LEARNING NAVIGABLE WORLD MODELS VIA LATENT ENERGY SHAPING

Luiz Facury[*], **Jose Geraldo Fernandes**[*], **Pedro Dutenhefner & Wagner Meira Jr.**
Department of Computer Science
Universidade Federal de Minas Gerais

ABSTRACT

Learning generalizable policies from large, unlabeled offline datasets is a key challenge in creating autonomous agents. While offline goal-conditioned reinforcement learning (GCRL) offers a powerful framework for this, existing methods often struggle with robust long-horizon planning. Model-free approaches can fail to generalize to novel scenarios, while traditional model-based planners must contend with compounding errors and complex, multimodal search spaces that make finding a solution difficult. In this work, we introduce a novel framework that harnesses the expressive power of Energy-Based Models (EBMs) to learn a robust and navigable latent world model. Our central contribution is to adapt a training objective that explicitly shapes the energy landscape; rather than just learning the distribution of plausible transitions, it also enable goal-conditioned planning. Our method encourages the composed energy between a start and goal state to form a convex-like basin. This ensures that gradient-based planning reliably converges to a meaningful next-step latent target. Our full pipeline integrates this latent EBM with a distance-preserving state encoder and a skill-conditioned actor to ground latent plans into actions. We evaluate our approach on a suite of challenging offline GCRL benchmarks, where our experiments demonstrate that shaping the energy landscape enables long-horizon planning.

1 INTRODUCTION

A central goal in reinforcement learning is the development of generalist agents capable of solving a wide array of tasks by learning from vast, unlabeled datasets. The paradigm of learning from reward-free offline data is particularly promising, as it allows agents to acquire diverse skills from pre-existing, suboptimal trajectories without the need for costly online interaction or expert supervision. A key testbed for such agents is offline goal-conditioned reinforcement learning (GCRL), where the agent must learn to navigate from any state to any other state, a fundamental capability for spatial reasoning and long-horizon decision-making.

Two dominant approaches have emerged to approach this challenge: model-free reinforcement learning and model-based optimal control. Model-free methods, such as those that learn distance-preserving representations or goal-conditioned value functions, excel at learning policies that can stitch together fragments of offline trajectories. Model-based approaches, in contrast, learn a dynamics model of the world and then plan actions to reach a desired goal, often demonstrating superior generalization to novel environments. However, both paradigms face a fundamental challenge: navigating the complex, often ambiguous space of possible futures. Model-free policies can struggle to generalize to out-of-distribution goals, while planners relying on learned dynamics can suffer from compounding model errors, leading to unrealistic or inefficient plans.

In this work, we propose a novel framework that harnesses the expressive power of Energy-Based Models (EBMs) to create a robust and navigable world model for goal-conditioned planning. Our key insight is that while EBMs are powerful at capturing the distribution of plausible state transitions, their resulting energy landscapes can be multimodal and difficult to optimize for long-horizon planning. We address this by adapting a training objective (Gladstone et al., 2025) to the problem

[*]Equal contribution

of latent space planning, so that it explicitly shapes the energy landscape between a start and goal state to be convex-like. Instead of merely penalizing implausible transitions, our method encourages the composed energy path between a start and goal state to have a unique, well-defined minimum that corresponds to the true intermediate state from the training data. This effectively transforms the complex, non-convex planning problem into a simple and reliable gradient-based search for the optimal next step.

We posit that a World Model consists of two parts: a State Representation (HILP) and a Transition Dynamics Model (Our EBM). The HILP encoder provides a static metric space, but lacks knowledge of causality or physical constraints. The EBM learns the causal structure of the environment, predicting valid next-step transitions by explicitly shaping the energy landscape to differentiate between physically impossible shortcuts and feasible trajectories.

Specifically, our full pipeline integrates this latent transition EBM with a distance-preserving encoder, from Hilbert foundation policy (HILP) (Park et al., 2024b), which provides a geometrically meaningful latent space. This creates a powerful planning module that, given a start and goal state, can generate a coherent sequence of latent waypoints that form a viable plan. This plan can, in turn, serve as a sequence of subgoals for a low-level policy. By pre-training these components on offline data, we create an agent that can, at test time, efficiently determine the next-step latent target at each stage by simply descending a smooth energy funnel toward the final goal, without sampling states from the training set. Our preliminary experiments on challenging navigation benchmarks suggest that this approach enables robust planning in complex environments.

To summarize, our contributions are:

- We introduce a latent transition EBM trained with an objective that *shapes* the energy surface between start and goal latents so that gradient-based refinement has a unique, meaningful minimizer located at the true next-step latent.
- We provide an empirical and preliminary study on challenging offline goal-conditional benchmarks that suggest competitive long-horizon planning without sampling states from the training set.

2 Related Work

Offline Goal-Conditioned Reinforcement Learning. Offline goal-conditioned RL aims to learn policies that can reach arbitrary goal observations from purely offline data. Value-based approaches such as Implicit Q-Learning (IQL) (Kostrikov et al., 2021) have been adapted to this setting. Hierarchical Implicit Q-Learning (HIQL) (Park et al., 2023) then builds on IQL by introducing a hierarchical decomposition with a high-level goal/subgoal proposer and a low-level goal-reaching controller. Representation-driven, policy-based methods like HILP (Park et al., 2024b) learn a goal metric in an embedding space so that distances reflect temporal reachability, and train a direction-conditioned "foundation" policy capable of composing long-horizon behaviors via latent planning. Complementing these, model-based methods plan in learned latent dynamics, PLDM (Sobal et al., 2025) fits a latent world model from offline trajectories and performs planning in that latent space to reach visual goals, demonstrating strong generalization without online interaction. Advancing the scalability of such latent planning, TD-MPC (Hansen et al., 2022; 2023) performs local trajectory optimization within an implicit, decoder-free world model. Recent extensions of this framework utilize robust architectures and task-agnostic hyperparameters, allowing them to scale effectively to massively multi-task continuous control domains.

Beyond these paradigms, sequence modeling approaches like Decision Transformer (Chen et al., 2021) recast offline RL as a conditional token generation problem, leveraging transformer architectures to predict actions based on return-to-go and state histories. Similarly addressing offline data structures, Latent Action Policies (LAPO) (Schmidt & Jiang, 2023) addresses the scarcity of action labels in large offline corpora by inferring *latent actions* directly from observation-only sequences. It jointly trains an inverse-dynamics model and a forward-dynamics model with a vector-quantized information bottleneck to enforce predictive consistency, producing a latent-action policy via behavior cloning that can be *decoded* into the true action space with a small labeled set or rapidly fine-tuned online. Orthogonal advances explore gradient-based planning through differentiable world models (SV et al., 2023), enabling backpropagation through the model to optimize goal achievement

under offline training. Leveraging stronger visual priors, DINO-WM (Zhou et al., 2024) shows that world models trained on pretrained DINOv2 features can support zero-shot planning to image-specified goals. On the representation side, contrastive formulations recast goal-reaching as learning state–goal embeddings whose inner product approximates goal-conditioned value (Eysenbach et al., 2022), while large-scale "play" datasets enable learning latent plans that can be retrieved and executed to reach diverse goals without task-specific supervision (Lynch et al., 2020). Together, these lines—hierarchical value learning, embedding-based foundation policies, latent-dynamics planning, action-free latent action discovery, differentiable planners, and self-supervised visual representations—provide complementary toolkits for scalable offline goal-conditioned control.

Energy Based Models. Energy-based models (EBMs) define an energy (unnormalized negative log-probability) over inputs (or input–output pairs) and are typically trained by contrastive objectives. In practice, one pushes down energy on observed data and up on off-manifold samples via e.g. Contrastive Divergence or Noise Contrastive Estimation (Hinton, 2002; Gutmann & Hyvärinen, 2010). Recently, Gladstone et al. (2025) propose a novel *Energy-Based Transformer* (EBT) that embeds self-attention into the EBM paradigm. In EBT, a transformer network assigns an energy to each (input, candidate) pair, and inference (prediction) is done by iteratively updating the input to minimize that energy, effectively performing gradient-based "System 2" reasoning. This design shows better scaling properties than conventional feed-forward EBMs. Beyond contrastive learning, other major EBM training paradigms include score matching and denoising methods. For example, denoising-score matching (Song et al., 2020) trains an EBM by matching the score (gradient of log density) using noise-perturbed data. Such methods, as well as related generative techniques like diffusion models (Sohl-Dickstein et al., 2015), sidestep explicit sampling during training. In the context of decision-making, Diffuser (Janner et al., 2022) leverages this equivalence by framing offline planning as a score-matching diffusion process over sequences of states and actions. Typically, sampling from an EBM requires MCMC-based inference: one runs Langevin dynamics (gradient steps plus noise) on the energy landscape to generate low-energy samples. These alternatives – contrastive, score-matching/denoising, and Langevin-based sampling – constitute the main families of modern EBM techniques.

Energy Based World Models. Energy-based models have also been applied to world modeling and planning. Florence et al. (2022) introduce Implicit Behavioral Cloning (IBC) where an expert policy is modeled implicitly by an EBM rather than an explicit policy network. The EBM assigns low energy to state–action pairs from demonstrations, and actions are sampled by Langevin inference. IBC finds that these implicit EBMs outperform standard behavioral cloning in complex robotic tasks and even rival offline RL methods, despite using no reward signal. Boney et al. (2020) consider model-based planning: they train an energy estimator on real state–transition samples (offline data) and then use it as a regularizer during planning with a learned dynamics model. The planner is penalized for generating transitions that receive high energy, which improves robustness to model errors.

Other recent works explore EBMs in world-model contexts. For instance, Du et al. (2020) show that an EBM can serve as a next-state dynamics model: they learn an energy function over successor states (given current state) and then plan in the state space by sampling via gradient-based inference. Since EBMs can capture multi-modal transition distributions, they naturally support diverse goal-reaching plans and maximum-entropy exploration. Zhang et al. (2023) propose *Energy-based Predictive Representations* (EPR) for RL: a neural EBM is trained to encode a predictive latent state from observations, which in turn permits implicit approximation of the Q-function and uncertainty-aware planning. These examples illustrate how EBMs can underpin world models and value estimation – either by modeling transition densities or by learning latent state features for control – complementing classical world-model approaches. Notably, these implicit and energy-based formulations offer a compelling alternative to dominant generative world models, such as DreamerV3 (Hafner et al., 2023), which rely on explicit observation decoding—a process that can dedicate excessive model capacity to visually complex but task-irrelevant details.

3 METHOD

We utilize an OGBench (Park et al., 2024a) offline dataset comprising multiple trajectories, $\mathcal{D} = \{\tau_k\}_k^K$, where each trajectory $\tau_k = \{s_{k,t}\}_t^{n_k}$ consists of a sequence of states. The objective is to train the EBM such that consecutive state pairs within these trajectories yield low energy scores, $E_\phi(s_{k,t}, s_{k,t+1})$ (from now on we suppress the index k for simplicity). Our pipeline has three core components: (i) a distance-preserving encoder $\phi : \mathcal{S} \to \mathcal{Z}$ that embeds states into a Hilbert space where $\|\phi(s) - \phi(g)\|$ correlates with temporal (step) distance; (ii) a latent *transition* energy-based model $E_\theta(z_t, z_{t+1})$ that scores candidate latent transitions; and (iii) leveraging this low-level transitions to a proper goal-conditioned for planning.

Encoder pretraining. We first learn ϕ using a distance-preserving objective, following HILP Park et al. (2024b): the encoder is trained so that Euclidean distances in $\mathcal{Z}$ reflect the temporal separation between states in the offline dataset. This makes nearby latent points meaningful subgoals and yields a latent geometry that supports both sampling of intermediate waypoints and local refinement.

Latent EBM with convex shaping. On top of ϕ we train a transition EBM $E_\theta(z_t, z_{t+1})$ that assigns low energy to plausible one-step latent transitions and higher energy to unlikely pairs, through convex shaping. Crucially, rather than leaving the EBM unconstrained and potentially multimodal, we explicitly shape the energy landscape along the segment between a start latent z_t and a goal latent z_{goal}. Concretely, for true transitions $(z_t, z_{t+1}, z_{\text{goal}})$ sampled from trajectories we minimize the energy for the observed pairs (z_t, z_{t+1}) with a predicted latent state just like Gladstone et al. (2025)

$$\hat{z}_{i+1} = \hat{z}_i - \alpha \nabla \mathcal{E}(z_t, \hat{z}_i)$$

to coincide with the true intermediate latent z_{t+1}, where α is a *gradient* step size and i its refinement iteration. This is a contrastive-free approach that reduces harmful multimodality along the start–goal corridor and produces a convex-like energy basin whose unique minimizer corresponds to the desired one-step latent. In practice the refinement-target term can be implemented by running a few gradient steps of $\mathcal{E}_{t\to\text{goal}}$ from the reference state z_t, not random initializations like EBT, and penalizing the distance between the refined candidate and z_{t+1}.

Long-Horizon Planning via Composed Energy Minimization. During planning, the agent leverages the locally-trained EBM to find a globally optimal path. Given a current latent state z_t and a final goal z_{goal}, the agent determines its next move by finding an intermediate subgoal $\hat{z}$ whose gradient points to lowest-energy path, after refinement in i iterations. This subgoal is now found by minimizing the composed energy, which is the sum of the energy from the current state to the subgoal and from the subgoal to the final goal:

$$\hat{z}_{i+1} = \hat{z}_i - \alpha \nabla (\mathcal{E}(z_t, \hat{z}_i) + \mathcal{E}(\hat{z}_i, z_{\text{goal}}))$$

This composed energy function is shaped to have a unique minimizer that represents the optimal next latent space. To ensure physical plausibility, the agent does not jump directly to this subgoal. Instead, it interprets z as a direction for travel. The next state in the plan, z_{t+1}, is determined by taking a fixed-size step from the current state in the direction of the optimal subgoal:

$$z_{t+1} = z_t + \beta \frac{\hat{z} - z_t}{||\hat{z} - z_t||}$$

Plan Execution via Low-Level Control. To ground the generated latent plans into physical actions, we utilize the pre-trained goal-conditioned policy $\pi_{\text{low}}(a|s, z)$ proposed by (Park et al., 2024b). While our Energy-Based World Model is responsible for the high-level reasoning—generating a sequence of navigable latent waypoints $\{z_1, z_2, \ldots, z_g\}$—the HILP policy acts as the local controller. At each timestep t, the current observation s_t and the immediate latent target z_{target} (derived from the EBM planner) are fed into π_{low} to sample the executable action $a_t \sim \pi_{\text{low}}(\cdot|s_t, z_{target})$. This reliance on the HILP actor serves as a strong validation of our world model: the quantitative success of the agent (Table 1) demonstrates that our EBM effectively produces subgoals that are not only geometrically consistent but also dynamically feasible for the low-level controller to reach.

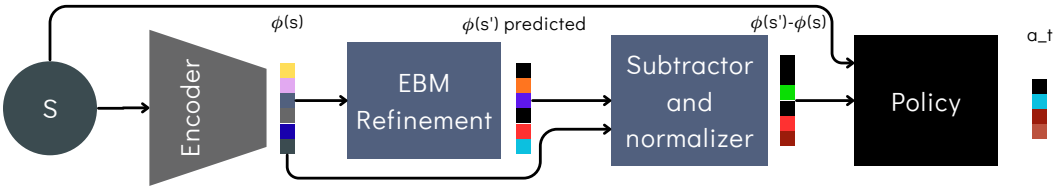

Figure 1: **Architecture of the Energy-Guided Planner.** The pipeline begins by mapping the input state S to a latent representation $\phi(s)$ using the pre-trained **Encoder**. The **EBM** defines a convex energy landscape, which guides the **Refinement** process (via gradient descent) to identify the optimal next latent state, $\phi(s')_{\text{predicted}}$ conditioned to the goal. To feed it to the policy network, a **Subtractor** and **Normalizer** compute the normalized direction vector between the current and predicted states (representing the term $\frac{z_{\text{subgoal}} - z_t}{\|z_{\text{subgoal}} - z_t\|}$). Finally, this directional goal is fed into the low-level **Policy**, which executes the action a_t required to move the agent along the planned trajectory.

3.1 Architecture and Training Stability

To ensure that our energy landscape remains well-conditioned for gradient-based planning, we parameterize the energy function $E_\theta(z_t, z_{t+1})$ as a Multi-Layer Perceptron (MLP) regularized with Spectral Normalization (Miyato et al., 2018). This constraint on the Lipschitz constant of the network is critical for stable energy-based training, preventing the energy surface from becoming too sharp or exhibiting pathological gradients that would hinder the planning process.

Furthermore, we adopt the robust training protocol proposed by the Energy-Based Transformer (EBT) framework (Gladstone et al., 2025) to regularize the optimization landscape. Specifically, we employ a **variable-horizon refinement scheme** during training, where the number of gradient steps, the step size α, and the noise scale for Langevin dynamics are randomized at each iteration. We also maintain a replay buffer of persistent negative samples to ensure the model continuously suppresses high-energy regions visited in previous iterations. This stochasticity during training forces the EBM to learn a smooth, convex-like basin around the target transition, ensuring that the inference-time planner is robust to initialization noise and capable of scaling to longer horizons.

4 Experiments

We evaluate our method on the OGBench navigation tasks (Park et al., 2024a). For all quantitative evaluations, we employ a hierarchical control scheme: the proposed Energy-Based planner replans the latent trajectory every step, providing a local subgoal to the HILP low-level actor, which executes the actions in the environment. We report the success rate, defined as the agent reaching within ϵ-distance of the physical goal state within the episode time limit.

4.1 Zero-Shot Latent Space Planning

To test the core capability of our approach, we task the planner with generating a latent trajectory from a start state to a distant goal state within the same episode. The planning process follows the iterative refinement procedure described in our Method section: at each step, the planner identifies an optimal intermediate subgoal by finding the minimizer of the composed energy landscape. It then takes a fixed-size step in the latent direction of this subgoal.

The result of this process is visualized in Figure 2. Since the planned waypoints z_t exist in a high-dimensional latent space without a direct decoder, we visualize the plan by plotting the nearest neighbor state from the offline dataset for each waypoint, measured by Euclidean distance in the latent space $\mathcal{Z}$. The sequence of these nearest neighbors forms a continuous and viable trajectory that successfully navigates the maze corridors from the start to the goal. This qualitatively demonstrates that our energy-shaping objective is effective, creating a navigable latent space by simply following the negative gradient of the energy.

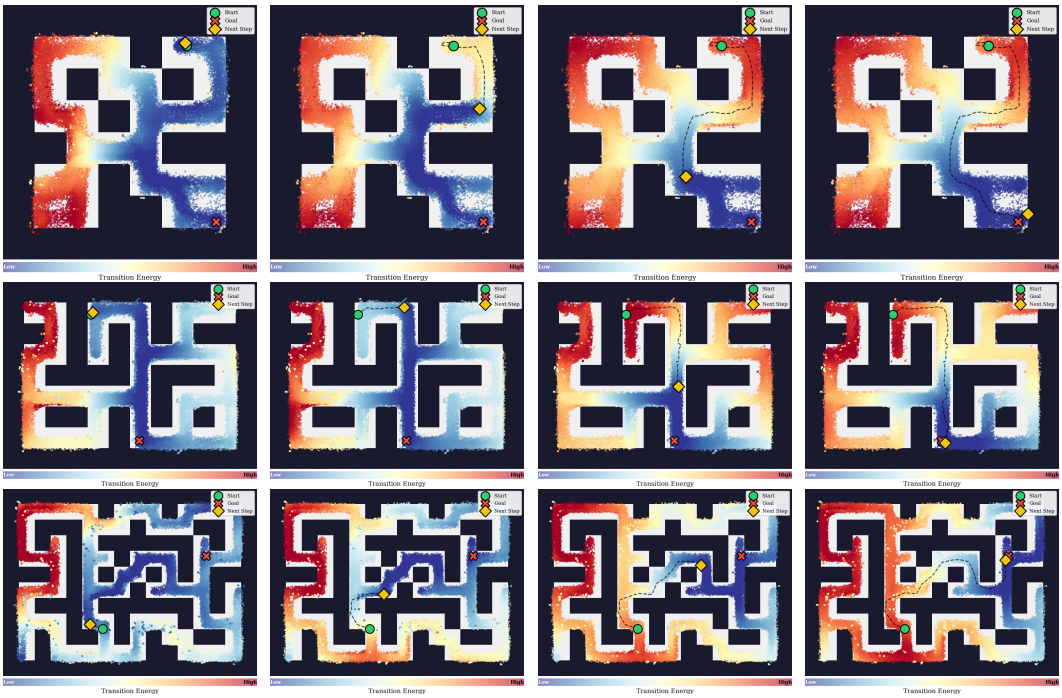

Figure 2: Energy landscape from intermediate planned states (z_t). The states are sampled from the offline dataset and colored by their energy relative to current state. This showed a tunnel of low energy between the start and the goal.

4.2 Analysis of the Navigable Energy Landscape

To understand *how* the planner makes its decisions at each step, we visualize the energy landscape it perceives. Figure 2 shows the one-step energy landscape from an intermediate point z_t along the planned path. As the full state space s is high-dimensional (including agent velocity, joint angles, etc.), a simple 2D grid evaluation is not possible. Instead, we sample a large corpus of states from the entire offline dataset and plot their 2D positions, colored by their energy relative to the current planned state z_t.

The figure clearly shows that the lowest-energy region (blue) is concentrated directly ahead of the agent's current position along the correct corridor toward the goal. The five sampled states with the lowest energy, highlighted in magenta, all lie within this forward-facing corridor. This provides strong evidence that our training objective was successful: the EBM has not just learned local dynamics, but a locally *goal-aware* energy landscape. When composed, this allows the planner to "funnel" its decisions, reliably choosing the correct direction at each step, even at ambiguous intersections.

4.3 Results

We present the quantitative performance of our energy-guided planner across various OGBench navigation tasks, comparing directly against multiple offline goal-conditioned baselines: GCBC (Lynch et al., 2020; Ghosh et al., 2019), GCIQL (Kostrikov et al., 2021), CRL (Eysenbach et al., 2022), and the HILP (Park et al., 2024b) foundation policy (Table 1). The results demonstrate that explicitly shaping the energy landscape significantly improves long-horizon planning capabilities, particularly in environments lacking structured offline trajectories.

Performance on Standard Navigation. In the standard `Navigate` tasks, our method achieves consistent and competitive performance. While baselines like CRL exhibit strong results on specific medium and large navigation tasks (e.g., $94 \pm 1\%$ on `Antmaze-Medium`), our planner scales more reliably across the board and strongly outperforms the HILP baseline. For instance, on the

Table 1: Quantitative results on OGBench navigation tasks. We compare GCBC, GCIQL, CRL, and HILP against our proposed method. Each goal is evaluated with 25 rollouts and averaged over 5 random seeds (mean $\pm$ std). Overall baseline averages are calculated as the unweighted mean across all 14 tasks.

Env	Type	Size	GCBC	GCIQL	CRL	HILP	Ours
PointMaze	Navigate	Medium	10.4 ± 6	51.8 ± 8	28.2 ± 7	54.1 ± 3	$\mathbf{59.4 \pm 4}$
		Large	27.8 ± 6	34.6 ± 3	40.5 ± 7	42.7 ± 6	$\mathbf{52.3 \pm 5}$
		Giant	1.2 ± 2	0.0 ± 0	$\mathbf{25.7 \pm 10}$	21.2 ± 5	20.4 ± 4
	Stitch	Medium	21.7 ± 18	22.4 ± 9	0.4 ± 1	$\mathbf{33.4 \pm 14}$	31.5 ± 12
		Large	8.3 ± 5	$\mathbf{29.8 \pm 2}$	0.0 ± 0	16.5 ± 15	23.1 ± 10
		Giant	0.0 ± 0	0.0 ± 0	0.0 ± 0	0.0 ± 0	$\mathbf{3.6 \pm 2}$
AntMaze	Explore	Medium	2.1 ± 1	13.7 ± 2	4.3 ± 2	32.7 ± 10	$\mathbf{34.4 \pm 8}$
		Large	0.0 ± 0	0.0 ± 0	0.0 ± 0	7.3 ± 8	$\mathbf{10.1 \pm 10}$
	Navigate	Medium	30.2 ± 4	69.5 ± 4	$\mathbf{94.1 \pm 1}$	82.1 ± 3	90.5 ± 5
		Large	24.6 ± 2	33.2 ± 4	$\mathbf{83.8 \pm 4}$	73.2 ± 4	80.7 ± 3
		Giant	0.0 ± 0	0.0 ± 0	16.7 ± 3	$\mathbf{43.1 \pm 8}$	42.8 ± 5
	Stitch	Medium	43.8 ± 11	28.3 ± 6	51.6 ± 6	90.2 ± 3	$\mathbf{91.8 \pm 4}$
		Large	3.9 ± 3	6.1 ± 2	11.8 ± 2	64.1 ± 7	$\mathbf{68.5 \pm 5}$
		Giant	0.0 ± 0	0.0 ± 0	0.0 ± 0	0.0 ± 0	$\mathbf{2.3 \pm 1}$
Average			12.4	20.7	25.5	40.0 ± 6.1	$\mathbf{43.7 \pm 5.6}$

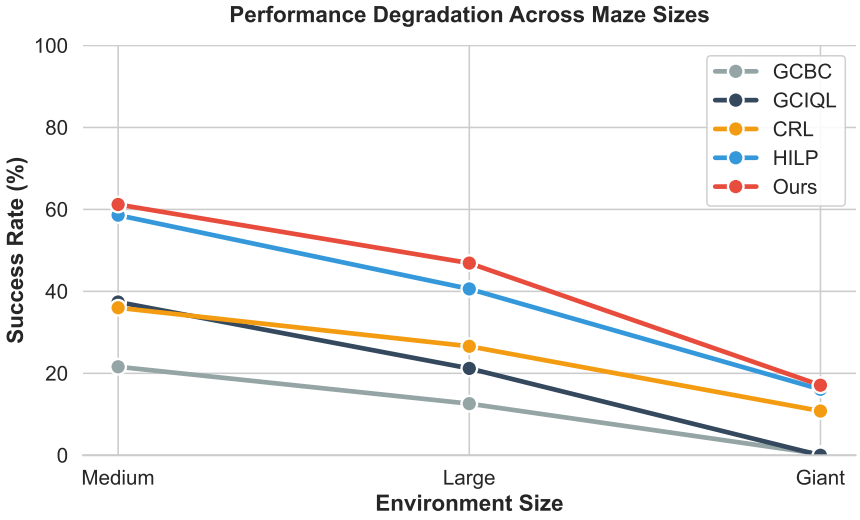

Figure 3: Performance degradation across maze sizes. The EBM-guided planner (Ours) retains higher success rates on long-horizon (Giant) tasks compared to standard offline baselines.

`Antmaze-Large` benchmark, our planner achieves a success rate of $80.7 \pm 3\%$, outperforming HILP's $73.2 \pm 4\%$. Figure 3 illustrates the performance degradation as environment size increases. Our method exhibits a notably shallower degradation slope from Medium to Giant mazes compared to traditional baselines like GCIQL and GCBC, validating that the convex-like energy basin successfully guides the agent through longer horizons where standard metrics struggle.

Robustness in Unstructured Environments. As noted in our limitations, datasets like `Stitch` and `Explore` lack the temporal structure required for perfect geometric embedding. This lack of structure causes standard offline RL methods to catastrophically fail; CRL, GCIQL, and GCBC all drop to near-zero success rates on the `Stitch` and `Explore` tasks, particularly in larger mazes.

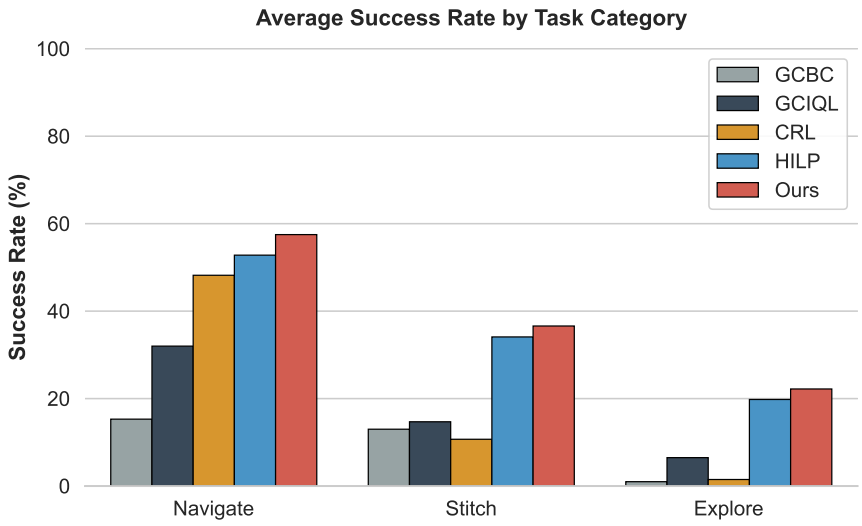

Figure 4: Average success rate grouped by task category. While standard baselines perform adequately on simple navigation, they collapse on unstructured datasets (Stitch/Explore). Our method significantly outperforms prior approaches.

In stark contrast, our method demonstrates superior robustness. As highlighted in Figure 4, our method substantially outperforms all baselines in these unstructured categories. For example, in `Antmaze-Explore` (Large), our method achieves $10.1 \pm 10\%$, setting the best score for the task. The energy-based refinement provides a critical gradient signal for planning that remains functional even when the underlying data is heavily fragmented.

Overall Performance. Across all evaluated tasks and difficulties, our method achieves the highest average success rate of $43.7 \pm 5.6\%$, compared to $40.0 \pm 6.1\%$ for HILP, 25.5% for CRL, 20.7% for GCIQL, and 12.4% for GCBC. This confirms that learning a navigable latent world model via energy shaping is a highly viable strategy for improving offline goal-conditioned planning, yielding superior aggregate results without the need for online interaction.

5 Limitations and Conclusion

Limitations. A foundational dependency of our pipeline is the pre-trained HILP state encoder, whose effectiveness hinges on learning a meaningful temporal distance metric from offline trajectories. While our energy-shaping planner significantly improves robustness on moderately unstructured datasets—allowing us to outperform baselines on many `stitch` and `explore` tasks—it remains bottlenecked by the underlying representation quality in extreme cases. On the most complex and expansive datasets (e.g., 'Giant' environments with highly disjointed or suboptimal paths), the temporal structure required for geometric embedding becomes highly fragmented. As shown in Figure 5, in these severe cases, the resulting latent space loses its geometric integrity, with latent distances no longer correlating with physical reachability. This breaks the core assumption of our gradient-based planner, which relies on a smooth, navigable energy landscape to find the 'downhill' path to the goal.

Conclusion. Despite these limitations, this work presents a step towards robust offline planning. We have demonstrated that by explicitly shaping the energy landscape of a latent dynamics model, an agent can successfully discover and generate valid, long-horizon plans in complex environments without suffering from the compounding errors that affect many model-based methods. The modularity of our planner is a key advantage. Future work will focus on two key areas: first, developing methods that constrain the step size in the planning while also keeping the latent targets in the data manifold; this could involve a weighted sum of the energies instead of using the predicted state direction; second, exploring alternative state representation learning schemes that are more robust to

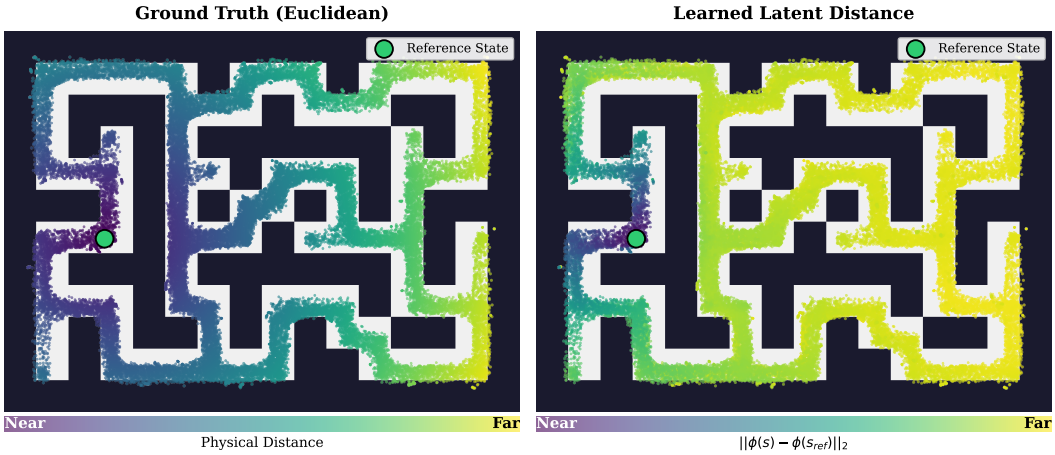

Figure 5: A visualization of the HILP encoder's learned distance metric on a dataset with poor temporal structure (e.g., `stitch` or `explore`). (Right) Ground-truth physical distance from a start point. (Left) The learned latent distance is unstructured, failing to capture the maze geometry. This makes gradient-based planning infeasible, since very similar distances are assigned to distant points.

unstructured data, thereby extending our planning framework to a wider range of challenging offline reinforcement learning problems.

REFERENCES

Rinu Boney, Juho Kannala, and Alexander Ilin. Regularizing model-based planning with energy-based models. In *Conference on Robot Learning*, pp. 182–191. PMLR, 2020.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Yilun Du, Toru Lin, and Igor Mordatch. Model-based planning with energy-based models. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 100 of *Proceedings of Machine Learning Research*, pp. 374–383. PMLR, 2020. URL `https://proceedings.mlr.press/v100/du20a.html`.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on robot learning*, pp. 158–168. PMLR, 2022.

Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated supervised learning. *arXiv preprint arXiv:1912.06088*, 2019.

Alexi Gladstone, Ganesh Nanduru, Md Mofijul Islam, Peixuan Han, Hyeonjeong Ha, Aman Chadha, Yilun Du, Heng Ji, Jundong Li, and Tariq Iqbal. Energy-based transformers are scalable learners and thinkers. *arXiv preprint arXiv:2507.02092*, 2025.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In *Conference on robot learning*, pp. 1113–1132. Pmlr, 2020.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. Hiql: Offline goal-conditioned rl with latent states as actions. *Advances in Neural Information Processing Systems*, 36:34866–34891, 2023.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. *arXiv preprint arXiv:2410.20092*, 2024a.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. *arXiv preprint arXiv:2402.15567*, 2024b.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. *arXiv preprint arXiv:2312.10812*, 2023.

Vlad Sobal, Wancong Zhang, Kynghyun Cho, Randall Balestriero, Tim GJ Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. *arXiv preprint arXiv:2502.14819*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in artificial intelligence*, pp. 574–584. PMLR, 2020.

Jyothir SV, Siddhartha Jalagam, Yann LeCun, and Vlad Sobal. Gradient-based planning with world models. *arXiv preprint arXiv:2312.17227*, 2023.

Tianjun Zhang, Tongzheng Ren, Chenjun Xiao, Wenli Xiao, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Energy-based predictive representations for partially observed reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 2477–2487. PMLR, 2023.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.