ELVIS: Ensemble-Calibrated Latent Imagination for Long-Horizon Visual MPC

Yurui Du[1,3], Pinhao Song[2,3], Yutong Hu[2,3], and Renaud Detry[1,2,3]

***Abstract*—A central challenge of visual control with model-based reinforcement learning (RL) is *reliable long-horizon planning*: Long rollouts with learned latent dynamics exhibit branching futures and multi-modal action-value distributions. In addition, compounding model error amplified by visual occlusions make deep imagination brittle. We present ELVIS, a latent model predictive controller (MPC) designed to make long-horizon planning practical. ELVIS plans in a Dreamer-style recurrent state space model (RSSM) and replaces standard unimodal model predictive path integral (MPPI) with a *Gaussian-mixture MPPI* that maintains multiple coherent hypotheses over long horizons, avoiding mode averaging under branching rollouts. In parallel, ELVIS stabilizes deep imagination with a *shared uncertainty-aware λ_t-return*: an ensemble of latent critics defines an upper-confidence-bound (UCB) score that gates a time-varying λ_t, adaptively trading off bootstrapping versus look-ahead to limit compounding error during planning. The same return is used both to train an actor-critic prior from imagined rollouts and to score candidate trajectories inside GMM-MPPI, aligning RL objectives with the planner's long-horizon optimization. On fourteen DeepMind Control Suite visual tasks, ELVIS establishes state-of-the-art performance compared with TD-MPC2 [8] and DreamerV3 [7]. Finally, ELVIS transfers zero-shot to a real-world sand spraying task with severe occlusions, improving surface-quality metrics and demonstrating robustness beyond simulation. Our code is publicly available at https://github.com/RILEY-REEDUS/ELVIS.**

I. Introduction

Model-based RL has emerged as a promising alternative to model-free methods, achieving state-of-the-art sample efficiency by learning a world model and reducing reliance on exhaustive trial-and-error. Typically, one first learns a predictive dynamics model via supervised training, then leverages it to either (i) generate imaginary rollouts that augment policy/value learning [5, 6, 7], or (ii) support an MPC that optimizes predicted trajectories online to provide world-informed control decisions [8, 9, 17]. Short-horizon planning/imagined rollouts provide effective "look-ahead" that reduces wasteful real-world exploration and improves data efficiency. Finally, when MPC is warm-started or guided by a learned policy prior, the planner can reliably choose actions that outperform the raw policy, improving sampling efficiency and stability [22].

Despite these advances, MPC in visual control is often limited to short receding horizons (e.g., 3–5 steps), largely because longer-horizon latent rollouts are both computationally expensive and increasingly error-prone. As the rollout horizon grows, model errors compound and imagined returns become less reliable; this problem becomes especially acute under partial observability, where occlusions reduce visual evidence and force the planner to rely more heavily on latent imagination [11]. Nonetheless, longer horizons should, in principle, substantially improve sample efficiency, provided planning can *reason about and react to uncertainty* so that unreliable imagined rollouts do not dominate decisions. A second limitation is the Markov assumption underlying many MPC instantiations. RL-based methods often heuristically *stack a short history of frames* as a Markov proxy—e.g., DQN stacks 4 Atari frames—then learn on this proxy state [16]. While this can mitigate partial observability in benign settings, it quickly breaks under occlusions, viewpoint shifts, or distractors [21], leading to state aliasing and drift over longer rollouts. In such regimes, agents need *memory/belief state* rather than short stacks [12].

To address these limitations, we propose ELVIS, an uncertainty-aware, memory-augmented MPC framework for visual model-based RL under partial observability. Two existing model-based RL lines are especially relevant to this setting. Dreamer-style agents [7] use recurrent state-space models (RSSMs) to maintain latent belief under partial observability, making them a natural foundation for visual control with occlusion. In parallel, TD-MPC [9] and TD-MPC2 [8] show that online trajectory optimization in the latent space of a learned world model can yield strong control performance. ELVIS combines these strengths: it uses a Dreamer-style recurrent latent belief for memory under occlusion, while extending TD-MPC2-style latent planning to longer horizons through multimodal proposals and uncertainty-aware return modulation.

Concretely, we employ an RSSM [5] to capture aleatoric uncertainty in high-dimensional observations through its stochastic latent state, while modeling epistemic uncertainty via an ensemble of value networks [14]. We instantiate informative

[1]Yurui Du and Renaud Detry are with KU Leuven, Dept. Electrical Engineering, Research unit Processing Speech and Images, B-3000 Leuven, Belgium. E-mail: {yurui.du, renaud.detry}@kuleuven.be.

[2]Pinhao Song, Yutong Hu, and Renaud Detry are with KU Leuven, Dept. Mechanical Engineering, Research unit Robotics, Automation and Mechatronics, B-3000 Leuven, Belgium. E-mail: {pinhao.song, yutong.hu}@kuleuven.be.

[3]Yurui Du, Pinhao Song, Yutong Hu, and Renaud Detry are with Flanders Make at KU Leuven, B-3000 Leuven, Belgium.

Funded by the European Union

This work was supported by the European Union (robetarme-project.eu). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or HADEA. Neither the European Union nor the granting authority can be held responsible for them.

data collection while guarding against brittle long-horizon predictions using uncertainty-based soft truncation that continuously trades off short-horizon bootstrapping against long-horizon rollout returns. This yields a unified interface between recurrent memory, long-horizon planning, and uncertainty-aware control.

Our paper makes the following contributions:

1) Adaptive-horizon, memory-augmented MPC for visual RL. A framework that adapts the effective return horizon online under controlled model error, using real-time confidence thresholds and compute–performance trade-offs, built on a recurrent latent world model for partial observability.

2) Uncertainty-regulated exploration and exploitation. A unified exploration-exploitation scheme bounded by soft truncation, yielding uncertainty-aware actions and more reliable long-horizon plans.

3) We evaluate ELVIS on both DeepMind visual control benchmarks and challenging real-world tasks featuring extreme partial occlusions.

Collectively, these elements produce a practical recipe for long-horizon, uncertainty-aware planning in visual model-based RL. Across various settings, our method consistently improves data efficiency, robustness under partial observability, and zero-shot transfer performance compared to state-of-the-art model-based RL baselines [7, 8].

II. Related Work

A. Sampling-Based MPC and Multimodal Latent Planning

Model Predictive Path Integral (MPPI) control provides a sampling-based framework for online trajectory optimization in nonlinear systems [23], and TD-MPC brings this idea into learned latent spaces for continuous control [9, 8]. ELVIS follows this latent-MPC lineage but targets a long-horizon regime where unimodal TD-MPC2-style proposals can become brittle: branching imagined futures may cause a single Gaussian to average over incompatible action-sequence hypotheses. We address this with multimodal MPPI in latent imagination. Existing multimodal MPPI methods in reactive task-and-motion planning and collision avoidance handle modes induced by high-level plan alternatives or maneuver choices [24, 1], while Bayesian MBRL methods such as PaETS also argue for maintaining multimodal trajectory hypotheses [18]. In contrast, ELVIS studies multimodal proposals inside a recurrent latent world model and couples them to a shared uncertainty-aware return used by both imagined learning and online planning.

B. Uncertainty Quantification in Model-Based RL

Uncertainty-aware model-based RL has been studied from several angles: PETS propagates epistemic uncertainty in learned dynamics [2], Plan2Explore uses world-model disagreement as an intrinsic reward [19], and recent latent-space methods use epistemic uncertainty for out-of-distribution detection and safety filtering [20]. ELVIS differs in the role assigned to uncertainty. Rather than using it primarily for dynamics propagation, exploration, or safety triggering, we use critic uncertainty to gate a time-varying λ_t, softly modulating imagined returns. This aligns the uncertainty signal with the return propagation, while avoiding the cost of maintaining uncertainty estimates inside the RSSM dynamics.

C. Optimism, Value-Guided Planning, and Trust-Aware Rollouts

ELVIS is also related to ensemble-based optimism and trust-aware rollout control. SUNRISE uses ensemble-UCB style optimism for exploration [14], and POLO combines online trajectory optimization with learned value functions and uncertainty-aware exploration [15]. However, neither SUNRISE nor POLO is designed to address when long-horizon imagined rollouts should be trusted during learning. MBPO and MACURA instead regulate model usage by scheduling, shortening, or adapting rollout lengths based on model uncertainty [13, 4]. ELVIS uses critic uncertainty for a different purpose: it modulates how strongly distant imagined returns propagate inside a fixed horizon. The resulting time-varying λ_t reduces the influence of uncertain distant returns during both imagined actor-critic learning and MPPI scoring. This distinction is computationally relevant in our online setting, where variable rollout lengths can introduce system overhead such as more frequent recompilation, whereas a fixed-horizon planner with soft return modulation preserves a stable execution path.

III. Preliminaries

This section summarizes the recurrent state-space model (RSSM) used throughout the paper and clarifies how it supports control under partial observability. We first describe RSSM as a latent-variable sequence model trained by variational inference, then explain how the learned latent dynamics support downstream return optimization in imagination.

We consider partially observable control from trajectories $\tau = \{o_t, a_t, r_t\}_{t=0}^{T}$, where o_t is the observation, a_t the action, and r_t the reward. The control objective is to maximize the discounted return

$$J(\pi) = \mathbb{E}_{p(\tau)}\left[\sum_{t=0}^{T} \gamma^t r_t\right]. \tag{1}$$

Following Dreamer-style latent imagination [7], RSSM represents the filtered belief at time t by $\hat{s}_t = (h_t, z_t)$, where h_t is a deterministic memory state summarizing observation history and z_t is a stochastic latent capturing aleatoric uncertainty.

Given actions $a_{0:T-1}$, RSSM learning seeks to maximize the conditional evidence

$$\max_{\theta,\phi} \log p_\theta(o_{0:T}, r_{0:T} \mid a_{0:T-1}), \tag{2}$$

where θ denotes the generative model parameters and ϕ the encoder parameters. The RSSM consists of a posterior encoder, a deterministic transition, a latent prior, an observation

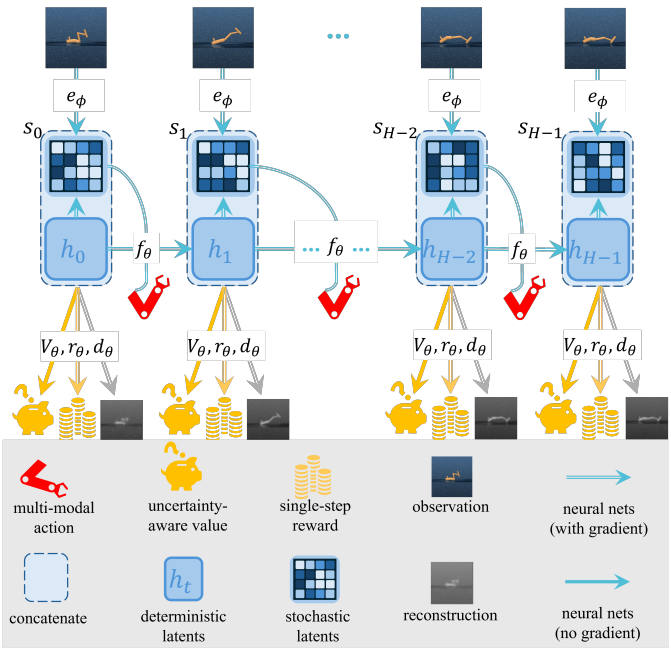


Fig. 1: **RSSM world model learning under partial observability.** An encoder infers stochastic latents z_t from observations conditioned on a recurrent memory state h_t, which is updated by a deterministic transition given actions. The learned prior predicts future latents, while decoders reconstruct observations and rewards, yielding a compact belief state $\hat{s}_t = (h_t, z_t)$ for latent imagination and downstream planning.

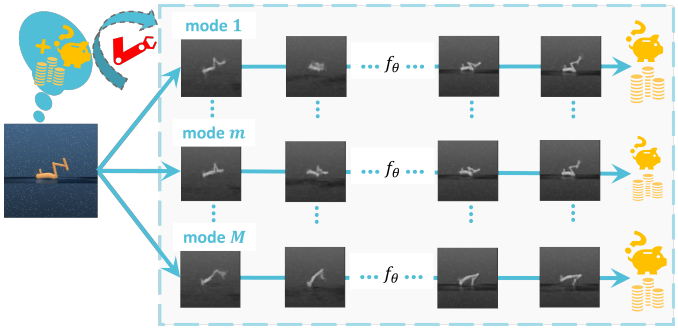


Fig. 2: **GMM-based long-horizon MPPI for multimodal trajectory distributions.** Over long horizons, sampled rollouts diverge and form multiple distinct high-reward modes. We fit a Gaussian Mixture Model (GMM) to the trajectory (or action-sequence) samples to capture this multimodality, then perform MPPI-style weighting and control extraction *per mode* before aggregating into a single action. This reduces mode collapse compared to a single-Gaussian update and improves robustness under partial observability.

decoder, and a reward model:

$$\begin{aligned} z_t &\sim e_\phi(z_t \mid h_t, o_t), \\ h_{t+1} &= f_\theta(h_t, z_t, a_t) = F_\theta(z_{:t}, a_{:t}) = F_\theta^{:t}, \\ z_{t+1} &\sim p_\theta(z_{t+1} \mid h_{t+1}), \\ o_t &\sim d_\theta(o_t \mid h_t, z_t), \\ r_t &\sim r_\theta(r_t \mid h_t, z_t). \end{aligned} \tag{3}$$

To incorporate the deterministic memory update into the probabilistic model, we write it as a Dirac transition. The resulting joint model is

$$\begin{aligned} &p_\theta(o_{0:T}, r_{0:T}, z_{0:T}, h_{0:T} \mid a_{0:T-1}) \\ &= p(h_0)\, p_\theta(z_0 \mid h_0)\, d_\theta(o_0 \mid h_0, z_0)\, r_\theta(r_0 \mid h_0, z_0) \\ &\times \prod_{t=0}^{T-1} \delta\big(h_{t+1} - f_\theta(h_t, z_t, a_t)\big)\, p_\theta(z_{t+1} \mid h_{t+1}) \\ &\times\ d_\theta(o_{t+1} \mid h_{t+1}, z_{t+1})\, r_\theta(r_{t+1} \mid h_{t+1}, z_{t+1}), \end{aligned} \tag{4}$$

while the filtering posterior is

$$\begin{aligned} &q_\phi(z_{0:T}, h_{0:T} \mid o_{0:T}, a_{0:T-1}) = p(h_0)\, e_\phi(z_0 \mid h_0, o_0) \\ &\times \prod_{t=0}^{T-1} \delta\big(h_{t+1} - f_\theta(h_t, z_t, a_t)\big)\, e_\phi(z_{t+1} \mid h_{t+1}, o_{t+1}). \end{aligned} \tag{5}$$

Applying the standard variational identity gives

$$\begin{aligned} &\log p_\theta(o_{0:T}, r_{0:T} \mid a_{0:T-1}) \\ &= \log \iint q_\phi \frac{p_\theta(o_{0:T}, r_{0:T}, z_{0:T}, h_{0:T} \mid a_{0:T-1})}{q_\phi(z_{0:T}, h_{0:T} \mid o_{0:T}, a_{0:T-1})}\, \mathrm{d}z_{0:H}\, \mathrm{d}h_{0:H} \\ &\geq \mathbb{E}_{q_\phi}[\log p_\theta(o_{0:T}, r_{0:T}, z_{0:T}, h_{0:T} \mid a_{0:T-1}) - \log q_\phi] \\ &=: \mathrm{ELBO}_{\mathrm{RSSM}}. \end{aligned} \tag{6}$$

Substituting (4) and (5), the Dirac terms cancel, yielding

$$\begin{aligned} &\mathrm{ELBO}_{\mathrm{RSSM}} = \mathbb{E}_{q_\phi}\Bigg[\sum_{t=0}^{T} \log d_\theta(o_t \mid h_t, z_t) \\ &+ \log r_\theta(r_t \mid h_t, z_t) + \log p_\theta(z_t \mid h_t) - \log e_\phi(z_t \mid h_t, o_t)\Bigg] \\ &= \sum_{t=0}^{T} \mathbb{E}_{q_\phi}\Big[\log d_\theta(o_t \mid F_\theta^{:t-1}, z_t) + \log r_\theta(r_t \mid F_\theta^{:t-1}, z_t)\Big] \\ &- \sum_{t=0}^{T} \Big[\mathrm{KL}\big(e_\phi(\cdot \mid F_\theta^{:t-1}, o_t) \,\|\, p_\theta(\cdot \mid F_\theta^{:t-1})\big)\Big]. \end{aligned} \tag{7}$$

Thus, RSSM learning maximizes $\mathrm{ELBO}_{\mathrm{RSSM}}$, which jointly encourages accurate observation reconstruction, reward prediction, and consistency between the filtered posterior and the latent prior.

After training, the posterior encoder maps the current history to a filtered latent belief $\hat{s}_t = (h_t, z_t)$, which serves as the starting state for imagination. Future latent rollouts then replace posterior updates with prior transitions under the learned RSSM, and predicted rewards are accumulated to approximate the return in Eq. 1. Dreamer-style methods use these imagined trajectories for actor-critic updates. In this paper, we focus on further return maximization under latent imagination using long-horizon MPC.

IV. METHODS

Our goal is long-horizon visual control under partial occlusion. Starting from the RSSM belief state $\hat{s}_t = (h_t, z_t)$, we make long-horizon MPC practical by addressing two coupled failure modes: (i) *trajectory branching* over long imagination horizons, which makes unimodal MPPI updates collapse; and (ii) *compounding epistemic uncertainty* from learned dynamics and missing observations, which makes deep rollouts unreliable unless planning and learning know when to truncate. We therefore introduce (1) **GMM-MPPI** to represent multi-modal long-horizon trajectory distributions, and (2) a **shared uncertainty-aware return** (UCB-gated λ_t) used for both imagined actor–critic learning and MPPI trajectory scoring, enabling an uncertainty-driven trade between short-horizon bootstrapping and long-horizon rollout returns (*soft truncation*) that is particularly effective under occlusions.

A. GMM-MPPI for Multi-Modal Long-Horizon Planning

At each environment step, we optimize an H-length action sequence in latent space using MPPI conditioned on the current belief $\hat{s}_k = (h_k, z_k)$. Standard MPPI maintains a unimodal (typically Gaussian) proposal over action sequences and updates its mean after reweighting samples by their predicted return. Prior stochastic-planning / MPC work has shown that optimal trajectory or action distributions can be multimodal, making unimodal Gaussian proposals limiting in some settings [18, 10]. In long-horizon latent planning under partial observability, this issue becomes especially acute: branching imagined futures can cause initially similar action sequences to lead to increasingly separated future states, while several such futures remain competitive in return. As illustrated in Fig. 2, a single-Gaussian proposal then becomes a poor approximation of the high-return distribution, averaging across incompatible hypotheses and producing conservative "mean" sequences that do not correspond to any coherent future. This motivates richer proposal families for planning.

To preserve multiple plausible long-term hypotheses, we maintain M *parallel Gaussian proposals* (modes) over action sequences:

$$q_m(a_{0:H-1}) \;=\; \mathcal{N}\big(a_{0:H-1}; \mu_m, \Sigma_m\big), \qquad m = 1, \dots, M, \tag{8}$$

where $\mu_m = \{\mu_{m,t}\}_{t=0}^{H-1}$ and $\Sigma_m = \{\Sigma_{m,t}\}_{t=0}^{H-1}$ are time-indexed moments (often diagonal). For each mode m, we draw K candidates, $\{a_{0:H-1}^{(k)}\}_{k=1}^{K} \sim q_m(\cdot)$, roll each sequence out under the learned RSSM latent dynamics starting from $\hat{s}_k$, and assign a scalar score $G_m^{(k)}$ using the uncertainty-aware return in Sec. IV-B (Eq. 17). To normalize scores consistently across modes, we compute the best score over *all* sampled rollouts,

$$G^\star \;=\; \max_{m \in \{1,\dots,M\},\, k \in \{1,\dots,K\}} G_m^{(k)}. \tag{9}$$

We then compute *mode-wise* weights using a relative-to-best softmax (normalized *within* each mode),

$$w_m^{(k)} \;=\; \frac{\exp\!\Big(\frac{1}{\tau}\,\frac{G_m^{(k)}}{G^\star + \delta}\Big)}{\sum_{j=1}^{K} \exp\!\Big(\frac{1}{\tau}\,\frac{G_m^{(j)}}{G^\star + \delta}\Big)}, \qquad k = 1, \dots, K, \tag{10}$$

where $\tau > 0$ is a temperature and $\delta > 0$ is a small constant for numerical stability. Each mode is updated by weighted moment matching:

$$\mu_{m,t} \leftarrow \sum_{k=1}^{K} w_m^{(k)}\, a_t^{(k)}, \qquad t = 0, \dots, H-1, \tag{11}$$

$$\Sigma_{m,t} \leftarrow \sum_{k=1}^{K} w_m^{(k)} \big(a_t^{(k)} - \mu_{m,t}\big)\big(a_t^{(k)} - \mu_{m,t}\big)^\top \;+\; \epsilon I, \tag{12}$$

with a small regularizer ϵI. Unlike a unimodal proposal, these parallel updates preserve multiple coherent hypotheses under branching latent trajectories and stabilize long-horizon optimization.

a) Policy–random initialization and warm-start.: We warm-start planning across time by shifting each mean sequence μ_m forward (receding horizon). To diversify hypotheses within a planning call, we initialize each mode m using a distinct policy–random mixing ratio $\alpha_m \in [0, 1]$:

$$\mu_m \leftarrow \alpha_m a_{0:H-1}^{\pi} + (1 - \alpha_m) a_{0:H-1}^{\mathrm{rand}}, \tag{13}$$

where $a_{0:H-1}^{\pi} \sim \pi_\psi(\cdot \mid \hat{s}_k)$ is a proposal from the learned actor prior and $a_{0:H-1}^{\mathrm{rand}}$ is a random sequence (e.g., Gaussian noise within action bounds). After L MPPI iterations, we select the highest-scoring rollout across modes (Eq. 9) and execute the first action of its corresponding mean sequence.

B. Uncertainty-aware learning and planning via UCB-gated λ_t

Long-horizon imagination is only useful if we can control compounding model error. This is particularly critical for *visual occlusions*: when some observations are missing, the belief must be propagated by the RSSM prior, increasing epistemic uncertainty and making deep rollouts unreliable unless the controller can decide *where the effective horizon should end*. We address this with a shared uncertainty-aware return used consistently for: (i) learning an imagined actor–critic prior and (ii) scoring MPPI rollouts.

a) Ensemble-UCB uncertainty: We maintain an ensemble of latent critics $\{V_i(\hat{s})\}_{i=1}^{M}$ and compute, along imagined rollouts,

$$\begin{aligned} \mu_t &:= \frac{1}{M} \sum_{i=1}^{M} V_i(\hat{s}_t), \\ \sigma_t &:= \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} \big(V_i(\hat{s}_t) - \mu_t\big)^2}. \end{aligned} \tag{14}$$

These define an optimism-flavored importance score, where β is used to deliver the exploration-exploitation tradeoff.

$$\mathrm{UCB}(\hat{s}_t) \;:=\; \mu_t + \beta\, \sigma_t. \tag{15}$$

UCB is high when a predicted future is both *valuable* (large mean) and *epistemically informative* (large dispersion), which is precisely the regime we wish to reach under occlusion-driven ambiguity via long-horizon exploration.

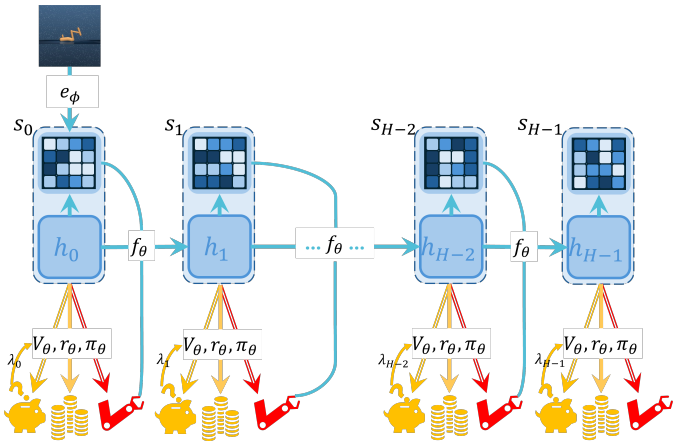


Fig. 3: **Imaginary TD learning with UCB-gated λ-returns.** We train actor–critic priors from RSSM-imagined rollouts using an ensemble-UCB score to set a time-varying λ_t in the λ-return targets. High-UCB states induce smaller λ_t (greater bootstrapping), while low-UCB states induce larger λ_t (deeper look-ahead), yielding stable yet exploratory value and policy learning for MPPI warm-starting.

b) Soft truncation with time-varying λ_t: Our uncertainty-based soft truncation is built upon standard λ return, except that we map UCB to a time-dependent λ_t to determine how far return information should propagate through imagination:

$$\lambda_t = \lambda_{\max} - (\lambda_{\max} - \lambda_{\min})\,\mathrm{norm}\big(\mathrm{UCB}(\hat{s}_t)\big), \quad \text{s.t.} \quad 0 \le \lambda_{\min} \le \lambda_{\max} \le 1, \tag{16}$$

so that *high UCB* induces *smaller* λ_t (more bootstrapping, shorter effective horizon), while *low UCB* induces *larger* λ_t (deeper look-ahead). This yields an uncertainty-triggered soft truncation: once a high-UCB future is reached, we reduce reliance on deeper (noisier) imagination; if the rollout remains uninformative, we keep propagating rewards further to search for better futures. Here $\mathrm{norm}(\cdot) \in [0,1]$ denotes a running-statistics normalization (e.g., exponential moving mean/variance with clipping) computed over recent imagined rollouts, following standard stabilization practices used in world-model and latent-MPC agents [7, 8].

c) Shared return for imagined TD learning and MPPI scoring: We define the time-varying λ-return recursively as

$$\begin{aligned} G_t &= \hat{r}_t + \gamma\Big((1-\lambda_t)\,\mu_{t+1} + \lambda_t\, G_{t+1}\Big), \\ G_{H-1} &= \mu_{H-1}, \end{aligned} \tag{17}$$

where $\hat{r}_t$ is the RSSM reward prediction. Crucially, we use the *same* UCB-gated return for *both* learning and planning: G_0 serves as the TD target for training the critic ensemble and an actor prior $\pi_\theta(a \mid \hat{s})$ as shown in Fig. 3. The learned actor then warm-starts MPPI by initializing the action-sequence mean (and/or injecting policy-sampled candidates), which improves sample efficiency and robustness under partial occlusions. This design explicitly aligns with the view that policy learning and MPC should collaborate: RL provides a strong prior that guides sampling-based planning, while MPC provides an online improvement operator that continually refines the policy's proposals [22].

We summarize our method, ELVIS, in Alg. 1.

Algorithm 1 ELVIS: Online learning and inference with GMM-MPPI and UCB-gated λ_t-returns

Initialize RSSM world model and posterior belief update (Dreamer-style) [7]
Initialize actor prior π_θ, critic ensemble $\{V_i\}_{i=1}^{E}$, replay buffer $\mathcal{D}$
Initialize per-mode moments $\{\mu_m, \Sigma_m\}_{m=1}^{M}$

Function PLAN($\hat{s}_k$)
  Warm-start all modes via receding-horizon shift
  Initialize each mode via Eq. 13
  for $\ell = 1$ **to** L **do**
    for $m = 1$ **to** M **do**
      Sample K sequences from Eq. 8
      Score rollouts using Eqs. 14–17
    end for
    Compute global best $G^\star$ via Eq. 9
    for $m = 1$ **to** M **do**
      Compute weights via Eq. 10
      Update (μ_m, Σ_m) via Eqs. 11–12
    end for
  end for
  Return first action of the best mode (Eq. 9)

for environment steps $k = 1, 2, \ldots$ **do**
  Update belief $\hat{s}_k$ using RSSM posterior [7]
  $a_k \leftarrow$ PLAN($\hat{s}_k$); execute a_k; store transition in $\mathcal{D}$
  Update RSSM (standard world-model learning) [7]
  Update critics and actor prior using shared return (Eq. 17)
end for

V. EXPERIMENTS

We structure experiments to answer two primary questions. **Q1: Does long-horizon, uncertainty-aware MPC improve performance on visual control?** To address this, we evaluate **TD-MPC2** [8], **DreamerV3** [7], and **ELVIS** on 14 DeepMind Control Suite (DMC) vision tasks. This comparison isolates complementary strengths: TD-MPC2 provides a strong online planner based on frame stacking instead of explicit recurrent memory, whereas DreamerV3 provides recurrent world-model learning without online planning. We follow standard protocols (10 evaluation episodes every 100k environment steps, 5 seeds).

We answer this in two steps. First, we report results under standard DMC visual control tasks, directly testing whether long-horizon planning and uncertainty awareness are effective for visual control in Sec. V-A. Second, we perform ablations on DMC to isolate the roles of *planning horizon* and *uncertainty-aware truncation* by sweeping horizons and removing uncertainty mechanisms (fixed-λ/no-UCB variants), quantifying how long-horizon gains depend on uncertainty control in Sec. V-A.

Q2: Under extreme partial occlusions in the real world, which components are essential for robust per-

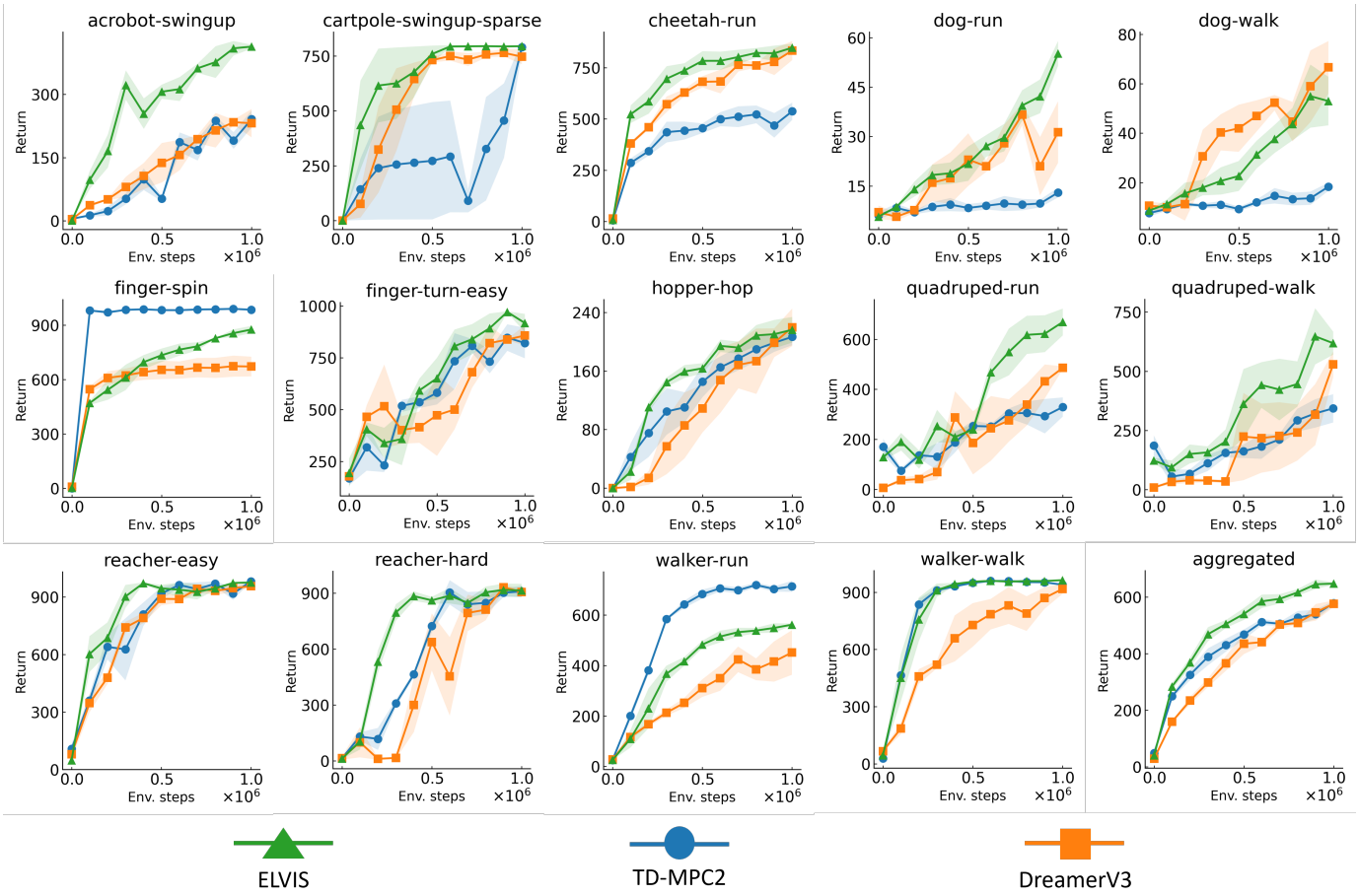


Fig. 4: **DMC visual control learning curves.** Per-task learning curves on 14 DeepMind Control (DMC) visual control benchmarks, together with an aggregated score that reports the mean episodic return averaged across all 14 tasks at each environment step. Shaded regions denote 95% confidence intervals over 5 random seeds. ELVIS achieves the strongest overall performance, ranking first or second on every task.

formance and sim-to-real transfer? To address this question, we evaluate **zero-shot sim-to-real transfer** on a real-world sand-spraying task featuring extreme occlusions and sensory noise [3]. This setting emphasizes the necessity of *explicit memory* and tests whether *uncertainty-aware planning* improves robustness beyond simulation. We use the surface-quality metrics and protocol of [3] and compare against the same baselines in Sec. V-B).

A. Visual Control on DMC

Across 14 visual control learning tasks, **ELVIS establishes state-of-the-art performance**, improving both sample efficiency and final return relative to the baselines in the majority of environments, as shown in Fig. 4. Notably, ELVIS is consistently competitive: it ranks *first or second* across tasks and is never the lowest-performing method among the three. This pattern is important because ELVIS was motivated by the need to handle more challenging visual control settings with severe partial observability and occlusions (Section V-B); yet the results here show that its design choices also translate to *standard* visual control benchmarks. In other words, the components introduced for robustness under partially observable settings, uncertainty awareness and long-horizon model-based planning, do not trade off performance in easier settings, but instead improve data efficiency and stabilize performance broadly. This suggests a promising direction for improving model-based RL: explicitly accounting for predictive uncertainty while planning over longer horizons can yield benefits even when observations are complete.

To isolate the key drivers of performance, we conduct controlled ablations on the same DMC suite and report *aggregated* performance (mean return averaged across the 14 tasks at each evaluation step) in Fig. 5. First, varying the MPC horizon shows that **long-horizon planning is crucial**: $H = 15$ **substantially outperforms** $H = 5$, indicating that deeper foresight is beneficial when paired with a learned world model. Second, replacing our uncertainty-based λ-return truncation with a non-uncertainty variant (e.g., fixed-λ / no-UCB) causes a **large drop** in aggregated score, showing that **uncertainty awareness is necessary to control compounding model error**. Third, replacing the proposed GMM action proposal with a **unimodal Gaussian** (**No GMM**) also lowers

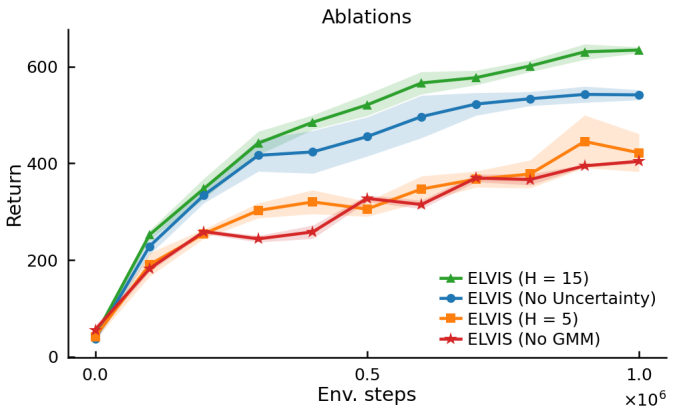
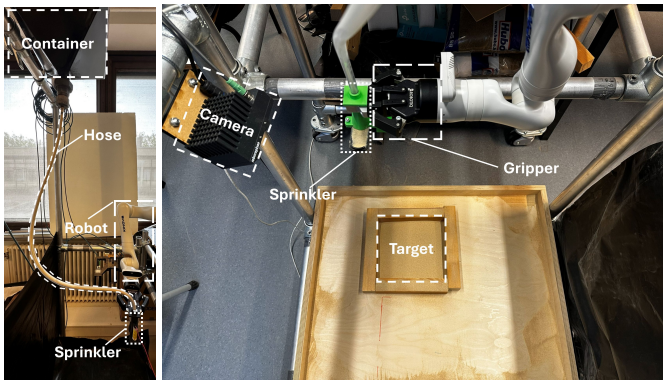


Fig. 5: **Ablations of ELVIS.** Aggregated learning curves on the same 14 DMC visual control tasks, where the score is the mean episodic return averaged across tasks at each environment step. Shaded regions denote 95% confidence intervals over 5 random seeds. Removing either GMM, uncertainty awareness or long-horizon planning degrades sample efficiency and final performance, indicating that all three components contribute and interact to produce ELVIS's gains.



(a) Side view (b) Top view

Fig. 6: Sand-spraying testbed used for zero-shot sim-to-real transfer. (a) Side view: sand stored in an overhead container is transmitted through a hose to a sprinkler positioned over the target, generating plumes that induce visual occlusion. (b) Top view: the sprinkler is held by a robotic arm and guided using heightmaps of the target derived from an overhead stereo camera.

performance, supporting our claim that long-horizon latent planning benefits from **multimodal action proposals** rather than a single Gaussian.

B. Sand Spray Sim-to-Real Zero Shot Transfer

We evaluate zero-shot sim-to-real transfer on a sand-spraying task that serves as a controlled proxy for real-world industrial operations in harsh, low-visibility settings, where airborne material and sensor noise induce severe occlusions. Following the experimental rationale of AREPO [3], a nozzle deposits granular material onto a target surface, and the control

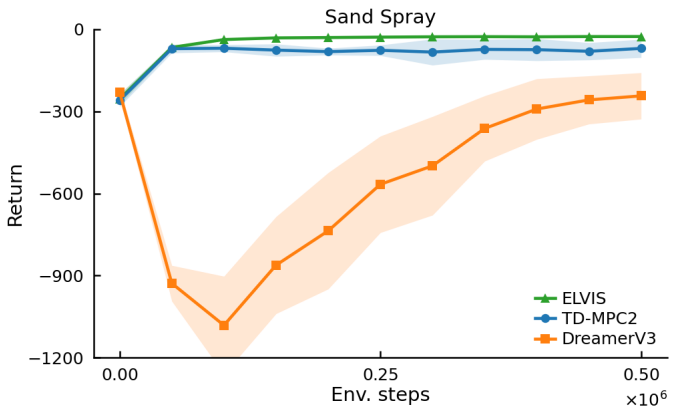


Fig. 7: **Learning curves of simulated sand spray task.** Scores of ELVIS, TD-MPC2, and DreamerV3 on the simulated sand spray task, which shows the mean episodic return at each environment step. Shaded regions denote 95% confidence intervals across 5 seeds. ELVIS significantly outperforms DreamerV3 while exhibiting more stable learning and reduced sensitivity to random seeds compared to TD-MPC2.

objective is to produce a uniform deposition profile despite partial observability and disturbances. Our simulator is a custom Gymnasium environment whose deposition dynamics are hand-designed from domain knowledge as detailed in AREPO, rather than built on top of a separate physics engine. In simulation, the agent observes simulated heightmaps and outputs planar nozzle-motion commands; after training, the learned policy is transferred without adaptation to the physical setup, where the robot is guided by real heightmaps reconstructed from an overhead stereo camera. Transfer does not rely on photorealistic rendering: the sim-to-real interface is primarily geometric through heightmaps, although a substantial gap remains due to dust, occlusion, sensor noise, and mismatch between simulated and real deposition dynamics. These factors reduce world-model prediction reliability, especially when observations are missing or corrupted over long-horizon rollouts. A detailed layout of the laboratory setup is shown in Fig. 6a and Fig. 6b.Performance is assessed using the surface-quality metrics defined in [3]. Specifically, we report three surface-quality metrics: i) *root-mean-square roughness* R_{rms}: the square root of the mean of the squares of the deviations of the surface height values from the mean surface height, ii) *peak-to-valley roughness* R_t: the difference in height between the highest point and the lowest point on a surface, iii) *waste volume ratio* r_{wv}: the ratio between the wasted volume and the desired volume to be fulfilled. The wasted volume is defined as the material volume that has been sprayed outside the target surface or that exceeds the target thickness.

a) Simulation learning performance: Fig. 7 shows that model-based planning is a decisive advantage on this simulated sand spraying task: ELVIS and TD-MPC2 consistently outperform DreamerV3, which learns a latent world model but does not perform online MPC-style trajectory optimization at control time. The gap indicates that, in challenging visual

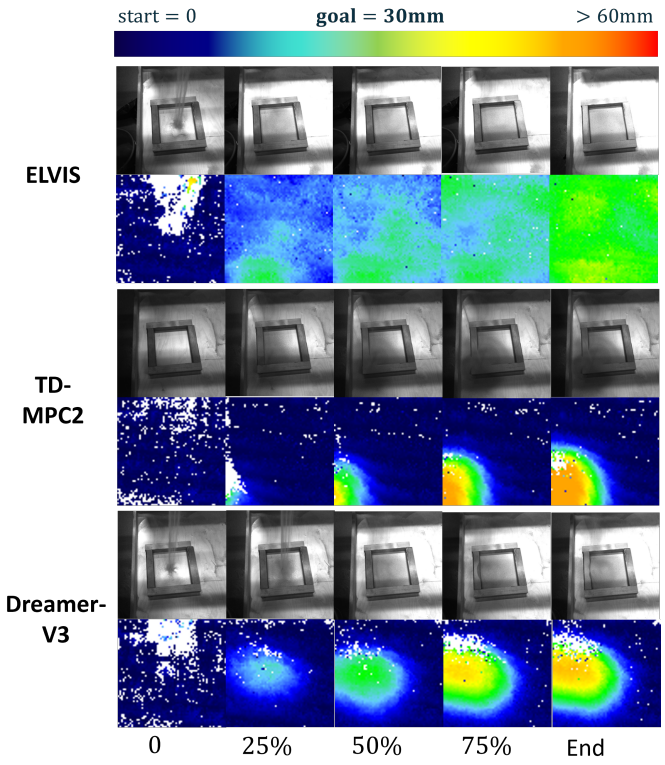


Fig. 8: **Zero-shot evaluation with real-world sand spray task.** For each method, the first row shows grayscale scene images for visualization only, while the policy itself acts on the corresponding heightmaps shown in the second row. In the sim-to-real experiment, ELVIS is most robust to partial observability caused by dust and sensory noise (unobservable white parts of the heightmaps) and achieves the best surface quality compared to other methods.

control under extreme visual occlusions, explicitly optimizing action sequences through a predictive world model yields substantially stronger closed-loop behavior than purely policy-based action selection.

Beyond the benefits of longer-horizon planning, ELVIS offers a measurable advantage over TD-MPC2 in terms of learning robustness, as it exhibits more consistent and stable training with reduced sensitivity to random seeds. This improved stability is aligned with the DMC results and supports the claim that uncertainty awareness, together with an effective extension of the planning horizon, can improve data efficiency and make model-based control more reliable.

b) Real-world evaluation: We transfer the trained policies to the laboratory sand-spraying testbed without adaptation and evaluate surface quality using the same metrics and protocol as [3]. Each method is evaluated over 5 real-world hardware trials. Table I reveals a marked change in relative ranking from simulation to deployment: during simulation training we observe ELVIS $>$ TD-MPC2 $\gg$ DreamerV3, whereas under zero-shot sim-to-real transfer the ordering becomes ELVIS $\gg$ DreamerV3 $>$ TD-MPC2. This reversal

TABLE I: In zero-shot sim-to-real transfer to the laboratory testbed, ELVIS achieves the best surface quality compared to baselines without uncertainty awareness.

Metrics	R_{rms} (mm)	R_t (mm)	r_{wv} (%)
ELVIS	**2.2 $\pm$ 0.4**	**17.5 $\pm$ 0.5**	**6.3 $\pm$ 1.1**
TD-MPC2	16.3 $\pm$ 0.3	69.6 $\pm$ 2.1	47.2 $\pm$ 1.4
DreamerV3	9.3 $\pm$ 0.4	52.3 $\pm$ 1.1	25.1 $\pm$ 0.7

highlights the importance of *recurrent belief tracking* under severe real-world occlusion and sensor noise: methods that rely on naive frame stacking can struggle to maintain coherent control when observations become intermittently unreliable. Moreover, ELVIS's advantage over DreamerV3 suggests that *recurrent memory alone is not sufficient*: coupling belief-based control with *uncertainty awareness* helps limit compounding model errors during long-horizon decision making and improves sim-to-real robustness.

These quantitative trends are reflected in qualitative behavior in Fig. 8. ELVIS produces the most uniform sand deposition throughout the entire spraying interval, yielding homogeneous coverage even under severe occlusions. In contrast, TD-MPC2 degrades sharply after transfer and collapses to repeatedly spraying a single corner, consistent with brittle planning under partial observability and unmodeled real-world perturbations. DreamerV3 transfers more reliably than TD-MPC2, supporting the benefit of recurrent memory, but its lack of online MPC refinement and explicit uncertainty handling limits its ability to recover from accumulated errors and to maintain consistent coverage. Together, these results demonstrate that ELVIS's combination of long-horizon planning, recurrent belief, and uncertainty-aware truncation is essential for robust zero-shot real-world performance beyond simulation.

VI. Conclusion

We presented ELVIS, a memory-based, long-horizon, uncertainty-aware latent visual MPC framework for control under partial observability. Built on RSSM belief states, ELVIS combines (i) GMM-MPPI to preserve multi-modal long-horizon hypotheses and avoid mode collapse, and (ii) an ensemble-UCB-gated, time-varying λ_t that softly truncates return propagation and links imagined actor-critic learning with online planning. Across standard DMControl vision benchmarks, ELVIS improves performance, and ablations support the roles of GMM-based long-horizon planning and uncertainty-aware truncation. ELVIS also achieves strong zero-shot sim-to-real transfer on a real-world sand-spraying task with severe visual occlusions, demonstrating robust deployment beyond simulation. The current study is limited to dense-reward continuous-control settings and incurs added compute cost from long-horizon latent MPC rollouts. Future work will study more efficient long-horizon world models and rollout-shortcut mechanisms, extend evaluation to sparse-reward manipulation, and develop tighter theory for mixture-based planning under model uncertainty.

References

[1] Alberto Bertipaglia, Dariu M. Gavrila, and Barys Shyrokau. Multi-Modal Model Predictive Path Integral Control for Collision Avoidance, 2025.

[2] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*. Curran Associates, Inc., 2018.

[3] Yurui Du, Louis Hanut, Herman Bruyninckx, and Renaud Detry. AREPO: Uncertainty-Aware Robot Ensemble Learning Under Extreme Partial Observability. *IEEE Robotics and Automation Letters*, 10(6), 2025.

[4] Bernd Frauenknecht, Artur Eisele, Devdutt Subhasish, Friedrich Solowjow, and Sebastian Trimpe. Trust the Model Where It Trusts Itself: Model-Based Actor-Critic with Uncertainty-Aware Rollout Adaption. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

[5] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565. PMLR, 2019.

[6] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

[7] Danijar Hafner et al. Mastering Diverse Control Tasks through World Models. *Nature*, 2025. DreamerV3 journal version.

[8] Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, Robust World Models for Continuous Control. In *International Conference on Learning Representations (ICLR)*, 2024. Spotlight.

[9] Nicklas A. Hansen, Hao Su, and Xiaolong Wang. Temporal Difference Learning for Model Predictive Control. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 8387–8406. PMLR, 2022.

[10] Kohei Honda, Naoki Akai, Kosuke Suzuki, Mizuho Aoki, Hirotaka Hosogaya, Hiroyuki Okuda, and Tatsuya Suzuki. Stein Variational Guided Model Predictive Path Integral Control: Proposal and Experiments with Fast Maneuvering Vehicles. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7020–7026, 2024.

[11] Kayalibay, Baris and Mirchev, Atanas and Agha, Ahmed and van der Smagt, Patrick and Bayer, Justin. Filter-aware model-predictive control. In *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pages 1441–1454, 2023.

[12] Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep Variational Reinforcement Learning for POMDPs. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2117–2126. PMLR, 2018.

[13] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to Trust Your Model: Model-Based Policy Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[14] Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 6131–6141. PMLR, 2021.

[15] Kendall Lowrey, Aravind Rajeswaran, Sham Kakade, Emanuel Todorov, and Igor Mordatch. Plan Online, Learn Offline: Efficient Learning and Exploration via Model-Based Control, 2018.

[16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level Control through Deep Reinforcement Learning. *Nature*, 518(7540):529–533, 2015.

[17] Anusha Nagabandi, Gregory Kahn, Ronald S. Fearing, and Sergey Levine. Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7559–7566. IEEE, 2018.

[18] Masashi Okada and Tadahiro Taniguchi. Variational Inference MPC for Bayesian Model-based Reinforcement Learning. In *Proceedings of the 3rd Conference on Robot Learning (CoRL)*, volume 100 of *Proceedings of Machine Learning Research*, pages 258–272. PMLR, 2020.

[19] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to Explore via Self-Supervised World Models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

[20] Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware Latent Safety Filters for Avoiding Out-of-Distribution Failures, 2025.

[21] Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The Distracting Control Suite – A Challenging Benchmark for Reinforcement Learning from Pixels. *arXiv preprint arXiv:2101.02722*, 2021.

[22] Yuhang Wang, Hanwei Guo, Sizhe Wang, Long Qian, and Xuguang Lan. Bootstrapped Model Predictive Control. In *International Conference on Learning Representations (ICLR)*, 2025.

[23] Grady Williams, Paul Drews, Brian Goldfain, James M.

Rehg, and Evangelos A. Theodorou. Aggressive Driving with Model Predictive Path Integral Control. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1433–1440. IEEE, 2016.

[24] Yuezhe Zhang, Corrado Pezzato, Elia Trevisan, Chadi Salmi, Carlos Hernandez Corbato, and Javier Alonso-Mora. Multi-Modal MPPI and Active Inference for Reactive Task and Motion Planning. *IEEE Robotics and Automation Letters*, 9(9):7461–7468, 2024.