensionality, this in turn requires a mechanism to compress the variable-length history into a fixed-dimensional summary, which we discuss next.

D.4 Recurrent Encoding and Hamiltonian Dynamics

Recurrent neural networks compress variable-length sequences into fixed-dimensional representations, making them a natural fit for conditioning the SIVI hierarchy on the history h_t. A recurrent network maintains a fixed-dimensional hidden state $\tilde{h}_t$ that is updated incrementally as new actions and observations arrive,

$$\tilde{h}_t = f_\theta(\tilde{h}_{t-1}, a_{t-1}, o_t), \tag{12}$$

where f_θ is a parameterized recurrent update function and $\tilde{h}_0$ is a learned initial state. Common instantiations of f_θ include LSTMs (Hochreiter & Schmidhuber, 1997), GRUs (Cho et al., 2014), linear recurrent units (Orvieto et al., 2023), and the gated linear recurrent units used in Griffin (De et al., 2024), each trading off memory capacity, training stability, and computational cost.

Gated linear recurrences. The linear recurrent unit (LRU) (Orvieto et al., 2023) updates its state by $\tilde{h}_t = \alpha \odot \tilde{h}_{t-1} + \sqrt{1-\alpha^2} \odot \hat{x}_t$ with a fixed diagonal decay α and an input $\hat{x}_t$; the decay is the same on every step, so the state cannot decide per step whether to keep or overwrite what it holds. The RG-LRU (De et al., 2024) makes the decay and the input scale depend on the input through two gates,

$$r_t = \mathrm{sig}(W_r \hat{x}_t + \beta_r), \qquad i_t = \mathrm{sig}(W_i \hat{x}_t + \beta_i), \qquad \alpha_t = \exp(-c \cdot \mathrm{softplus}(\Lambda) \odot r_t)\,, \tag{13}$$

with sigmoid sig, element-wise product $\odot$, learnable $W_r, W_i, \beta_r, \beta_i$, a learnable vector Λ with one entry per hidden unit, and a scalar c. The recurrence gate r_t sets the decay α_t: near zero it makes α_t close to one and the state is held, near one it lets the input overwrite the state at the rate set by Λ. The input gate i_t scales the input before it enters. Section 3.3 uses the recurrence of Eq. (5) with these gates, and Appendix G.1 gives their initialization.

Hamiltonian dynamics. Beyond compressing the history, sequential belief approximation also requires a mechanism for evolving the components of the generative process across timesteps while keeping them distinguishable. Hamiltonian dynamics (Neal, 2012) provides such a mechanism for systems whose underlying dynamics evolve smoothly, covering both linear and nonlinear regimes (Appendix E.1 delineates the scope). Given a position variable $\psi \in \mathbb{R}^{2d}$ — which in our setting plays the role of the SIVI auxiliary variable from Section D.3 — and an auxiliary momentum variable $p \in \mathbb{R}^{2d}$, a Hamiltonian system is defined by the energy function

$$H(\psi, p) = U(\psi) + \tfrac{1}{2}\|p\|^2, \tag{14}$$

where $U(\psi)$ is the potential energy and $\frac{1}{2}\|p\|^2$ is the kinetic energy under unit mass. Hamilton's equations $\dot{\psi} = p, \dot{p} = -\nabla_\psi U(\psi)$ move (ψ, p) along trajectories of constant energy. In practice they are integrated by the Leapfrog scheme (Neal, 2012; Leimkuhler & Reich, 2005), which advances one step of size ϵ by a half-step of momentum, a full step of position, and a second half-step of momentum,

$$\begin{aligned} p_{1/2} &= p - \tfrac{\epsilon}{2}\nabla_\psi U(\psi), \\ \psi' &= \psi + \epsilon\, p_{1/2}, \\ p' &= p_{1/2} - \tfrac{\epsilon}{2}\nabla_\psi U(\psi'), \end{aligned} \tag{15}$$

and is applied L times per timestep. The map $(\psi, p) \mapsto (\psi', p')$ is symplectic: it preserves phase-space volume exactly, for any ϵ and L, because each of its three sub-steps is a shear. This is the property Section 3.2 relies on: two components at distinct positions are never mapped to the same position, so a population of components stays distinguishable across arbitrarily long trajectories. Section 3.2 conditions the potential on the memory state, $U_\theta(\psi; \tilde{h}_t)$, and draws a fresh momentum per component at each timestep.

E Theoretical Analysis

E.1 Assumptions

Before stating our assumptions, we briefly delineate the class of dynamical systems to which Hamiltonian flow applies, since this scope underlies several of the assumptions below. Hamiltonian dynamics describes the evolution of a system through a smooth energy function on phase space, and applies to systems whose underlying state evolves continuously in time (Neal, 2012; Leimkuhler & Reich, 2005). This regime covers both linear dynamics, such as harmonic oscillators, and nonlinear dynamics, such as n-body gravitational systems and molecular dynamics, provided the energy landscape is differentiable and trajectories do not exhibit jumps. Symplectic integrators such as Leapfrog discretize these continuous trajectories while preserving phase-space volume, which keeps a population of representative components distinguishable across arbitrarily long horizons (Leimkuhler & Reich, 2005). Two regimes fall outside this scope: *hybrid or discontinuous dynamics*, in which the latent state undergoes abrupt jumps (e.g., contact events in rigid-body systems, mode switches in switched systems), and *dissipative or stochastic-dominated dynamics*, in which energy is not approximately conserved over the relevant horizon. GBVI's formulation targets the smooth regime, which covers the Atari occlusion POMDPs studied in this paper, where the underlying game state evolves continuously between frames. We now state the assumptions under which our analysis proceeds.

Assumption 4 (Explicit Conditional Distribution)**.** The per-component conditional distribution $q(z_t \mid \psi_t)$ admits a closed-form density that is differentiable in ψ_t. In our instantiation, $q(z_t \mid \psi_t) = \mathcal{N}\big(z_t \mid \mu_t, \operatorname{diag}(\sigma_t)^2\big)$ with $\sigma_t > 0$ elementwise, where $\psi_t = (\mu_t, \log \sigma_t)$.

Remark 5. This assumption is the only closed-form density requirement that GBVI imposes on its variational family, and it is intentionally placed on the conditional rather than the marginal. The SIVI surrogate ELBO (Yin & Zhou, 2018) evaluates densities only through $q(z_t \mid \psi_t)$ and never

through the marginal $h_\phi(z_t \mid \tilde{h}_t)$, so the explicit-density requirement is fully discharged at the conditional level. This stands in contrast to existing variational POMDP methods—including diagonal-Gaussian approaches (Krishnan et al., 2015; Han et al., 2019; Hafner et al., 2020), Gaussian-mixture variants (Tschiatschek et al., 2018), and normalizing flows (Chen et al., 2022)—which require the entire belief distribution to admit a tractable density. By restricting closed-form structure to the conditional, GBVI lifts the density restriction at the level of the per-step belief while retaining the tractability needed for optimization.

Assumption 6 (Sampleable Mixing Distribution)**.** The history-conditioned mixing distribution $q_\phi(\psi_t \mid \tilde{h}_t)$ is sampleable: for any $\tilde{h}_t$, the generative process in Eq. (2) produces samples ψ_t in finite computation. We do not require q_ϕ to admit a closed-form density.

Remark 7. This assumption is what lets GBVI's per-step belief $h_\phi(z_t \mid \tilde{h}_t)$ take shapes beyond closed-form variational families. The SIVI surrogate ELBO needs only samples from q_ϕ, not its density, so the transport of Section 3 can be any generative process that carries the components across steps: the memory-conditioned Hamiltonian flow of our main instantiation, the Langevin SDE variant, or the Langevin chain of Appendix F.3. The comparable returns across the three flows (Figure 6) reflect this breadth: any flow that moves the components along the memory-conditioned potential without letting two collapse suffices, whether or not its marginal has a closed form. This is the relaxation that separates GBVI's variational family from those of prior POMDP belief methods, which fix a closed-form family before training (Section 2).

Assumption 8 (Smooth Energy Landscape)**.** The memory-conditioned potential $U_\theta(\psi;\, \tilde{h})$ is continuously differentiable in ψ for every $\tilde{h}$, and its gradient is L_U-Lipschitz in ψ uniformly over $\tilde{h}$: there exists $L_U > 0$ such that for all $\psi_1, \psi_2 \in \mathbb{R}^{2d}$ and all $\tilde{h}$,

$$\left\|\nabla_\psi U_\theta(\psi_1;\, \tilde{h}) - \nabla_\psi U_\theta(\psi_2;\, \tilde{h})\right\| \;\le\; L_U \left\|\psi_1 - \psi_2\right\|. \tag{16}$$

Remark 9. The C^1 regularity guarantees that the Leapfrog map (half-kick / drift / half-kick) is a diffeomorphism on phase space, which directly supports the distinguishability claim in contribution (ii): distinct initial component configurations remain distinct under integration for arbitrary trajectory lengths, since a diffeomorphism cannot map two distinct points to the same image. The Lipschitz condition further ensures stability of the Leapfrog update under our chosen step size. In our instantiation, U_θ is a 3-layer MLP with SiLU activations (Table 6); SiLU is selected over ReLU precisely because it yields a C^1 potential, satisfying this assumption by construction. Architectures with non-differentiable activations (e.g., ReLU) would produce a potential whose gradient is discontinuous at activation kinks, technically violating this assumption—though Leapfrog typically remains numerically stable in practice.

Assumption 10 (Stable Integrator Step Size)**.** The Leapfrog step size ϵ and the Lipschitz constant L_U from Assumption 8 satisfy $\epsilon L_U < 2$.

Remark 11. This is the standard linear-stability condition for symplectic Leapfrog integration on a smooth potential (Neal, 2012; Leimkuhler & Reich, 2005): under it the integrator's update map has bounded operator norm, so component trajectories do not blow up over the L Leapfrog steps applied per environment timestep. The condition is mild given GBVI's hyperparameters ($\epsilon = 0.1$, $L = 2$ Leapfrog steps; see Table 6), and is consistent with Figure 6, where the Hamiltonian flow has the narrowest standard-error bands of the three flows. This assumption is a stability guarantee, not a distinguishability guarantee: the latter follows from Assumption 8 alone, so violating $\epsilon L_U < 2$ would degrade trajectory accuracy without compromising the property Section 3.2 relies on, that two components are never mapped to the same position.

Assumption 12 (Smooth Latent Dynamics)**.** The environment's underlying state evolution, projected to the latent space $\mathcal{Z}$, follows trajectories that are continuous in time without abrupt jumps. Formally, there exists $L_{\mathrm{dyn}} > 0$ such that $\|z_{t+1} - z_t\| \le L_{\mathrm{dyn}}$ for any consecutive latent states along an episode trajectory.

Remark 13. This assumption places GBVI within the smooth-dynamics regime described in the preamble and excludes hybrid systems with contact events, mode switches, or other abrupt latent transitions—which would require event-driven integrators rather than smooth Hamiltonian flow. The Atari occlusion POMDPs studied in this paper satisfy this assumption: game state evolves continuously between frames under the underlying ALE dynamics, and the curtain occlusion mechanism modulates only the *observation* of state, not the state itself. Domains with discontinuous latent transitions—rigid-body contact, switched control systems, abrupt regime changes in financial time series—fall outside GBVI's current scope and are discussed in Appendix A.

E.2 PROOFS

Proof of Proposition 1. Write the belief as $b_t = \sum_{k=1}^{K} \pi_k \, q_k$ with $q_k = q(\cdot \mid \psi_t^{(k)})$, and let T and O be the transition and observation functions of Eq. (7) for the action a_t. The prediction step is linear in the belief,

$$(Tb_t)(s') = \sum_s T(s' \mid s, a_t)\, b_t(s) = \sum_k \pi_k \, (Tq_k)(s'),$$

so it acts on each component separately and leaves the weights unchanged. This is the transport, and $\bar{q}_k = Tq_k$ is the predicted component. The correction step multiplies by $O(o_{t+1} \mid s', a_t)$ and renormalizes. Let $\ell_k = \sum_{s'} O(o_{t+1} \mid s', a_t)\, \bar{q}_k(s')$ be the evidence of o_{t+1} under $\bar{q}_k$. Then

$$b_{t+1}(s') = \frac{\sum_k \pi_k \, O(o_{t+1} \mid s', a_t)\, \bar{q}_k(s')}{\sum_j \pi_j \, \ell_j} = \sum_k \frac{\pi_k \, \ell_k}{\sum_j \pi_j \, \ell_j} \, \tilde{q}_k(s'), \qquad \tilde{q}_k = \frac{O(o_{t+1} \mid \cdot, a_t)\, \bar{q}_k}{\ell_k},$$

where each $\tilde{q}_k$ is a normalized distribution. Multiplying by O and renormalizing changes the mass at each state without moving it, so together the new weights $\pi'_k = \pi_k \ell_k / \sum_j \pi_j \ell_j$ and the reshaped components $\tilde{q}_k$ are the reweighting. If o_{t+1} carries no information about the state, $O(o_{t+1} \mid s', a_t)$ is the same constant c for every s', so $\ell_k = c$ for every k, $\pi' = \pi$, and $\tilde{q}_k = \bar{q}_k$. The decomposition is exact. GBVI keeps each component in the Gaussian family of Eq. (1): the flow's move of $\psi^{(k)} = (\mu, \log \sigma)$ realizes both $\bar{q}_k$ and $\tilde{q}_k$ by a projection onto this family, and the memory realizes the weight update. The projection is the approximation. □

Proof of Corollary 2. Fix a period of n steps without informative observations. A component is dropped when the representation retains no mass within the cutoff of it, and we assume that a dropped component is not re-created during the period. Let D_n be the set of components dropped by the end of the period. Dropping is irreversible, so $D_n \subseteq D_{n+1}$. By Proposition 1, the reweighting is the identity on every step of the period, and the prediction preserves the mass of each component. The mass of the exact posterior on the hypotheses of D_n at the end of the period is therefore $m_n = \sum_{k \in D_n} \pi_k$, the mass those hypotheses held when they were dropped, and m_n is nondecreasing in n. On the step before the next informative observation arrives, the representation's belief places no mass within the cutoff of the region where the exact posterior holds m_n, since every dropped hypothesis is farther than the cutoff from every retained component (on Area Search, on a different road). In the Wasserstein–Fisher–Rao distance with cutoff, mass that has no source within the cutoff cannot be transported and has to be created, at a cost of at least c per unit of mass, where c is the reweighting cost of one unit of mass (Appendix G.3). Hence $\mathrm{WFR}^2 \geq c\, m_n$, and the bound is nondecreasing in n. □

Lemma 14 (Bounded contraction under Leapfrog)**.** *Under Assumptions 8 and 10, with $\epsilon \leq 1$, one Leapfrog step at a fixed memory state and a common momentum p maps two components ψ_1, ψ_2 to ψ'_1, ψ'_2 with*

$$\|\psi'_1 - \psi'_2\| \;\geq\; \left(1 - \tfrac{\epsilon^2 L_U}{2}\right) \|\psi_1 - \psi_2\| \;>\; 0.$$

Proof. The position update of Eq. (15) is $\psi' = \psi + \epsilon p - \frac{\epsilon^2}{2} \nabla_\psi U_\theta(\psi; \tilde{h})$. For a common p the momentum term cancels in the difference, and by the triangle inequality and the L_U-Lipschitz gradient,

$$\|\psi'_1 - \psi'_2\| \;\geq\; \|\psi_1 - \psi_2\| - \tfrac{\epsilon^2}{2} \|\nabla_\psi U_\theta(\psi_1; \tilde{h}) - \nabla_\psi U_\theta(\psi_2; \tilde{h})\| \;\geq\; \left(1 - \tfrac{\epsilon^2 L_U}{2}\right) \|\psi_1 - \psi_2\|.$$

With $\epsilon \leq 1$, Assumption 10 gives $\epsilon^2 L_U \leq \epsilon L_U < 2$, so the factor is positive. After the first step the two components carry different momenta, so the bound above applies to a single step. Across steps, distinctness follows from two facts. For any fixed momenta the Leapfrog map is symplectic and hence a diffeomorphism of phase space (Remark 9), so two components with distinct (ψ, p) are mapped to distinct (ψ', p'). With independent momenta $p_1 \neq p_2$ the position difference gains the term $\epsilon(p_1 - p_2)$, whose distribution is absolutely continuous, so $\psi'_1 = \psi'_2$ has probability zero. This is the volume-preservation statement used in Section 3.2. □

Proof of Theorem 3. Condition on ψ_{t-1}, $\tilde{h}_{t-1}$ and a_{t-1} throughout; all densities below are conditional on these, and we suppress the conditioning. Write $P = (p_t^{(1)}, \ldots, p_t^{(K)})$ for the momenta.

Given P and o_t, every component $\psi_t^{(k)} = \Phi(\psi_{t-1}^{(k)}, p_t^{(k)}; U_\theta(\,\cdot\,; \tilde{h}_t))$ is a deterministic function of P, since $\tilde{h}_t$ is a deterministic function of $\tilde{h}_{t-1}$, o_t and a_{t-1}. The belief is then the explicit mixture $m_P(z_t) = \sum_k \pi_k(\tilde{h}_t)\, q(z_t \mid \psi_t^{(k)})$, and the belief of Eq. (3) is its expectation, $h_\phi = \mathbb{E}_P[m_P]$. Drawing $k \sim \pi(\tilde{h}_t)$ and $z_t \sim q(\,\cdot \mid \psi_t^{(k)})$ draws $z_t \sim m_P$, so the objective of Eq. (6) is

$$\mathcal{L}^{(t)} = \mathbb{E}_P\, \mathbb{E}_{z_t \sim m_P}[\log p(o_t, z_t) - \log m_P(z_t)] = \mathbb{E}_P\big[\mathrm{ELBO}(m_P)\big].$$

For every P, $\mathrm{ELBO}(m_P) = \log p(o_t) - \mathrm{KL}(m_P \,\|\, p(\,\cdot \mid o_t)) \le \log p(o_t)$, which gives $\mathcal{L}^{(t)} \le \log p(o_t)$. Moreover, since the KL divergence is convex in its first argument and $h_\phi = \mathbb{E}_P[m_P]$, Jensen's inequality gives $\mathrm{KL}(h_\phi \,\|\, p(\,\cdot \mid o_t)) \le \mathbb{E}_P[\mathrm{KL}(m_P \,\|\, p(\,\cdot \mid o_t))]$, that is, $\mathcal{L}^{(t)} \le \mathrm{ELBO}(h_\phi) \le \log p(o_t)$. No step used the distribution of P or independence between the $\psi_t^{(k)}$ or across timesteps, so the bound holds for any distribution of the momenta and for components carried from their previous positions. □

F Additional Results

This section reports five additional experiments that complement the main results in Section 4. Each one isolates one part of the decomposition of Section 3. A synthetic crossing-modes task turns off transport and reweighting one at a time (Section F.1). An extended Bowling curtain-width sweep adds the transport-only variant to the recurrent baselines (Section F.2). A comparison of three instantiations of the flow shows that the transport does not depend on the Hamiltonian choice (Section F.3). A sweep over the number of components K traces how many hypotheses the belief needs to hold (Section F.4). A controlled-gap study on Area Search measures belief error against the number of steps the vehicle is out of view (Section F.5).

F.1 Transport and Reweighting on Synthetic Crossing Modes

The decomposition of Section 3 says that a belief needs two operations: transport, which carries the components while observations are uninformative, and reweighting, which moves weight between them when an informative observation arrives. This experiment turns each one off in turn on a task where the exact belief is known. Two Gaussian modes ($\sigma = 0.3$) start at $(\pm 3, 0)$, move toward each other until they overlap at $t = 50$, and continue until they have swapped positions at $t = 70$. Only one mode is observed per timestep, alternating, so the position of the other has to be carried by the belief. We initialize each method's components at the true positions and measure $\mathrm{KL}(Q \| P)$ at every step; the KL is available here because P is an explicit mixture. The crossing is the hard part: while the two modes overlap, the two hypotheses cannot be told apart by the observations, and a belief that merges them cannot separate them again when the modes move on.

The two variants use analytical updates, so the comparison is about the dynamics and not about learning. The *reweighting-only* variant keeps the memory-driven weighting but replaces the Hamiltonian flow with a contractive update that pulls each component toward the position the memory currently favours, as a recurrent state does. The *transport-only* variant keeps the volume-preserving flow but fixes the weights, conditioning on the equally weighted history instead of on recent observations, as a particle representation without a learned weight update does.

Each variant fails on the operation it lacks. Figure 4 shows the two failures. The reweighting-only variant tracks both modes before the crossing, but from $t \approx 30$ the contractive update pulls its two components together, and once they have collapsed the hypothesis that reverses direction is gone. Its error rises again after the crossing, when the modes separate and the belief has only one component to follow them with. This is the situation of Corollary 2: a hypothesis dropped during the gap cannot be recovered when the observation that supports it arrives. The transport-only variant keeps its components apart through the crossing, but its weights cannot follow the reversal, so the stale pre-crossing evidence keeps pulling the belief toward the wrong mode and its error peaks at $t \approx 70$. GBVI combines both: the flow carries the two hypotheses through the overlap without letting them collapse, and the memory moves weight onto the one the next observation supports. Neither operation suffices alone. Reweighting without volume preservation loses the hypothesis; transport without reweighting cannot correct the belief when the observation arrives.

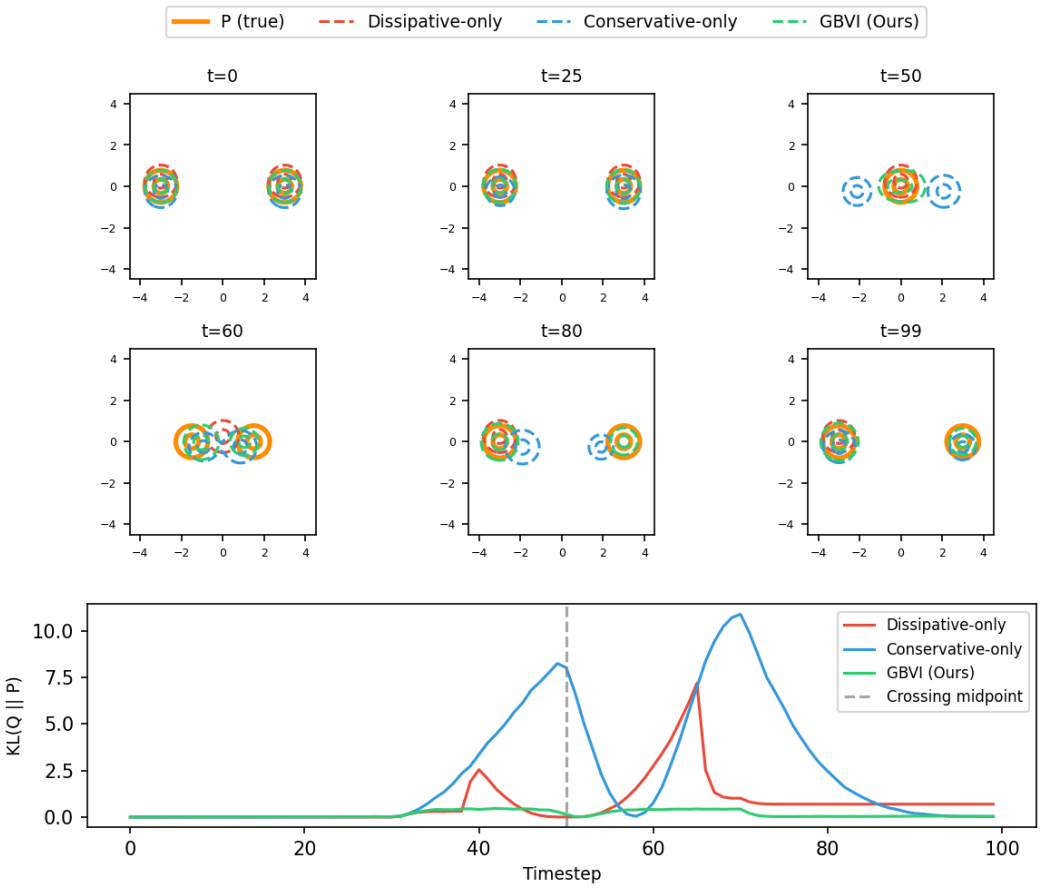


Figure 4: Crossing modes. Top: the true distribution P (solid) and the belief Q of each method (dashed) at key timesteps. Bottom: $\mathrm{KL}(Q\|P)$ over time. The two modes overlap at $t = 50$ and have swapped positions by $t = 70$. The reweighting-only variant collapses its components before the crossing and cannot recover the hypothesis that reverses. The transport-only variant keeps its components apart but cannot move weight onto the reversing one. GBVI keeps the error near zero throughout.

F.2 Bowling Sweep: Recurrent Baselines and Transport-only Variant

The curtain width on Bowling sets how long the ball stays hidden, so sweeping it lengthens the gap while everything else stays fixed. The curtain spans the full frame height, and widths of $\{12, 24, 36, 48, 60\}$ pixels cover progressively more of the lane and of the ball's trajectory; width 24 is the configuration of Table 1. We compare GBVI against the recurrent encoders RG-LRU and LRU and against the transport-only variant, which keeps the Hamiltonian flow but fixes the weights. Each method is trained for 5M environment steps and evaluated over 100 episodes per width.

The widths span short and long gaps. At width 12 the curtain hides only a small part of the lane, the ball reappears after a few steps, and any history summary suffices. At width 60 the curtain hides nearly the whole trajectory, so the agent sees the release and then brief glimpses of the pins. Scoring at this width requires the belief to carry hypotheses about the ball's position through a long stretch of uninformative frames and to correct them in a single update when the ball reappears. The intermediate widths interpolate between these two regimes.

GBVI's return falls the least as the gap grows. Figure 5 shows that all four methods score about 35 at width 12 and that GBVI has the highest return at every larger width. At width 60 GBVI scores 28.9 ± 0.45, while RG-LRU and LRU fall to 22.3 ± 0.87 and 21.6 ± 0.79, a gap of about 7 points. GBVI loses about 18% of its return from width 12 to width 60, RG-LRU about a third.

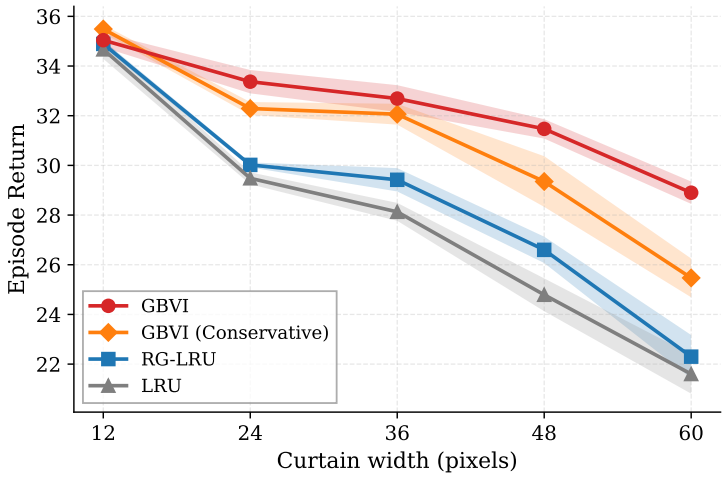


Figure 5: Episode return on Bowling under the fixed curtain as the curtain width grows from 12 to 60 pixels, for GBVI, its transport-only variant, RG-LRU and LRU. Solid lines show the mean over 100 episodes and shaded bands the standard error. All four methods are close at the smallest width, where the ball is hidden for only a few steps. As the width grows and the ball stays hidden longer, GBVI's return falls the least, the transport-only variant sits between GBVI and the recurrent encoders, and the recurrent encoders fall the most.

The gap widens as the ball stays hidden longer. A recurrent vector holds one hypothesis, so once the ball's position behind the curtain is ambiguous the vector commits to one and cannot correct it when the ball reappears elsewhere; GBVI carries several positions and moves weight onto the one the reappearance supports.

Transport alone carries part of the advantage, and reweighting the rest. The transport-only variant sits between GBVI and the recurrent encoders across the whole sweep. The volume-preserving flow keeps the K components from collapsing while the ball is hidden, which is why the variant beats RG-LRU and LRU as the curtain widens. But with fixed weights the belief cannot shift mass onto the component the reappearance supports, so the policy receives the hypotheses without knowing which one the observation favours. The distance between the transport-only variant and GBVI is the contribution of the reweighting: about 3–4 points of return at width 60. GBVI's standard error is also the narrowest across the sweep, ± 0.45 at width 60 against ± 0.87 and ± 0.79 for the recurrent encoders.

F.3 Bowling Sweep: Choice of the Transport Flow

Section 3 realizes the transport with a Hamiltonian flow on a memory-conditioned potential, integrated by Leapfrog over K persistent components. The transport itself does not require this choice: any map that carries the components along the potential without letting two collapse serves it. To test how much depends on the Hamiltonian flow, we replace it with two alternatives that keep the same potential and the same reweighting and differ only in how they move the components. The alternatives go from more to less explicit structure on the components, so the comparison also places GBVI on the spectrum of implicit inference.

Three flows with decreasing structure. The first is the *Hamiltonian Leapfrog flow over K persistent components* of Section 3: each component is pushed by a fresh momentum and bent by the gradient of U_θ, and the symplectic integrator preserves volume, so two components are never mapped to the same position. The second is an *overdamped Langevin SDE over K persistent components*,

$$\psi_t^{(k)} = \psi_{t-1}^{(k)} - \epsilon \nabla_\psi U_\theta(\psi_{t-1}^{(k)}; \tilde{h}_t) + \sqrt{2\epsilon}\, \xi_t^{(k)}, \quad \xi_t^{(k)} \sim \mathcal{N}(0, I_{2d}),$$

which keeps the K components and the same potential but gives up the volume-preservation guarantee. The third is a *Langevin chain* that gives up the K persistent components too: the components at each step are draws from a single Langevin chain on the memory-conditioned potential, advanced

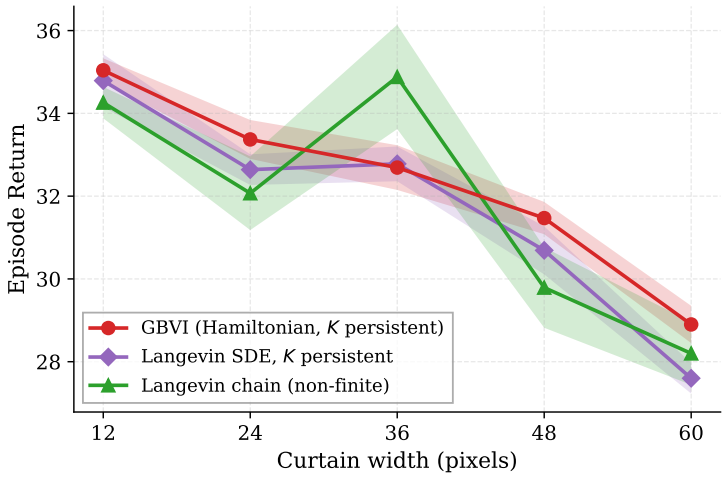


Figure 6: Three flows for the transport on Bowling under the fixed curtain at varying widths. All three share the encoder, the RG-LRU memory, the memory-conditioned potential and the reweighting, and differ only in how the components are moved. Solid lines show mean return over 100 episodes; shaded bands show the standard error. The three returns overlap throughout the sweep, and the Hamiltonian flow has the narrowest bands.

by the same overdamped update. The Hamiltonian flow imposes component identity and symplectic structure, the Langevin SDE keeps identity but drops symplecticity, and the chain keeps neither. All three leave the conditional $q(z_t \mid \psi_t)$ unchanged, so the objective of Section 3 applies to each.

All three reach comparable returns. Figure 6 shows overlapping mean returns at every curtain width, with standard-error bands that overlap throughout. This is consistent with the decomposition: what the transport has to do is carry the components through the gap without letting two collapse, and none of the three flows collapses them. The Hamiltonian flow prevents collapse by preserving volume. The two Langevin variants prevent it by the injected noise $\sqrt{2\epsilon}\,\xi$, which keeps the components from being pulled together by the gradient alone. The flow that does collapse components is a deterministic contraction toward what the memory favours, as in the reweighting-only variant of Section F.1 or in a recurrent vector, and that is the failure the transport is built to avoid. The choice among the three flows is therefore a design dimension on which structure can be traded for implementation simplicity, not a condition for the transport to work.

The Hamiltonian flow is the most reproducible. The three flows differ in the width of their bands. The Hamiltonian flow's bands are noticeably narrower than those of the two Langevin variants, most of all at the larger curtain widths where the gap is longest. Volume preservation guarantees that distinct components stay distinct under integration, so the belief has a similar shape across evaluation seeds. The Langevin variants have no such guarantee, and in long gaps their components occasionally drift toward one another, which raises the seed-to-seed variance even when the mean return is unchanged. Leapfrog is kept in GBVI for this stability, not because the transport requires it.

Position on the spectrum of implicit inference. The three flows trace a spectrum that runs through approximate inference more broadly: explicit families with a closed-form density, such as Gaussians, Gaussian mixtures and normalizing flows; representations whose mixing distribution is implicit but whose carrier is a set of K persistent components; and fully implicit samplers in which the mixing distribution is an SDE or MCMC chain (Yin & Zhou, 2018). GBVI sits at the second level by default. The experiment shows that it keeps working when pushed toward the third, and the results of Section F.4 show what is lost when it is pushed back toward the first, at $K = 1$.

F.4 Bowling Sweep: Number of Components

The number of components K is the number of hypotheses the belief can hold at once. With $K{=}1$ the belief of Eq. (3) is a single Gaussian with a closed-form density; as K grows it approaches a

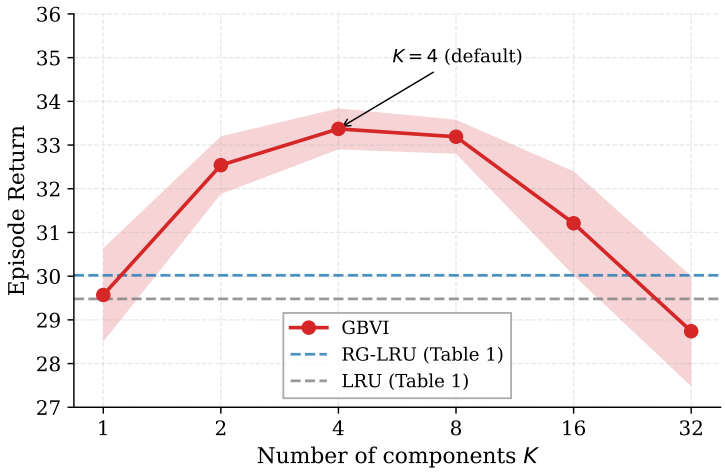


Figure 7: Episode return on Bowling at curtain width 24 as the number of components K varies over $\{1, 2, 4, 8, 16, 32\}$ (5M steps, 100 episodes per K). Solid red line shows mean return; shaded band shows the standard error. Horizontal dashed lines mark the RG-LRU and LRU baselines from Table 1 at the same setting. The curve is an inverted U: $K{=}1$ reduces GBVI to a single Gaussian and matches the recurrent baselines, $K{=}4$–8 is the peak, and $K{=}32$ falls below the baselines.

nonparametric representation. We sweep $K \in \{1, 2, 4, 8, 16, 32\}$ on Bowling under the fixed curtain at width 24, training each configuration for 5M environment steps and evaluating over 100 episodes. All other hyperparameters are held at their values in Table 6, including the latent dimension $d{=}8$, the Leapfrog steps $L{=}2$ and step size $\epsilon{=}0.1$, and the potential network, so the only variable across runs is K.

The sweep has three regimes. Figure 7 shows an inverted U. $K{=}1$ scores 29.57 ± 1.07, $K{=}2$ jumps to 32.54 ± 0.66, the curve peaks at $K{=}4$ with 33.37 ± 0.47 and holds at $K{=}8$ with 33.19 ± 0.39, then falls to 31.21 ± 1.19 at $K{=}16$ and to 28.74 ± 1.26 at $K{=}32$. The baselines of Table 1 at this setting are RG-LRU 30.02 and LRU 29.48. At $K{=}1$ GBVI ties the recurrent baselines. From $K{=}2$ to $K{=}8$ it is clearly above them. From $K{=}16$ it falls back, and at $K{=}32$ below them.

$K{=}1$ is a single closed-form Gaussian and ties the recurrent baselines. At $K{=}1$ the only weight is $\pi_1(\tilde{h}_t) \equiv 1$, so there is nothing to reweight, and the belief is a single Gaussian of the family used by Krishnan et al. (2015) and Han et al. (2019). The component is still transported by the flow, so the belief is carried across steps, but it holds one hypothesis. Its return of 29.57 ± 1.07 sits between RG-LRU (30.02) and LRU (29.48). This is what Section 3 predicts: a belief with one hypothesis cannot hold several positions of the hidden ball, so it gives the policy no more than a recurrent vector does. It also isolates what the semi-implicit construction adds. If holding one hypothesis were not the limitation, letting the belief hold several could not produce the gains at $K \geq 2$.

$K{=}2$ to $K{=}8$ is a plateau, not unbounded growth. The largest step in the sweep is from $K{=}1$ to $K{=}2$, a gain of 2.97 points that puts GBVI above both baselines. $K{=}4$ adds 0.83 over $K{=}2$, and $K{=}8$ is within standard error of $K{=}4$. The jump from one component to two is the gain from holding a second hypothesis. The plateau says that at width 24 the ball is at one of a few plausible positions along the lane, so a few hypotheses cover them, and further components duplicate ones already there. The default $K{=}4$ of the main paper is near the top of this plateau, and $K{=}2$ recovers most of the gain at lower cost.

$K{=}16$ and beyond fall below the baselines. Past the plateau the return declines, and at $K{=}32$ it is below both recurrent baselines. Three mechanisms compound. The weights are a softmax over K outputs of a linear readout from the recurrent state, and as K grows the softmax is harder to move away from near-uniform, so the reweighting cannot concentrate mass on the hypotheses an observation supports. The potential network is shared by all components and has fixed capacity, and as K grows it has to hold more separate low-energy regions than it can, after which components drift toward shared minima. And the weight spread over thirty-two components is too thin for the policy

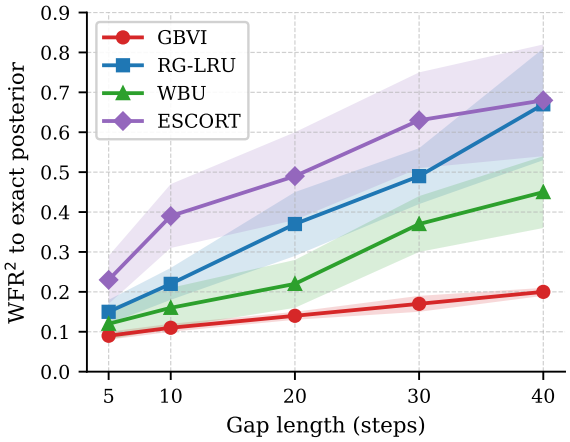


Figure 8: Belief error on Area Search on the last step before the vehicle returns to view, against the number of steps it was out of view. Solid lines show the mean WFR^2 to the exact posterior over 10 scenarios per gap length and shaded bands the standard error. The exact posterior itself spreads along the roads as the gap grows, so a flat curve means the belief spreads with it, not that the vehicle's position is known.

Table 5: Belief error on the last step before the vehicle returns to view, $\mathrm{mean}_{\pm\text{standard error}}$ over 10 scenarios per gap length. Bold marks the lowest error in each column.

Method	$L = 5$	$L = 10$	$L = 20$	$L = 30$	$L = 40$
GBVI	$\mathbf{0.09}_{\pm 0.01}$	$\mathbf{0.11}_{\pm 0.01}$	$\mathbf{0.14}_{\pm 0.01}$	$\mathbf{0.17}_{\pm 0.02}$	$\mathbf{0.20}_{\pm 0.01}$
RG-LRU	$0.15_{\pm 0.03}$	$0.22_{\pm 0.04}$	$0.37_{\pm 0.08}$	$0.49_{\pm 0.07}$	$0.67_{\pm 0.14}$
WBU†	$0.12_{\pm 0.04}$	$0.16_{\pm 0.05}$	$0.22_{\pm 0.06}$	$0.37_{\pm 0.07}$	$0.45_{\pm 0.09}$
ESCORT	$0.23_{\pm 0.06}$	$0.39_{\pm 0.08}$	$0.49_{\pm 0.11}$	$0.63_{\pm 0.12}$	$0.68_{\pm 0.14}$

to read from the belief summary. GBVI's optimum in K is therefore bounded. A particle filter improves as $K \to \infty$ under standard conditions (Arulampalam et al., 2002); GBVI is not a cheaper particle filter but a representation in which a small number of persistent components is enough, and adding components past that number hurts. This is the limitation noted in Appendix A: K has to match the ambiguity of the domain.

The standard error follows the same shape. The standard error is ± 1.07 at $K{=}1$, ± 0.66 at $K{=}2$, ± 0.47 at $K{=}4$, ± 0.39 at $K{=}8$, ± 1.19 at $K{=}16$ and ± 1.26 at $K{=}32$, so the plateau at $K{=}4$–8 is both the highest and the most reproducible setting on Bowling.

F.5 Area Search: Belief Error against Gap Length

Table 2 averages the belief error over every step of an episode, so it cannot say how the error depends on how long the vehicle has been out of view. To measure that dependence directly we design scenarios with a controlled gap. The vehicles' paths are drawn once per scenario and held fixed, and the UAV's route is found by breadth-first search over (cell, step) pairs so that it sees the tracked vehicle for four steps, loses it for exactly L steps, and sees it again on step $L + 1$; scenarios in which the vehicle crosses no intersection during the gap are discarded, since the belief would then have a single hypothesis to carry. We use $L \in \{5, 10, 20, 30, 40\}$ with 10 scenarios each, evaluate every method on the same route and readings within a scenario, and report the WFR^2 on the last gap step, the same step on which Table 2 reports the mass at return. The methods are GBVI and one baseline from each group of Table 1: RG-LRU for a single hypothesis, ESCORT for hypotheses that are carried but can collapse, and WBU for state-supervised training.

Every baseline's error grows with the gap, and GBVI's grows far more slowly. From $L = 5$ to $L = 40$, RG-LRU's error rises from 0.15 to 0.67, ESCORT's from 0.23 to 0.68, and WBU's from 0.12 to 0.45, whereas GBVI's rises from 0.09 to 0.20, one fifth of RG-LRU's slope, with the narrowest standard error at every length. The ordering at $L = 40$ matches the failure modes of Section 2:

RG-LRU holds one hypothesis and so has nothing on the roads the vehicle actually took, ESCORT carries several but its regulariser lets them collapse during long gaps, and WBU's state supervision helps at short gaps but does not keep the hypotheses apart as the gap lengthens. ESCORT's error also saturates between $L = 30$ and $L = 40$, consistent with Corollary 2: once the hypotheses have collapsed, further steps out of view cannot make the belief worse.

The advantage is in carrying the hypotheses, not in the first steps after the vehicle is lost. At $L = 5$ WBU is within 0.03 of GBVI and RG-LRU within 0.06, so all methods start the gap with a comparable belief; the difference opens as the vehicle crosses intersections and the number of roads it could be on multiplies. GBVI's components are transported along those roads, so its belief spreads with the exact posterior, while the baselines' beliefs fall behind it on one part or the other of the split in Table 2. This is the belief-error counterpart of the return curves in Figure 3a, where widening the curtain lengthens the gap and opens the same widening advantage.

G Experiment Details

Table 6: Hyperparameters used in all experiments.

Hyperparameter	**Value**
Observation Encoder	
Input frame size	84×84 grayscale
Frame stack	4
Conv channels	$(32, 64, 32)$
Conv kernel sizes	$(8, 4, 3)$
Conv strides	$(4, 2, 1)$
Activation	ReLU
FC hidden width	512
Latent dimension d	8
RG-LRU Memory	
Input vector $\hat{x}_t$	$[x_t;\, a_{t-1}]$ (one-hot action)
Hidden dimension D	256
Decay initialization range	$[0.9, 0.999]$
Decay scaling factor c	8.0
Gate weight initialization std	$1/\sqrt{D}$
Initial hidden state $\tilde{h}_0$	$\mathbf{0}$
Memory Readout	
Mixture-weight head	Linear $D \to K$, softmax
Energy-conditioning head	Linear $D \to D$
Belief Stream	
Number of mixture components K	4
Component dimension	$2d = 16$
Leapfrog steps L	2
Leapfrog step size ϵ	0.1
Initial component means	$\mathcal{N}(0, I_d)$
Initial component log-stds	$\mathbf{0}$ (unit standard deviation)
Momentum sampling distribution	$\mathcal{N}(0, I_{2d})$ at each timestep
Potential Network U_θ	
Architecture	3-layer MLP
Hidden width	64
Activation	SiLU
Weight initialization	Kaiming-normal
Decoder D_ξ (training only)	
Architecture	3-layer MLP
Hidden width	256

Table 6 – continued

Hyperparameter	**Value**
Activation	SiLU
Stage 1: Belief Pretraining	
Optimizer	Adam
Learning rate	1×10^{-4}
Objective	SIVI surrogate ELBO
Stage 2: PPO Training	
Total environment steps	10^7
Parallel environments	8
Frame skip	4
Rollout length per env (n_{steps})	128
Minibatch size	256
Optimization epochs per rollout	4
Learning rate	2.5×10^{-4} (constant)
Clip range	0.1 (constant)
Discount factor γ	0.99
GAE λ	0.95
Entropy coefficient	0.01
Value coefficient	0.5
Actor / critic head	MLP, 2 hidden layers width 64, tanh
Evaluation episodes per (game, mode)	100

G.1 Implementation Details

Hyperparameter values for the components described in this section are summarized in Table 6; this section specifies the architectures and training procedures behind those values.

Observation encoder. For Atari, observations are pre-processed through the standard Stable-Baselines3 (Raffin et al., 2021) `AtariWrapper` (84×84 grayscale, frame-skip 1, reward clipping) and stacked across 4 consecutive timesteps before encoding. The encoder E_ϕ consists of three 2D convolutional layers with channels $(32, 64, 32)$, kernel sizes $(8, 4, 3)$, and strides $(4, 2, 1)$, each followed by a ReLU activation, then flattened and projected through two fully-connected layers of width 512 and d to produce the encoder feature $x_t \in \mathbb{R}^d$. The CNN backbone follows the convention adopted by DVRL (Igl et al., 2018) and DPFRL (Ma et al., 2020) for fair comparison; all methods in our evaluation share this encoder structure and differ only in how the encoded feature is consumed.

RG-LRU memory. The recurrent memory f_θ instantiates Eq. 5. The memory consumes the input vector $\hat{x}_t = [x_t; a_{t-1}]$ obtained by concatenating the encoder feature with a one-hot encoding of the previous action, and projects it linearly to the recurrence dimension $D = 256$. The recurrence and input gate weights W_r, W_i are initialized from $\mathcal{N}(0, 1/D)$ with zero biases, and the base decay parameter Λ is initialized so that the effective decay α_t at $r_t = 1$ falls uniformly in $[0.9, 0.999]$ following De et al. (2024), with decay scaling factor $c = 8.0$. The hidden state $\tilde{h}_0$ is set to zero at the start of each episode.

Memory readout. Two linear heads read out from $\tilde{h}_t$. The mixture-weight head maps $\tilde{h}_t \in \mathbb{R}^D$ to $\mathbb{R}^K$ followed by a softmax, producing the weight vector $\pi(\tilde{h}_t) \in \Delta^{K-1}$ introduced in Section 3.3. The energy-conditioning head maps $\tilde{h}_t$ to the conditioning input of the potential network. Both heads are jointly trained with the rest of the belief network in the variational pretraining stage described below.

Potential network. The memory-conditioned potential $U_\theta(\psi; \tilde{h}_t)$ is a three-layer MLP of hidden width 64 with SiLU activations, mapping the concatenation $[\psi; \tilde{h}_t]$ to a scalar energy. SiLU is chosen over ReLU because the Hamiltonian flow of Eq. 15 requires the gradient $\nabla_\psi U_\theta$ to behave smoothly. Linear layers are initialized with Kaiming-normal weights and zero biases.

Belief stream. The K persistent components are initialized at the start of each episode with means sampled from $\mathcal{N}(0, I_d)$ and log-standard-deviations set to zero, giving K initially distinct unit Gaussians. At every timestep, fresh momenta $p_t^{(k)} \sim \mathcal{N}(0, I_{2d})$ are sampled independently for each component and discarded after integration. Component positions evolve under Leapfrog integration with $L = 2$ steps and step size $\epsilon = 0.1$ (Eq. 15); the symplectic guarantee that distinct components cannot collapse onto each other holds for any $L \geq 1$, so $L = 2$ is chosen as the minimum that retains a full half-kick / drift / kick / drift / half-kick cycle while keeping the per-step compute overhead small.

Decoder (training only). For variational pretraining we use a decoder D_ξ that maps a latent sample $z \in \mathbb{R}^d$ back to encoder feature space. The decoder is a three-layer MLP with hidden width 256 and SiLU activations, trained against the squared error to the encoder feature x_t as the SIVI reconstruction term. The decoder is discarded after pretraining and is not used during PPO.

Belief summary and policy heads. At each timestep, the K component pairs $\{(\mu_t^{(k)}, \sigma_t^{(k)})\}_{k=1}^K$ together with the mixture weights $\pi(\tilde{h}_t)$ are flattened into the belief summary $b_t \in \mathbb{R}^{K(2d+1)}$ as defined in Section 3.4. The actor π_ζ and critic V_ζ each receive the concatenation $[x_t; b_t]$ and process it through an MLP with two hidden layers of width 64 and tanh activations following the Stable-Baselines3 default. Both heads are jointly optimized by PPO; the belief network is kept frozen throughout PPO training.

Two-stage training. GBVI is trained in two decoupled stages. In Stage 1, the encoder, RG-LRU memory, mixture-weight and energy-conditioning heads, potential network, and decoder are jointly optimized on environment trajectories using the Adam optimizer at learning rate 1×10^{-4}, maximizing the SIVI surrogate ELBO (Yin & Zhou, 2018) on the per-step observation likelihood. Decoupling the variational pretraining from policy optimization avoids the well-known instability of jointly optimizing reconstruction and policy objectives, and lets the policy condition on a stationary belief representation. In Stage 2, all belief-network parameters are frozen and PPO trains only the actor and critic heads on top of the frozen belief summary b_t for 10^7 environment steps.

Hyperparameter selection. GBVI architecture hyperparameters (K, d, L, ϵ, potential MLP width, RG-LRU dimension) were chosen on Bowling under the fixed-curtain mode and held fixed across the remaining environments and occlusion modes. We searched $K \in \{2, 4, 8\}$ and $d \in \{4, 8, 16\}$; $K = 4$, $d = 8$ gave the best return-per-parameter trade-off and is reported. PPO hyperparameters follow the Stable-Baselines3 Atari defaults; only the rollout length ($n_{\text{steps}} = 128$), learning rate (2.5×10^{-4}), and clip range (0.1) were adjusted to the values shared across all methods to ensure that downstream comparisons isolate the belief representation rather than the optimizer schedule.

Compute. Experiments were run on three workstations with NVIDIA RTX 4090 GPUs (24 GB) running Ubuntu (WSL) with PyTorch 2.11 and CUDA 12.x–13.x. CPU configurations across the three machines are AMD Ryzen 9 9950X3D (16 cores), Intel Core i9-14900KS (24 cores), and AMD Ryzen Threadripper PRO 5995WX (64 cores), with 30 GB to 503 GB of system RAM. Each (game, mode, method) configuration with 10^7 environment steps takes approximately 8 hours of wall-clock time on a single GPU; the full main-results sweep of 8 configurations $\times$ 6 methods totals roughly 400 GPU-hours.

G.2 Experimental Setup

This section expands on the experimental setup summarized in Section 4 with the per-game geometry, occlusion mechanics, and training/evaluation protocol details needed to reproduce our results.

Curtain mechanics. Both occlusion modes blank a fixed rectangular region of the frame; they differ in how the blanking evolves across timesteps. In *fixed curtain* mode, the rectangle is opaque on every timestep throughout the episode: pixels inside the rectangle are set to $(0, 0, 0)$, pixels outside pass through unchanged, and the rectangle stays in fixed raw-frame coordinates regardless of game state. In *flickering curtain* mode, the same rectangle alternates between opaque and transparent in fixed-length intervals: an internal counter advances once per environment step, and at every $T = 4$ steps a Bernoulli draw with probability $p = 0.5$ decides whether the rectangle is transparent for the next T-step interval. When transparent, the frame passes through unchanged and gives the agent a brief glimpse of whatever lies behind the curtain; when opaque, the region is blanked as in the fixed mode.

The number of consecutive opaque intervals is therefore geometric with mean $1/p = 2$, so a typical gap under the flickering curtain is T to $2T$ steps and rarely exceeds $4T$, whereas under the fixed curtain the gap lasts as long as the occluded object stays in the region. The curtain RNG is seeded per episode so flicker decisions are reproducible across resets, and the curtain position itself does not move with any game entity within an episode.

Per-game curtain placement. For Pong, Bowling, BeamRider, and Asteroids, the curtain rectangle is specified directly in raw $(210, 160, 3)$ ALE pixel coordinates by a top-left corner $(\text{top_row}, \text{left_col})$ and dimensions $(\text{height}, \text{width})$; for MsPacman, the rectangle is specified in maze grid coordinates and converted to pixel coordinates at frame-blackout time using the maze bounding box. Curtain placements per game (with $H_f = 210$ and $W_f = 160$ denoting raw frame dimensions for the four pixel-coordinate games) are:

- **MsPacman**: 18×18 grid-cell rectangle centered on a 31×28 maze grid. The maze is mapped from the raw ALE frame using a fixed pixel-to-grid bounding box (top-left of the maze region in pixels, total cell width / height in pixels), shared with the RAM-based ground-truth extraction so that grid-cell occlusion checks and pixel-level blackout agree exactly.
- **Bowling**: vertical strip across the lane.
- **BeamRider**: horizontal band over the enemy descent region, between the spawn zone at the top of the playfield and the player ship at the bottom; descending enemies cross the band rather than the band moving.
- **Asteroids**: vertical strip over the playfield. The screen wrap-around of Asteroids means a single asteroid may pass through the curtain region more than once per episode, exercising the cross-step belief mechanism more heavily than a one-shot occlusion would.
- **Pong**: vertical strip placed between the two paddles so the ball must cross the occluded region during volleys.

Across all five games the rectangle is positionally fixed within an episode and does not chase or follow any moving game entity. The MsPacman grid uses column-axis tunnel wrap-around to match the game's horizontal corridor wrap, while the row axis does not wrap; this is consistent with the BFS connectivity computed for the maze.

Atari preprocessing. All Atari environments share the standard preprocessing pipeline. Raw ALE frames pass first through the per-game POMDP wrapper that applies the curtain to the $(210, 160, 3)$ uint8 RGB frame, then through Stable-Baselines3's `AtariWrapper`: 84×84 grayscale resize, frame-skip 4 (with max-pooling over the last two skipped frames), no-op reset (up to 30 no-ops at episode start), reward clipping to $\{-1, 0, +1\}$, and life-loss-as-terminal during training. We then apply frame stacking with stack size 4 at the per-environment level, so each encoder input is an $84 \times 84 \times 4$ tensor. The same preprocessing chain is used for all methods in our comparison; the only architectural difference between methods is the belief representation that consumes the encoded feature.

Training protocol. All methods are trained for 10^7 environment steps using 8 parallel environments collected through Stable-Baselines3's `DummyVecEnv`. PPO is run with discount factor $\gamma = 0.99$, GAE $\lambda = 0.95$, a constant learning rate of 2.5×10^{-4}, constant clip range 0.1, rollout length $n_{\text{steps}} = 128$ per environment, minibatch size 256, 4 optimization epochs per rollout, entropy coefficient 0.01, and value coefficient 0.5. All methods share the same PPO hyperparameters and the same Stable-Baselines3 implementation; the only difference between methods is how each one maintains the belief representation passed to PPO. For GBVI, the belief network is pretrained as described in Section G.1 and then frozen during PPO training so the policy builds on a stable belief representation.

Evaluation protocol. For every (game, mode, method) combination we run 100 evaluation episodes using 100 fixed seeds. The same 100 seeds are shared across all combinations so that any pair of methods is evaluated on the same set of episode initializations and curtain RNG draws. Evaluation environments use the same preprocessing chain as training but with the life-loss-as-terminal flag disabled, so reported episode returns reflect full-life episodes. We report the mean and standard error of the episode return over the 100 episodes per combination.

Seeds and reproducibility. Each parallel training environment uses a distinct base seed offset by its environment index, and the curtain RNG is seeded independently per episode so that flicker decisions are reproducible across reset calls. Evaluation seeds are drawn from a fixed list shared across methods to ensure paired comparisons. All randomness sources — PPO sampling, curtain flicker, observation noise, and parameter initialization — use seeded generators so that results from a given configuration are reproducible up to nondeterminism in the underlying CUDA kernels.

G.3 Area Search

This section specifies the Area Search domain of Section 4.3: the map, the vehicle and sensor models, the exact posterior, the evaluation protocol, and the belief-error metric. The error as a function of gap length is in Appendix F.5.

Map. The AirSim city (Zhang et al., 2024) provides a 1000×1000 segmentation map and an obstacle-height map. Road cells are those whose segmentation class is road or crossing; obstacle cells are those whose height exceeds the UAV's flying altitude. Both are downsampled by 10 to a 100×100 grid, a coarse cell being an obstacle when more than half of its fine cells are; the free cells are the roads. Distances between free cells are shortest-path lengths along free cells, computed once by breadth-first search, so two cells on either side of a building are far apart even when adjacent in the plane (Figure 2).

Vehicles. A vehicle's state is a free cell and one of four headings. On each move it continues straight or turns into an open side road with equal probability, and reverses with probability 0.01, or always at a dead end; the new heading is the direction moved. Vehicles move once every two UAV steps. The two vehicles move independently and are distinguishable, so a reading says which vehicle produced it. The transition model is held as an explicit sparse row-stochastic matrix over (cell, heading) states, from which the simulator samples and which the exact filter multiplies by.

UAV and sensor. The UAV moves one cell per step along the four headings or stays, and catches a vehicle by reaching its cell. Its sensor covers the free cells within a Euclidean radius of 5 cells that have a clear line of sight, tested by Bresenham's line in either direction. A vehicle in this region produces a reading, the center of its cell plus isotropic Gaussian noise with standard deviation 0.5 cells; a vehicle outside it produces none. An empty reading is negative information: it has likelihood one on unseen cells and zero on visible ones, so during a gap the observation likelihood is constant over every cell the vehicle could be in. Episodes last up to 400 steps, with reward 1 per catch and a step cost of 0.01.

Exact posterior. Because the vehicles move independently and are observed separately, the exact posterior factorizes into one grid filter per vehicle over its $4\times$(number of free cells) states. Each step mirrors the simulator: the vehicle is not on the UAV's cell, on a move step the belief is pushed through the transition matrix, and the reading or its absence is applied as a likelihood; the position posterior is the sum over headings. The prior is uniform over free cells and headings.

Protocol. Scenarios fix the UAV's route and the vehicles' start states, so every method sees the same action-observation sequence and the exact posterior is the same for all. Each scenario runs for up to the 400-step budget, and how long a vehicle stays out of view is set by the route and the vehicles' paths, so the gap length varies across scenarios and within an episode; the controlled-gap scenarios of Appendix F.5 fix it instead. All methods are the belief networks of Section 4.1, trained on Area Search trajectories with the two-stage procedure of Appendix G.1, with a linear probe followed by a softmax over cells read out from the belief passed to the actor and critic (the summary b_t for GBVI). The probe has the same form for every method (Wang et al., 2023), is fit on training episodes with the true cell as target, and is frozen before evaluation. Caught vehicles are excluded from the error.

Belief-error metric. We measure the Wasserstein–Fisher–Rao distance (Chizat et al., 2010; Liero et al., 2017) between the probed belief and the exact posterior in its Hellinger–Kantorovich form. The ground distance between two cells is the shortest road path, scaled so that the cutoff of one block length, the distance between adjacent intersections, maps to $\pi/2$; the transport cost is $-\log\cos^2$ of the scaled distance and is infinite beyond the cutoff, where removing mass and re-creating it is cheaper than moving it. The optimal coupling γ is computed by unbalanced Sinkhorn in the log domain with entropic step 0.01, at which the exact posterior's error against itself is below 10^{-3}. Its objective splits into the transport part $\sum \gamma \cdot \mathrm{cost}$ and the reweighting part $\mathrm{KL}(\gamma\mathbf{1} \,\|\, \mathrm{belief}) +$

Table 7: Inference cost per timestep (GFLOPs) across LightDark and the five Atari games. The five Atari games share the same observation pipeline (84×84 grayscale, frame-stack 4) and the same CNN encoder, which dominates the per-step cost (see Table 8); per-method differences across Atari are therefore at the percent level. Atari values for each method are reported by architectural equivalence: the `torch.profiler` sweep was run on the games for which fixed-curtain checkpoints were available, and the remaining cells are filled in from the same method's measurement on a different Atari game, since the per-step architecture differs across Atari games only in the actor head's action-space dimension — a contribution at the 10^{-5} GFLOPs level that is below the precision reported. ESCORT's LightDark figure reflects its full SVGD update including kernel-matrix evaluation; on Atari, the SVGD update operates on a fixed number of latent-space components and contributes negligibly above the shared encoder cost.

Environment	GBVI	RG-LRU	LRU	HVAE	SIVAE	ESCORT
LightDark	3.02×10^{-4}	1.40×10^{-5}	1.19×10^{-5}	3.33×10^{-4}	3.87×10^{-4}	6.72
Pong	1.94×10^{-2}	1.92×10^{-2}	1.92×10^{-2}	1.89×10^{-2}	1.97×10^{-2}	1.92×10^{-2}
MsPacman	1.94×10^{-2}	1.92×10^{-2}	1.92×10^{-2}	1.89×10^{-2}	1.97×10^{-2}	1.92×10^{-2}
Bowling	1.94×10^{-2}	1.92×10^{-2}	1.92×10^{-2}	1.89×10^{-2}	1.97×10^{-2}	1.92×10^{-2}
BeamRider	1.94×10^{-2}	1.92×10^{-2}	1.92×10^{-2}	1.89×10^{-2}	1.97×10^{-2}	1.92×10^{-2}
Asteroids	1.94×10^{-2}	1.92×10^{-2}	1.92×10^{-2}	1.89×10^{-2}	1.97×10^{-2}	1.92×10^{-2}

$\mathrm{KL}(\gamma^\top \mathbf{1} \,\|\, \text{posterior})$, which sum to the reported WFR^2. Table 2 averages all three over steps and episodes and reports the belief mass on the vehicle's true cell on the last step before it returns to view.

H Computational Complexity

This section reports the inference-time computational cost of GBVI and the five baselines, both as a per-method comparison across the benchmark suite (Table 7) and as a per-component breakdown for GBVI (Table 8). All values are reported in GFLOPs for a single `model.predict` call at one timestep, profiled on CPU with `torch.profiler` (`with_flops=True`) after a five-step warm-up so the recurrent state has reached a steady-state configuration. The Atari numbers use the fixed-curtain checkpoints; flickering-curtain checkpoints yield identical per-step inference cost since the curtain mode is an environment-side construct and does not affect the model's forward pass.

H.1 Cross-method comparison

Across the Atari configurations measured, all six methods sit within $\pm 2\%$ of one another, with GBVI at 1.94×10^{-2} GFLOPs and the cheapest baseline (HVAE) at 1.89×10^{-2}. The absolute gap between GBVI and the deterministic recurrent baseline RG-LRU is 0.02×10^{-2} GFLOPs — about a 1% overhead — which quantifies the additional cost of GBVI's belief-side computation (memory readout heads, K Hamiltonian-flow components, and the $L = 2$ Leapfrog gradient evaluations per component) on top of the shared encoder and recurrent core.

The picture changes on LightDark, which uses a small MLP encoder instead of the Atari CNN. Without the encoder amortizing the per-step cost, the belief-side computation dominates: GBVI sits at 3.02×10^{-4} GFLOPs and RG-LRU at 1.40×10^{-5} GFLOPs, so GBVI's per-step cost is roughly $22\times$ that of RG-LRU, versus $1.01\times$ on Atari. The same scaling holds for HVAE, SIVAE, and ESCORT relative to RG-LRU: differences in belief representation that look small in absolute terms on Atari are visible directly on LightDark, but in both regimes the absolute cost of the GBVI-specific belief computation is bounded by the cost of $K(L{+}1)$ gradient evaluations of a small potential MLP, which is independent of the input dimensionality.

H.2 GBVI per-component breakdown

Table 8 reports GBVI's per-step FLOPs broken down by component, computed analytically from the hyperparameters in Table 6 (latent dimension $d = 8$, RG-LRU hidden dimension $D = 256$, $K = 4$ components, $L = 2$ Leapfrog steps, potential-network hidden width 64). The breakdown uses the standard convention of 2 FLOPs per multiply-add and counts only matmul-style operations;

Table 8: GBVI per-component inference cost on Atari, derived analytically from the hyperparameters in Table 6. The breakdown sums to 1.65×10^{-2} GFLOPs; the small gap to the measured 1.94×10^{-2} GFLOPs in Table 7 is attributable to operations counted by `torch.profiler` but not included in the analytical accounting (e.g., `softmax`, weight transpose copies, and activation kernels).

Component	FLOPs	GFLOPs
CNN encoder ($84\times84\times4 \to d$)	15,282,176	1.5282×10^{-2}
RG-LRU memory step	39,936	3.9936×10^{-5}
Memory readout (π + energy heads)	133,120	1.3312×10^{-4}
Hamiltonian flow ($K=4$, $L=2$ steps)	1,035,264	1.0353×10^{-3}
Actor + critic heads	38,272	3.8272×10^{-5}
Total	16,528,768	1.6529×10^{-2}

activation functions and elementwise updates contribute $O(\text{width})$ FLOPs and are absorbed into the rounding.

Of GBVI's 1.65×10^{-2} GFLOPs per timestep on Atari, 92.5% is the CNN encoder, 6.3% is the Hamiltonian flow over the four components, and the remaining 1.2% is split among the RG-LRU recurrence, the two memory-readout heads, and the actor/critic heads. GBVI's belief-side computation thus contributes a small additive term on top of the encoder cost, with the dominant term within that contribution being the $K(L+1)=12$ gradient evaluations of the potential network per timestep. Scaling either K or L moves GBVI's belief-side cost linearly while leaving the encoder cost untouched, so practitioners interested in trading per-step cost against multi-hypothesis capacity can do so by tuning K alone, with the encoder amortizing the trade-off across all values of K considered.