BRAIN-INSPIRED PRIOR REGULATION FOR EFFICIENT PLANNING WITH WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in world models enable predictive reasoning about action consequences, supporting goal-directed decision-making in robotic manipulation and visual navigation. However, action search remains computationally expensive, and existing prior-guided methods improve sampling efficiency through initialization but lack persistent regulation of subsequent optimization, leaving search vulnerable to deviations from the learned action prior. To address this limitation, we propose Brain-Inspired Prior Regulation (BIPR), a framework for efficient planning with world models through coordinated prior regulation across multiple timescales. Specifically, prior anchoring penalizes deviations from the action prior throughout search, uncertainty–conflict arbitration adapts regularization to current search evidence, and metaplastic calibration uses offline learning history to modulate this adaptive response. Experiments on established benchmarks for world-model-based control show that BIPR improves macro-average success rates with LeWorldModel and Sub-JEPA by 30.33 and 26.33 percentage points, respectively, over the MPPI baseline, while complementary efficiency evaluations demonstrate a 46.73% reduction in evaluation wall time relative to CEM under a matched candidate budget. Beyond task success, computational cost analysis shows that BIPR requires no additional trainable parameters compared with prior-guided planning, while incurring only 0.82% additional evaluation overhead. These findings suggest that brain-inspired prior regulation offers a promising approach to improving the performance–efficiency trade-off of world-model-based planning.

1 INTRODUCTION

Visual world models turn observations and candidate actions into predictions of future states, enabling an agent to compare possible behaviors before acting. Planning in learned latent spaces is particularly attractive when pixel reconstruction is unnecessary: the same predictive representation can support different visual goals without training a separate policy for each goal. Recent joint-embedding world models demonstrate this capability with compact end-to-end architectures and pretrained visual representations (Maes et al., 2026b; Zhao et al., 2026; Zhou et al., 2025). These advances offer a path toward reusable predictive representations, connecting visual understanding with goal-directed behavior in robotic manipulation and autonomous navigation.

LeWorldModel (LeWM), a recent end-to-end visual world model with competitive performance against state-of-the-art JEPA baselines, uses the cross-entropy method (CEM) to optimize action sequences through learned latent dynamics (Maes et al., 2026b). Iterative candidate evaluation can provide strong control performance, but repeatedly rolling out many sequences makes planning expensive. Improving the predictive backbone alone does not remove this search cost, while reducing the search budget lowers computation but cannot ensure that the retained candidates include useful actions. More efficient sampling and informative proposals are therefore important complements to representation learning (Pinneri et al., 2021; Williams et al., 2018), offering a way to improve planning efficiency by directing model evaluations toward promising action sequences.

To address this challenge, recent methods introduce learned action priors, subgoal-conditioned action generation, and latent planners to guide candidate proposals or decompose difficult searches into tractable decisions (Wang et al., 2026; Cheng et al., 2026; Masip et al., 2026). Among them, PRISM combines a lightweight Gaussian action prior with fixed-covariance model predictive path integral

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.

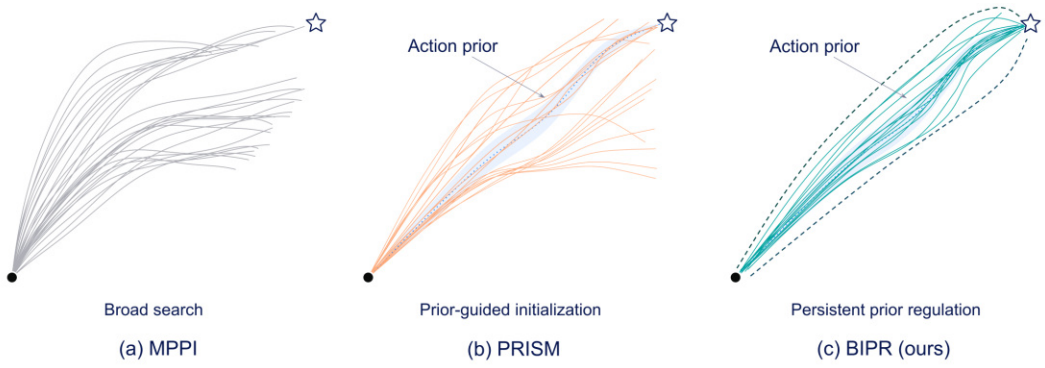

Figure 1: **Concentrating useful search without removing candidates.** MPPI without an action prior searches broadly; PRISM supplies a prior-guided start, but model-only scoring can favor off-prior directions. BIPR persistently penalizes displacement, concentrating high-weight candidates near the prior-typical region. PRISM and BIPR use the same K and fixed sampling covariance.

control (MPPI) instead of CEM's iterative covariance refitting, preserving state-dependent sampling precision and enabling effective planning with smaller sampling budgets (Wang et al., 2026). However, prior-guided MPPI that scores candidates solely by predicted goal cost can still move its search mean far from prior-supported behavior, allowing model-favored but unreliable actions to receive high weights, and limiting the benefit of informed initialization. Penalizing prior deviations throughout planning can counter this drift, yet a uniform penalty risks suppressing valid exploration. Large prior variance may reflect either insufficient learning or multiple reasonable actions, while model–prior disagreement alone cannot distinguish useful departures from prediction errors. This raises our research question: how can a planner preserve useful action priors while adapting their influence to uncertain search evidence, thereby improving control performance under a limited computational budget (Figure 1)?

To address this question, we propose Brain-Inspired Prior Regulation (BIPR), a framework for efficient planning with world models. Drawing on cognitive principles of habitual and goal-directed control, reliability-based arbitration, and history-dependent adaptation (Daw et al., 2005; Lee et al., 2014; Abraham & Bear, 1996), the framework separates three decisions: how far a candidate deviates from a learned action relation, when current search evidence warrants stronger regularization, and how offline learning history should temper that response. Specifically, Prior Anchoring translates the stabilizing influence of learned action habits into a persistent prior-relative penalty; Uncertainty–Conflict Arbitration draws on reliability-based arbitration to adjust this penalty using current uncertainty and model–prior conflict, while Metaplastic Calibration draws on metaplasticity to calibrate the additional response using held-out learning progress. Without online training, additional networks, or additional world-model rollouts, BIPR augments existing candidate scores with a variance-normalized action penalty and a history-calibrated scalar weight, addressing search drift and inflexible prior enforcement within the existing rollout budget.

Experiments on established benchmarks for world-model-based control show that BIPR improves macro-average success rates with LeWorldModel and Sub-JEPA by 30.33 and 26.33 percentage points, respectively, over the MPPI baseline, while complementary efficiency evaluations demonstrate a 46.73% reduction in evaluation wall time relative to CEM under a matched candidate budget. Beyond task success, computational cost analysis shows that BIPR requires no additional trainable parameters compared with prior-guided planning, while incurring only 0.82% additional evaluation overhead. These findings identify adaptive use of learned behavioral knowledge as a practical complement to predictive representation learning, offering a route toward more capable world-model-based control within the computational constraints of embodied systems.

2 Related Work

World Models. World models support control by predicting action consequences. PlaNet plans through latent dynamics (Hafner et al., 2019), and PETS accounts for predictive uncertainty with probabilistic ensembles (Chua et al., 2018). DreamerV3 learns behavior through imagination (Hafner et al., 2025), while TD-MPC and TD-MPC2 combine latent planning with value estimation (Hansen et al., 2022; 2024); iVideoGPT explores action-conditioned video prediction (Wu et al., 2024). These advances strengthen predictive control, but efficient action search remains a complementary challenge.

JEPA-Based World Models. I-JEPA and V-JEPA learn visual representations through embedding prediction (Assran et al., 2023; Bardes et al., 2024). DINO-WM adds predictive dynamics to pretrained DINOv2 features for visual-goal planning (Zhou et al., 2025; Oquab et al., 2024). LeWorldModel stabilizes end-to-end predictive learning with Gaussian regularization (Maes et al., 2026b), while Sub-JEPA applies this constraint in subspaces (Zhao et al., 2026). These methods improve predictive representations but leave action-prior guidance largely to the planner. BIPR addresses this gap through adaptive prior constraints on candidate evaluation.

Prior-Guided Planning with World Models. CEM and MPPI refine sampled action distributions (Rubinstein, 1999; Williams et al., 2018); iCEM and Biased-MPPI improve sampling through sample reuse and controller guidance, respectively (Pinneri et al., 2021; Trevisan & Alonso-Mora, 2024). Skill priors, diffusion models, and mean-flow policies provide learned behavioral structure (Pertsch et al., 2021; Janner et al., 2022; Chi et al., 2025; Zhan et al., 2026). For world-model planning, PRISM fuses a Gaussian action prior with MPPI's initial distribution (Wang et al., 2026), SAGE conditions action generation on latent subgoals (Cheng et al., 2026), and FF-JEPA decomposes long-horizon behavior using a latent planner (Masip et al., 2026). These approaches improve proposals or temporal decomposition, but prior-initialized MPPI still lacks an explicit penalty for subsequent search drift. BIPR extends prior guidance to candidate scoring throughout optimization.

3 Brain-Inspired Prior Regulation

In human cognition, habitual and goal-directed systems coordinate action selection through reliability-based arbitration (Daw et al., 2005; Lee et al., 2014), while hippocampal prospective sequences and hippocampal–prefrontal interactions support the imagination and evaluation of future trajectories (Pfeiffer & Foster, 2013; Jensen et al., 2024). Memory processes operating across timescales preserve experience that can shape subsequent decisions (Benna & Fusi, 2016). Together, these mechanisms balance learned behavioral tendencies with prospective evidence to regulate action. Inspired by this organization, we design BIPR as a sequential prior-regulation mechanism: Prior Anchoring measures each candidate's departure from learned behavior, Uncertainty–Conflict Arbitration adjusts the response to current uncertainty and model–prior disagreement, and Metaplastic Calibration uses offline learning history to calibrate that response. By linking these mechanisms through a shared candidate penalty and a single final scalar weight, the complete framework regulates prior influence throughout optimization and feeds the calibrated scores directly into the planning update.

3.1 Planning Setup

At environment time t, let o_t and o_g be the current and goal observations. An encoder E produces representations $z_t = E(o_t)$ and $z_g = E(o_g)$ in $\mathbb{R}^{d_z}$, where d_z is the latent dimension. A predictor F maps a candidate action sequence $\mathbf{a} = (\mathbf{a}_0, \ldots, \mathbf{a}_{H-1}) \in \mathbb{R}^D$ and z_t to the terminal prediction $\widehat{z}_H(\mathbf{a}; z_t)$. Here H is the planning horizon and D is the total number of normalized scalar action coordinates; $a[d]$ denotes coordinate d. Any observation history required by the predictor is fixed within a planning call and omitted from the notation. The goal cost is

$$C_{\mathrm{WM}}(\mathbf{a}) = \frac{1}{d_z} \left\| \widehat{z}_H(\mathbf{a}; z_t) - z_g \right\|_2^2 . \tag{1}$$

Backbones with structured latent representations retain their native cost aggregation. All action coordinates use the training data's normalization; the chosen actions are converted back to environment units before execution.

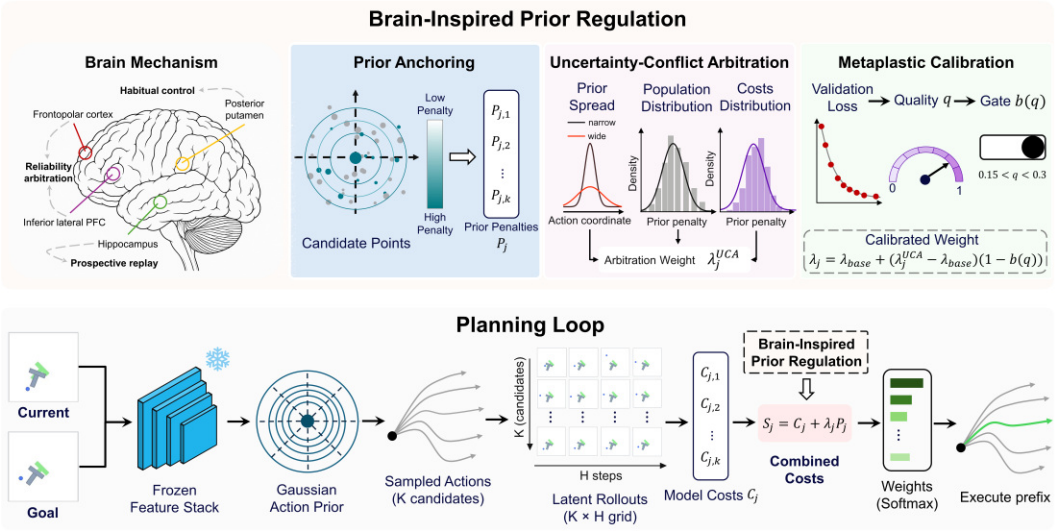

Figure 2: **Brain-inspired regulation from visual goals to action updates.** Prior Anchoring computes the penalty vector $\mathbf{p}_j$; Uncertainty–Conflict Arbitration combines prior spread and model-cost elite disagreement into λ_j^{UCA}; Metaplastic Calibration uses checkpoint progress q to obtain λ_j. With model-cost vector $\mathbf{c}_j$, the combined vector $\mathbf{s}_j = \mathbf{c}_j + \lambda_j \mathbf{p}_j$ yields weights and the next mean. The prefix is executed after rounds.

A prior head G predicts the mean $\boldsymbol{\mu}_p \in \mathbb{R}^D$ and coordinatewise standard deviation $\boldsymbol{\sigma}_p \in \mathbb{R}^D_{>0}$ of the Gaussian action prior $\mathcal{N}(\boldsymbol{\mu}_p, \mathrm{diag}(\boldsymbol{\sigma}_p^2))$. The head is trained offline with the shared prior-guided baseline's objective (Wang et al., 2026; Seitzer et al., 2022); the encoder E, world-model predictor F, and prior head G are all frozen.

Let $\boldsymbol{\mu}_b$ and $\boldsymbol{\sigma}_b$ be the base planner's initial mean and standard deviation, and define $\boldsymbol{\tau}_i = (\boldsymbol{\sigma}_i^2 + \epsilon_f)^{-1}$ for $i \in \{b, p\}$. The shared Gaussian fusion initializes the search mean $\boldsymbol{\mu}_0$ and fixed sampling standard deviation $\boldsymbol{\sigma}_s$:

$$\boldsymbol{\mu}_0 = \frac{\boldsymbol{\tau}_b \odot \boldsymbol{\mu}_b + \boldsymbol{\tau}_p \odot \boldsymbol{\mu}_p}{\boldsymbol{\tau}_b + \boldsymbol{\tau}_p}, \qquad \boldsymbol{\sigma}_s = \max\left\{(\boldsymbol{\tau}_b + \boldsymbol{\tau}_p)^{-1/2}, \sigma_{\min}\right\}. \tag{2}$$

Here $\epsilon_f = 10^{-8}$ stabilizes fusion, $\sigma_{\min} = 0.05$ floors the sampling standard deviation, and $\odot$ denotes elementwise multiplication; the remaining operations are coordinatewise. A call performs J optimization rounds with K candidates per round. We write $\mathbf{a}_j^{(k)}$ for candidate k in round j and $\boldsymbol{\mu}_j$ for the current search mean. The prior moments and $\boldsymbol{\sigma}_s$ remain fixed within the call.

3.2 Prior Anchoring

The coexistence of experience-based action tendencies and prospective evaluation reveals a principle of persistent but defeasible behavioral guidance (Daw et al., 2005). Inspired by this principle, we design Prior Anchoring so that learned actions continue to influence candidate selection after initialization, while predicted goal improvement can justify departures from the prior. The module limits search drift without removing candidate actions, thereby preserving exploratory alternatives while keeping optimization tied to learned behavioral structure.

Define the floored prior scale $\widetilde{\sigma}_p[d] = \max\{\sigma_p[d], \sigma_{\min}\}$. Prior Anchoring assigns each candidate a variance-normalized penalty and adds it to the model cost:

$$P(\mathbf{a}) = \min\left\{\frac{1}{D}\sum_{d=1}^{D}\left(\frac{a[d] - \mu_p[d]}{\widetilde{\sigma}_p[d]}\right)^2, P_{\max}\right\}, \qquad P_{\max} = 9,$$
$$s_{j,k} = c_{j,k} + \lambda_j p_{j,k}. \tag{3}$$

Here $c_{j,k} = C_{\mathrm{WM}}(\mathbf{a}_j^{(k)})$, $p_{j,k} = P(\mathbf{a}_j^{(k)})$, and $\lambda_j \geq 0$ is shared by all candidates in round j. Prior Anchoring alone uses the fixed baseline weight $\lambda_{\mathrm{base}} = 0.5$; Uncertainty–Conflict Arbitration and Metaplastic Calibration determine λ_j for the complete framework. Averaging over D avoids a penalty that grows merely with sequence length, while normalization permits larger departures along broader prior coordinates. The cap bounds the added cost but cannot distinguish candidates beyond saturation. Because the reference stays at $\boldsymbol{\mu}_p$ as $\boldsymbol{\mu}_j$ moves, the penalty persists throughout search.

3.3 Uncertainty–Conflict Arbitration

A prominent cognitive account of reliability arbitration proposes that control shifts between habitual and model-based systems according to the relative reliability of their predictions (Daw et al., 2005; Lee et al., 2014). This account motivates Uncertainty–Conflict Arbitration: disagreement and uncertainty should modulate the strength of behavioral guidance. In our implementation, broad prior scales can weaken Prior Anchoring's normalized distance, while model-preferred candidates can depart disproportionately from the prior; either condition therefore increases the additional penalty.

Let $\overline{\sigma}_p = D^{-1} \sum_d \widetilde{\sigma}_p[d]$ be the mean prior scale. To detect disagreement, let $\mathcal{E}_j$ contain the lowest-cost $\max\{1, \mathrm{round}(\alpha K)\}$ candidates, where $\alpha = 0.1$, and let $\overline{p}_j^E$ and $\overline{p}_j$ be the mean prior penalties within $\mathcal{E}_j$ and across all candidates. Spread, conflict, and their resulting arbitration weight are

$$\begin{aligned} u &= \left[1 + \exp(-k_u(\overline{\sigma}_p - \tau_u))\right]^{-1}, \\ v_j &= 1 - \exp\left[-k_c \max\left\{\frac{\overline{p}_j^E}{\max(\overline{p}_j, \epsilon_c)} - 1, 0\right\}\right], \\ \lambda_j^{\mathrm{UCA}} &= \lambda_{\mathrm{base}} + (\lambda_{\max} - \lambda_{\mathrm{base}})\big[1 - (1-u)(1-v_j)\big]. \end{aligned} \tag{4}$$

Here $(k_u, \tau_u, k_c) = (30, 0.9, 8)$, $\epsilon_c = 10^{-6}$, and $\lambda_{\max} = 1.0$. The spread signal u is fixed within a planning call, whereas v_j is recomputed each round. Selecting $\mathcal{E}_j$ by model cost before adding Prior Anchoring avoids circularity. Predicted action spread can reflect several uncertainty sources and is not itself a calibrated estimate of epistemic error. Thus, Uncertainty–Conflict Arbitration reuses the penalties from Prior Anchoring to identify conflict and returns a bounded increase above the persistent baseline. The next module calibrates whether to retain that increase.

3.4 Metaplastic Calibration

Neural plasticity supports rapid adaptation to changing experience, while memory retains information that can improve subsequent behavior (Abraham & Bear, 1996; Kirkwood et al., 1996; Benna & Fusi, 2016). We transfer this history-dependent regulation principle to planning through Metaplastic Calibration: offline learning progress changes the response to current arbitration evidence. When instantaneous prior spread alone is insufficient to determine the strength of additional regularization, the module uses learning history to calibrate that response.

Let $L_{\mathrm{init}} > 0$ and $L_{\mathrm{best}} \geq 0$ be the prior head's initial and best held-out action-prediction mean squared errors. The checkpoint progress, interval gate, and final weight are

$$\begin{aligned} q &= \min\left\{1, \max\left\{0, \frac{L_{\mathrm{init}} - L_{\mathrm{best}}}{L_{\mathrm{init}}}\right\}\right\}, \\ b(q) &= \mathbb{I}[\tau_{\mathrm{low}} < q < \tau_{\mathrm{high}}], \\ \lambda_j &= \lambda_{\mathrm{base}} + (\lambda_j^{\mathrm{UCA}} - \lambda_{\mathrm{base}})(1 - b(q)). \end{aligned} \tag{5}$$

The scalar q is fixed for each trained checkpoint, $(\tau_{\mathrm{low}}, \tau_{\mathrm{high}}) = (0.15, 0.30)$, and $\mathbb{I}[\cdot]$ is the indicator function. The binary gate $b(q)$ removes only the additional arbitration weight: when it is one, $\lambda_j = \lambda_{\mathrm{base}}$; otherwise, $\lambda_j = \lambda_j^{\mathrm{UCA}}$. If valid progress metadata are unavailable, the implementation retains Uncertainty–Conflict Arbitration, equivalent to setting $b = 0$. Equations 3, 4, and 5 form a single dependency: Prior Anchoring supplies $p_{j,k}$, Uncertainty–Conflict Arbitration proposes the weight, and Metaplastic Calibration determines how much of its increase is retained in the final score.

3.5 Planning Update

In BIPR, the search mean carries the planning update: candidate evaluation changes the next proposal distribution, while the prior and network parameters remain unchanged. This closes the loop from learned behavior through calibrated evaluation to the next imagined action set.

At round j, draw K sequences from $\mathcal{N}(\boldsymbol{\mu}_j, \mathrm{diag}(\boldsymbol{\sigma}_s^2))$ and replace one sample with $\boldsymbol{\mu}_j$. After scoring, let $\mathcal{S}_j$ contain the $M = 30$ lowest-score candidates; unlike $\mathcal{E}_j$, it is selected by the regularized score. With temperature $T = 0.5$, the normalized weights and updated mean are

$$w_{j,k} = \frac{\exp(-s_{j,k}/T)}{\sum_{\ell \in \mathcal{S}_j} \exp(-s_{j,\ell}/T)}, \qquad \boldsymbol{\mu}_{j+1} = \sum_{k \in \mathcal{S}_j} w_{j,k} \mathbf{a}_j^{(k)}, \tag{6}$$

where ℓ is a dummy candidate index and weights are zero outside $\mathcal{S}_j$. After J rounds, the planner executes the first $H_{\text{exec}} \leq H$ actions or action blocks of $\boldsymbol{\mu}_J$ and then replans if needed. No executed rewards enter Prior Anchoring, Uncertainty–Conflict Arbitration, or Metaplastic Calibration.

All three modules reuse the same K action sequences and model costs. Their additional arithmetic is $O(JKD)$, plus elite selection, and they require neither additional trainable parameters nor additional dynamics rollouts. Figure 2 summarizes the information flow; pseudocode and a notation table are provided in Appendix A.

4 Experiments

We evaluate BIPR on standardized world-model benchmarks to determine whether it improves control under a fixed rollout budget, transfers across predictive backbones, achieves a favorable success–cost trade-off against CEM, and derives its gains from the proposed components.

4.1 Evaluation Protocol

Tasks and backbones. The evaluation covers four complementary control environments. Cube is an OGBench goal-conditioned manipulation task that requires an agent to move a cube to a target configuration (Park et al., 2025). PushT is a contact-rich planar task in which a circular end-effector must push a T-shaped object to a target position and orientation (Chi et al., 2025; Maes et al., 2026a). Reacher is a continuous-control task from the DeepMind Control Suite in which an articulated arm must reach a visual target under MuJoCo dynamics (Tunyasuvunakool et al., 2020; Todorov et al., 2012). TwoRoom is a visual navigation task that requires an agent to reach a target in the opposite room through constrained doorways (Maes et al., 2026a;b). We evaluate frozen LeWM (Maes et al., 2026b) and official pretrained Sub-JEPA (Zhao et al., 2026) checkpoints.

Environment and metric provenance. The complete evaluation uses the standardized environments and planning interface provided by stable-worldmodel (Maes et al., 2026a). We retain the evaluator's task-specific success predicates and initial–goal sampling. Success rate is the fraction of episodes marked successful by that evaluator. For the Average column, task success is first averaged within each seed and then aggregated across seeds.

Planning budget. The main protocol uses $K = 128$ candidate action sequences per update and $J = 30$ updates, with five action blocks of five primitive actions each. Goals are selected at a 25-step offset, the standard execution prefix contains five blocks, and the evaluator permits up to 50 primitive steps. We keep the world model frozen throughout prior-head training and evaluation and freeze the prior head during evaluation. Experiments were conducted on NVIDIA L40S GPUs with 48 GB of memory. Wall-clock measurements include the complete evaluation workload to ensure fair comparison.

Comparisons. We compare BIPR with MPPI (Williams et al., 2018), CEM (Rubinstein, 1999), PRISM (Wang et al., 2026), SAGE (Cheng et al., 2026), and FF-JEPA (Masip et al., 2026). MPPI uses no learned action prior, whereas PRISM uses the shared prior head and proposal fusion without candidate regularization. For fair comparison, we adapt SAGE and FF-JEPA to the same environment interface, candidate budget, planning horizon, evaluation seeds, and episode protocol while retaining their respective learned-action-proposal and deterministic latent-planning mechanisms.

Table 1: **Main results with frozen LeWM.** Task success rates (%) are reported as mean ± sample standard deviation across three evaluation seeds, with 50 episodes per seed. All rows use $K = 128$ candidate action sequences and 30 planning updates; SAGE retains its elite-CEM optimizer.

Planner	Cube	PushT	Reacher	TwoRoom	Average
FF-JEPA	36.67 ± 4.16	6.67 ± 3.06	14.67 ± 7.02	38.00 ± 5.29	24.00
SAGE	58.67 ± 12.06	15.33 ± 5.77	16.67 ± 5.03	58.00 ± 5.29	37.17
MPPI	44.67 ± 2.31	58.00 ± 0.00	40.00 ± 2.00	68.00 ± 4.00	52.67
PRISM	78.67 ± 7.02	88.00 ± 3.46	49.33 ± 7.02	84.00 ± 4.00	75.00
BIPR	$\mathbf{84.00 \pm 5.29}$	$\mathbf{89.33 \pm 1.15}$	$\mathbf{68.00 \pm 2.00}$	$\mathbf{90.67 \pm 4.16}$	**83.00**

Table 2: **Transfer to frozen Sub-JEPA on the same four environments.** Protocol and reporting match Table 1.

Planner	Cube	PushT	Reacher	TwoRoom	Average
FF-JEPA	42.00 ± 7.21	7.33 ± 4.16	11.33 ± 4.16	34.67 ± 1.15	23.83
SAGE	43.33 ± 3.06	10.00 ± 5.29	14.67 ± 5.03	48.00 ± 3.46	29.00
MPPI	51.33 ± 2.31	66.00 ± 6.00	44.00 ± 4.00	70.00 ± 2.00	57.83
PRISM	$\mathbf{92.67 \pm 3.06}$	$\mathbf{90.00 \pm 8.00}$	45.33 ± 2.31	89.33 ± 3.06	79.33
BIPR	$\mathbf{92.67 \pm 2.31}$	88.00 ± 5.29	$\mathbf{60.67 \pm 6.11}$	$\mathbf{95.33 \pm 1.15}$	**84.17**

4.2 Main Results

Table 1 shows that BIPR reaches 83.00% average success, improving over MPPI by 30.33 percentage points and over PRISM by 8.00 points. The largest improvement over PRISM occurs on Reacher, from 49.33% to 68.00%. Cube and TwoRoom improve by 5.33 and 6.67 points, respectively, while PushT improves by 1.33 points.

The gain is largest on Reacher because precise goal-directed control is particularly sensitive to search drift: persistent prior-relative scoring keeps repeatedly updated action proposals aligned with plausible reaching behavior. Cube and TwoRoom also benefit because their solutions require structured spatial progress toward the goal. PushT shows a smaller gain because PRISM already reaches 88.00%, leaving limited headroom, and contact-rich pushing depends strongly on accurately predicting local interactions. The consistent improvement across the four environments indicates that maintaining prior guidance throughout optimization benefits both manipulation and navigation under a fixed rollout budget.

Table 2 evaluates transfer to a different predictive representation. BIPR reaches 84.17% average success, improving over Sub-JEPA with MPPI by 26.33 points and over PRISM by 4.83 points. Reacher and TwoRoom improve by 15.33 and 6.00 points, Cube ties, and PushT decreases by 2.00 points. These results show that the same prior-regulation mechanism transfers across the two end-to-end JEPA backbones.

4.3 Success–Cost Trade-off Against CEM

As shown in Table 3, when the candidate count is $K = 128$, BIPR with MPPI reaches 78.33% average success and obtains the best Reacher result, while CEM with prior-mean initialization reaches 79.00%. Compared with standard CEM, BIPR increases average success from 76.67% to 78.33% and reduces evaluation wall-clock time from 290.28 to 154.63 seconds, a 46.73% reduction. Relative to MPPI at the same budget, the gain is 29.33 points with only a 3.71% increase in wall-clock time. Thus, BIPR raises MPPI to performance comparable with CEM at a small additional evaluation cost while preserving MPPI's substantial timing advantage.

The behavior changes when the candidate count increases to $K = 300$. The average success of BIPR decreases slightly from 78.33% to 76.33%, whereas the other planners improve modestly; nevertheless, BIPR obtains the best PushT result at 92.67%. Relative to MPPI, it improves average

Table 3: **Solver-budget comparison on PushT and Reacher.** K denotes the number of candidate action sequences evaluated per planning update. Results use three seeds and 50 episodes per task; time is mean evaluation wall-clock time in seconds. "Prior mean" denotes CEM initialized by the learned prior mean. Bold marks the highest task and average success and the lowest time within each candidate budget.

Guidance	Solver	K	PushT	Reacher	Average	Time
None	MPPI	128	58.00 ± 0.00	40.00 ± 2.00	49.00	**149.10**
None	CEM	128	88.67 ± 6.11	64.67 ± 4.62	76.67	290.28
BIPR	MPPI	128	88.67 ± 5.03	$\mathbf{68.00 \pm 2.00}$	78.33	154.63
Prior mean	CEM	128	$\mathbf{92.00 \pm 0.00}$	66.00 ± 8.72	**79.00**	291.77
None	MPPI	300	66.67 ± 8.33	40.67 ± 5.03	53.67	**167.76**
None	CEM	300	90.00 ± 4.00	$\mathbf{71.33 \pm 4.16}$	**80.67**	464.55
BIPR	MPPI	300	$\mathbf{92.67 \pm 3.06}$	60.00 ± 5.29	76.33	169.65
Prior mean	CEM	300	91.33 ± 5.77	68.67 ± 9.02	80.00	462.82

success by 22.66 points while increasing wall-clock time by only 1.13%, less than the 3.71% increase at $K = 128$. It also reduces evaluation time by 63.34% relative to prior-mean CEM. A larger candidate set provides broader coverage of the contact-sensitive PushT action space, which is consistent with the improved PushT result, but it also exposes more model-favored candidates and can amplify objective error on precise reaching, where BIPR decreases from 68.00% to 60.00%. These results indicate that BIPR is most useful for low-cost, fast planning: it makes MPPI competitive with CEM while retaining a large cost advantage, rather than relying on a high-cost large-candidate optimizer.

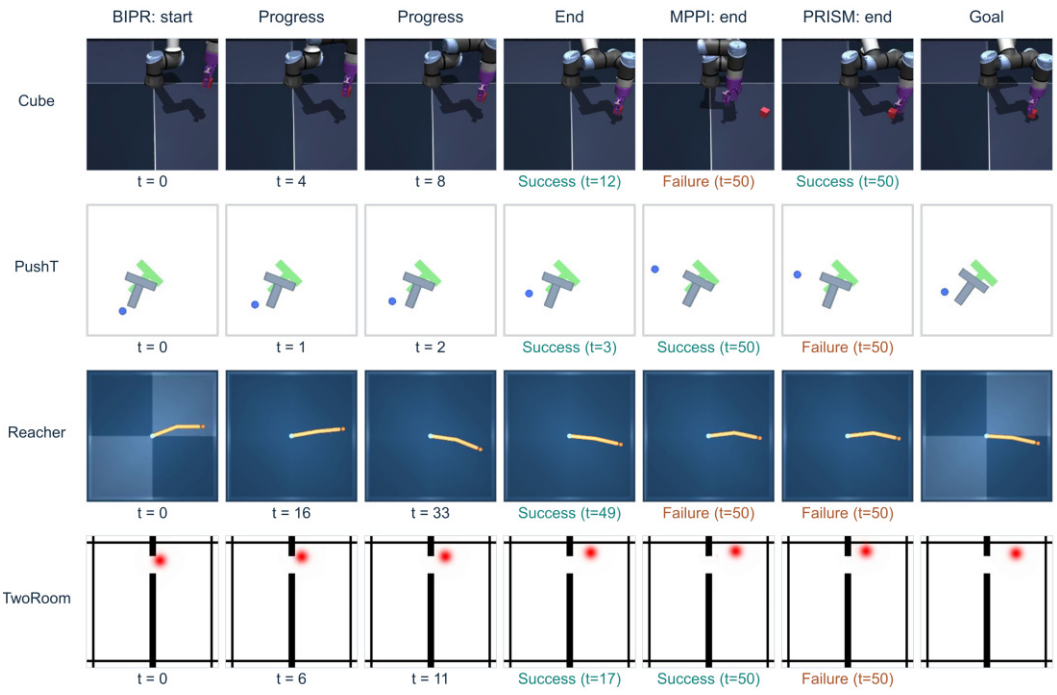

Figure 3: **Successful BIPR execution and recorded baseline outcomes.** Each row uses a matched initial–goal pair with LeWM. Four BIPR snapshots span the initial state to its first recorded success; the MPPI and PRISM panels show their final states after the full evaluation budget. Labels give episode outcomes and execution steps t.

Table 4: **Component ablations on LeWM.** Three evaluation seeds, 50 episodes per task per seed; $K = 128$, 30 updates. The first four rows add components; the remaining rows intervene on the full model.

Configuration	PushT	Reacher	Average
Prior head	88.00 ± 3.46	49.33 ± 7.02	68.67
+ Prior Anchoring	88.67 ± 1.15	62.67 ± 4.16	75.67
+ Uncertainty–Conflict Arbitration	88.00 ± 3.46	68.00 ± 2.00	78.00
+ Metaplastic Calibration (Full)	$\mathbf{89.33 \pm 1.15}$	$\mathbf{68.00 \pm 2.00}$	**78.67**
Full − fixed anchor floor	87.33 ± 4.62	66.00 ± 0.00	76.67
Full − uncertainty branch	88.67 ± 1.15	63.33 ± 3.06	76.00
Full − conflict branch	88.67 ± 1.15	64.67 ± 6.43	76.67
Full + shuffled quality	88.00 ± 3.46	62.67 ± 4.16	75.33

4.4 Visualizations

Figure 3 presents planning examples from different methods across all four environments. BIPR succeeds in every displayed case. In Cube, PushT, and TwoRoom, BIPR reaches the goal in substantially fewer execution steps, and it also completes the more difficult Reacher task within the prescribed budget. These examples show that BIPR maintains goal-directed search while suppressing drift from plausible behavior, enabling faster and more accurate task completion than the compared planners in the displayed cases and providing further empirical support for the brain-inspired regulation design.

4.5 Component and Counterfactual Ablations

Table 4 distinguishes incremental gains from full-model counterfactuals. Prior Anchoring produces the largest incremental improvement, increasing average success by 7.00 points, followed by 2.33 points from Uncertainty–Conflict Arbitration and 0.67 points from Metaplastic Calibration. Among the counterfactual interventions, removing the uncertainty branch causes the largest performance decrease, while shuffling the quality signal also substantially reduces success. These results identify persistent prior anchoring as the dominant source of improvement and show that the full model also depends particularly on uncertainty-based arbitration. Shuffling quality breaks the alignment between learning history and the current regulation response, causing penalties to be applied at inappropriate times and degrading performance.

The three modules introduced by BIPR require no additional training. Across all four environments, BIPR has a mean evaluation wall-clock time of 319.15 seconds, only 0.82% higher than PRISM; detailed cost accounting is reported in Appendix A. Additional hyperparameter-sensitivity experiments and qualitative visualizations are provided in Appendices C and D.

5 Conclusion

We introduced BIPR, a brain-inspired planning framework that extends learned action priors from proposal initialization to persistent, adaptively calibrated candidate regulation. On standard world-model control benchmarks, BIPR improves average success over MPPI by 30.33 and 26.33 percentage points with LeWM and Sub-JEPA, respectively, while reducing evaluation wall-clock time by 46.73% relative to CEM under a matched candidate budget. Meanwhile, BIPR introduces no additional trainable parameters and incurs only 0.82% additional evaluation overhead.

Future work will extend BIPR to longer-horizon manipulation, embodied navigation, and real-robot control, while evaluating it with a broader range of world-model architectures and action priors. Replacing fixed reliability signals with closed-loop reachability calibration and dynamically adjusting prior influence under distribution shift could further improve robustness and enable efficient planning in more complex deployment settings.

GENERATIVE AI DISCLOSURE

Generative AI tools were used solely to improve the clarity and readability of the manuscript. The authors reviewed all AI-assisted edits and take full responsibility for the final content.

REFERENCES

Wickliffe C. Abraham and Mark F. Bear. Metaplasticity: the plasticity of synaptic plasticity. *Trends in Neurosciences*, 19(4):126–130, 1996. doi: 10.1016/S0166-2236(96)80018-X. URL https://doi.org/10.1016/S0166-2236(96)80018-X.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Marcus K. Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19(12):1697–1706, 2016. doi: 10.1038/nn.4401. URL https://www.nature.com/articles/nn.4401.

Letian Cheng, Qi Zhang, and Yisen Wang. SAGE: Subgoal-conditioned action generation for latent world model planning. *arXiv preprint arXiv:2607.17973*, 2026. doi: 10.48550/arXiv.2607.17973. URL https://arxiv.org/abs/2607.17973.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10–11):1684–1704, 2025. doi: 10.1177/02783649241273668. URL https://doi.org/10.1177/02783649241273668.

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*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

Nathaniel D. Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 2005. doi: 10.1038/nn1560. URL https://doi.org/10.1038/nn1560.

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*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

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

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8387–8406. PMLR, 2022. URL https://proceedings.mlr.press/v162/hansen22a.html.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915. PMLR, 2022. URL https://proceedings.mlr.press/v162/janner22a.html.

Kristopher T. Jensen, Guillaume Hennequin, and Marcelo G. Mattar. A recurrent network model of planning explains hippocampal replay and human behavior. *Nature Neuroscience*, 27(7):1340–1348, 2024. doi: 10.1038/s41593-024-01675-7. URL https://www.nature.com/articles/s41593-024-01675-7.

Alfredo Kirkwood, Marika G. Rioult, and Mark F. Bear. Experience-dependent modification of synaptic plasticity in visual cortex. *Nature*, 381(6582):526–528, 1996. doi: 10.1038/381526a0. URL https://www.nature.com/articles/381526a0.

Sang Wan Lee, Shinsuke Shimojo, and John P. O'Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699, 2014. doi: 10.1016/j.neuron.2013.11.028. URL https://doi.org/10.1016/j.neuron.2013.11.028.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lucas Maes, Quentin Le Lidec, Luiz Facury, Nassim Massaudi, Ayush Chaurasia, Francesco Capuano, Richard Gao, Taj Gillin, Dan Haramati, Damien Scieur, Yann LeCun, and Randall Balestriero. stable-worldmodel: A platform for reproducible world modeling research and evaluation. *arXiv preprint arXiv:2605.21800*, 2026a. doi: 10.48550/arXiv.2605.21800. URL https://arxiv.org/abs/2605.21800.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026b. doi: 10.48550/arXiv.2603.19312. URL https://arxiv.org/abs/2603.19312.

Sergi Masip, Jonathan Swinnen, Yutong Hu, Renaud Detry, and Tinne Tuytelaars. FF-JEPA: Long-horizon planning in world models with latent planners. *arXiv preprint arXiv:2606.09311*, 2026. doi: 10.48550/arXiv.2606.09311. URL https://arxiv.org/abs/2606.09311.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ecd92623ac899357312aaa8915853699-Abstract-Conference.html.

Karl Pertsch, Youngwoon Lee, and Joseph Lim. Accelerating reinforcement learning with learned skill priors. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 188–204. PMLR, 2021. URL https://proceedings.mlr.press/v155/pertsch21a.html.

Brad E. Pfeiffer and David J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79, 2013. doi: 10.1038/nature12112. URL https://www.nature.com/articles/nature12112.

Cristina Pinneri, Shambhuraj Sawant, Sebastian Blaes, Jan Achterhold, Joerg Stueckler, Michal Rolinek, and Georg Martius. Sample-efficient cross-entropy method for real-time planning. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 1049–1065. PMLR, 2021. URL https://proceedings.mlr.press/v155/pinneri21a.html.

Reuven Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1(2):127–190, 1999. doi: 10.1023/A:1010091220143. URL https://doi.org/10.1023/A:1010091220143.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antić, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aPOpXlnV1T.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109. URL https://doi.org/10.1109/IROS.2012.6386109.

Elia Trevisan and Javier Alonso-Mora. Biased-MPPI: Informing sampling-based model predictive control by fusing ancillary controllers. *IEEE Robotics and Automation Letters*, 9(6):5871–5878, 2024. doi: 10.1109/LRA.2024.3397083. URL https://doi.org/10.1109/LRA.2024.3397083.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. doi: 10.1016/j.simpa.2020.100022. URL https://doi.org/10.1016/j.simpa.2020.100022.

Yuhai Wang, Jiawei Xia, Rongxuan Zhou, Xiao Hu, Yongliang Shi, Jing Du, and Yang Ye. PRISM: PRior-guided imagination sampling in world models. *arXiv preprint arXiv:2606.07974*, 2026. doi: 10.48550/arXiv.2606.07974. URL https://arxiv.org/abs/2606.07974.

Grady Williams, Paul Drews, Brian Goldfain, James M. Rehg, and Evangelos A. Theodorou. Information-theoretic model predictive control: Theory and applications to autonomous driving. *IEEE Transactions on Robotics*, 34(6):1603–1622, 2018. doi: 10.1109/TRO.2018.2865891. URL https://doi.org/10.1109/TRO.2018.2865891.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are scalable world models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2173. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7dbb5bfab324e3b86af9bd0df15498dd-Abstract-Conference.html.

Guojian Zhan, Letian Tao, Pengcheng Wang, Yixiao Wang, Yuxin Chen, Yiheng Li, Hongyang Li, Masayoshi Tomizuka, and Shengbo Li. Mean flow policy with instantaneous velocity constraint for one-step action generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fe7f375ef01e43f17d2c32b28a176577-Abstract-Conference.html.

Kai Zhao, Dongliang Nie, Yuchen Lin, Zhehan Luo, Yixiao Gu, Deng-Ping Fan, and Dan Zeng. Sub-JEPA: Subspace gaussian regularization for stable end-to-end world models. *arXiv preprint arXiv:2605.09241*, 2026. doi: 10.48550/arXiv.2605.09241. URL https://arxiv.org/abs/2605.09241.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

A Algorithm and Notation

Algorithm 1 gives one planning call of BIPR. Every call uses the same frozen encoder, predictor, and prior head; only the search mean and scalar regulation signals change. The candidate actions and costs computed in each round are shared by Prior Anchoring, Uncertainty–Conflict Arbitration, and Metaplastic Calibration.

Algorithm 1 BIPR planning

Require: Current/goal observations o_t, o_g and any required observation history;
frozen E, F, G; base moments $\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b$;
checkpoint errors $L_{\text{init}}, L_{\text{best}}$ if available;
$H, H_{\text{exec}}, K, J, M, T$ and regulation settings (Table 5).
Ensure: Executable prefix containing H_{exec} action blocks.
Encode $z_t = E(o_t)$ and $z_g = E(o_g)$.
Predict $(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p) = G(z_t, z_g)$.
Initialize $\boldsymbol{\mu}_0, \boldsymbol{\sigma}_s$ using Equation 2.
Set $\widetilde{\sigma}_p[d] = \max\{\sigma_p[d], \sigma_{\min}\}$ for $d = 1, \ldots, D$.
Compute $\overline{\sigma}_p$ and u using Equation 4.
if both checkpoint errors are available and $L_{\text{init}} > 0$ **then**
    Compute q and $b(q)$ using Equation 5.
else
    Set the Metaplastic Calibration gate to zero (retain Uncertainty–Conflict Arbitration).
end if
for $j = 0, \ldots, J-1$ **do**
    Sample $\mathbf{a}_j^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_j, \mathrm{diag}(\boldsymbol{\sigma}_s^2))$ for $k = 1, \ldots, K$.
    Set $\mathbf{a}_j^{(1)} \leftarrow \boldsymbol{\mu}_j$ (replace one sample; do not increase K).
    Roll out F to obtain $c_{j,k} = C_{\text{WM}}(\mathbf{a}_j^{(k)})$ for all k.
    Prior Anchoring: compute $p_{j,k} = P(\mathbf{a}_j^{(k)})$ using Equation 3.
    Select $\mathcal{E}_j$ using the lowest $\max\{1, \mathrm{round}(\alpha K)\}$ model costs $c_{j,k}$.
    Uncertainty–Conflict Arbitration: compute v_j and λ_j^{UCA} using Equation 4.
    Metaplastic Calibration: obtain λ_j from the progress gate using Equation 5.
    Compute $s_{j,k} = c_{j,k} + \lambda_j p_{j,k}$ for all k.
    Select $\mathcal{S}_j$ using the $\min(M, K)$ lowest combined scores $s_{j,k}$.
    Compute $w_{j,k}$ and $\boldsymbol{\mu}_{j+1}$ using Equation 6; keep $\boldsymbol{\sigma}_s$ fixed.
end for
return the first H_{exec} blocks of $\boldsymbol{\mu}_J$, converted to environment action units.

Table 5: Default configuration of the archived final LeWM planner.

Quantity	Value
Candidate count / optimization rounds	$K = 128$ / $J = 30$
Planning horizon / action block	5 blocks / 5 primitive actions
Execution prefix / episode budget	5 blocks / at most 50 primitive actions
MPPI retained candidates / temperature	30 / 0.5
Prior standard-deviation floor / penalty cap	0.05 / 9
Anchor floor / maximum weight	0.5 / 1.0
Spread threshold / gain	0.9 / 30
Conflict elite fraction / gain	0.1 / 8
Quality interval	$(0.15, 0.30)$
Prior MLP hidden width / likelihood exponent	512 / $\beta = 0.5$
episodes	50

Shared prior-head training. The shared head is a three-layer multilayer perceptron with hidden width 512, trained offline on cached current/goal features and normalized action targets (Wang et al., 2026). It uses a variance-weighted Gaussian negative log-likelihood (Seitzer et al., 2022), with variance-weighting exponent $\beta = 0.5$, and AdamW (Loshchilov & Hutter, 2019). These are baseline training choices, not additional BIPR modules.

Table 6: Notation used by BIPR. Dimensions refer to normalized planning coordinates.

Symbol	Definition
$t;\ o_t, o_g$	Environment time; current and goal observations.
E, F, G	Encoder, predictor, and prior head.
$z_t, z_g;\ d_z$	Current/goal latent representations and latent dimension.
H, H_{exec}, D	Planning horizon, executed prefix, and total scalar action-sequence dimension.
$\mathbf{a};\ a[d]$	Candidate action sequence and its d-th flattened scalar coordinate.
$J, j;\ K, k$	Optimization-round count and index; candidate count and index.
$\mathbf{a}_j^{(k)}$	Candidate sequence k at round j.
$\widehat{z}_H, C_{\text{WM}}$	Terminal latent prediction and backbone-specific goal-cost function.
$\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b$	Base planner mean and standard deviation before prior fusion.
$\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p$	Action-prior mean and standard deviation, fixed within a call.
$\boldsymbol{\tau}_b, \boldsymbol{\tau}_p$	Stabilized base and prior precision vectors used for fusion.
$\boldsymbol{\mu}_j, \boldsymbol{\sigma}_s$	Search mean at round j and fixed fused sampling standard deviation.
$\sigma_{\min}, \widetilde{\sigma}_p[d]$	Standard-deviation floor and floored prior scale at coordinate d.
$P, P_{\max};\ p_{j,k}$	Prior-penalty function, cap, and candidate penalty.
$c_{j,k}, s_{j,k}$	Model cost and combined regularized score of candidate k at round j.
$\overline{\sigma}_p, u;\ k_u, \tau_u$	Mean floored prior scale and spread signal; spread gain and threshold.
$\alpha, \mathcal{E}_j$	Model-elite fraction and candidate-index set selected by model cost.
$\overline{p}_j, \overline{p}_j^E$	Population and model-elite mean prior penalties.
v_j, k_c	Bounded conflict signal and conflict gain.
λ_j^{UCA}	Proposed arbitration weight.
$\lambda_{\text{base}}, \lambda_{\max}$	Persistent baseline and maximum regularization weights.
$L_{\text{init}}, L_{\text{best}};\ q$	Initial/best validation action MSE and clipped checkpoint progress.
$\tau_{\text{low}}, \tau_{\text{high}};\ b(q)$	Strict progress-interval boundaries and binary Metaplastic Calibration gate.
λ_j	Final calibrated weight used in candidate scoring.
$M, \mathcal{S}_j$	Retained-candidate count and set selected by combined score.
$T, w_{j,k}$	Cost-weighting temperature and normalized candidate weight.
ϵ_f, ϵ_c	Fusion stabilizer (10^{-8}) and conflict-denominator floor (10^{-6}).

Table 7: Four-task LeWM cost accounting. Cache and training times are totals across environments. Evaluation time is the report's mean wall time per evaluation unit. Backbone pretraining is excluded.

Method	Parameters (M)	Cache (h)	Train (h)	Cold total (h)	Eval (s)
FF-JEPA (local)	9.46	1.35	0.06	1.41	609.64
SAGE (local)	34.30–34.45	1.35	0.47	1.82	602.98
PRISM	0.51–0.59	1.35	1.36	2.71	316.56
BIPR	0.51–0.59	1.35	1.36	2.71	319.15

B Training and Evaluation Costs

PRISM and BIPR share the same trained head; the modules add no learnable parameters or preparation stage. As shown in Table 7, compared with the state-of-the-art PRISM baseline, BIPR incurs only a slight increase in planning time while leaving all other costs unchanged.

C Sensitivity and Additional Planner Experiments

C.1 Hyperparameter Sensitivity

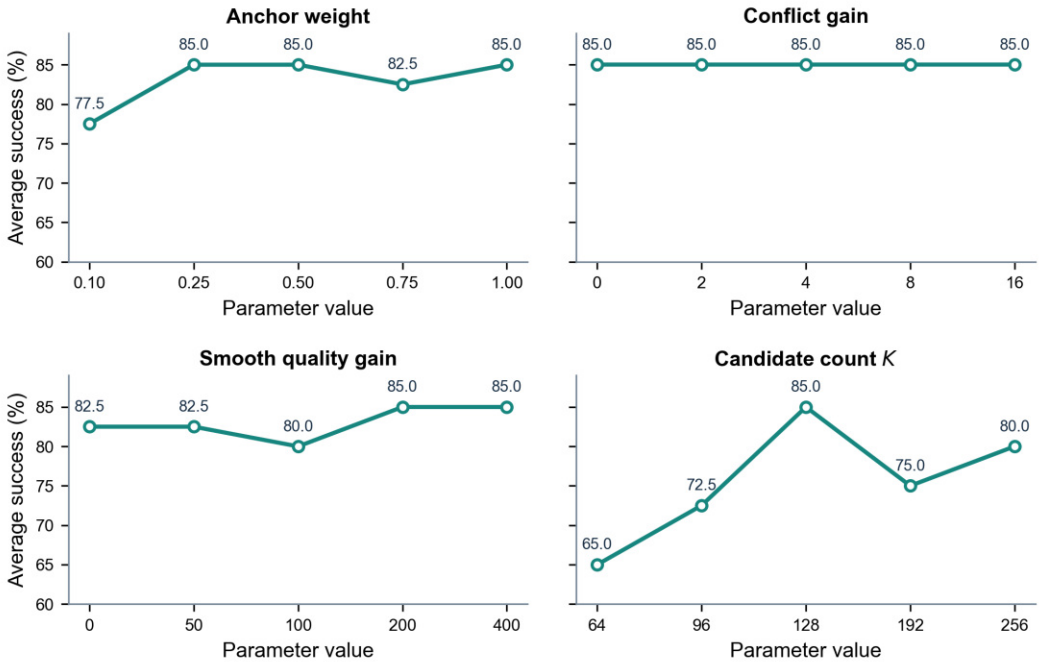

Figure 4: Hyperparameter sensitivity of BIPR to the anchor weight, conflict gain, smooth quality gain, and candidate count K. Each panel varies one parameter and reports the average success rate (%) across four environments.

As shown in Figure 4, the anchor weight and smooth quality gain produce broadly consistent performance across parameter settings, the candidate count K performs best at 128, and changing the conflict gain does not affect performance.

Table 8: Sensitivity to the number of optimization rounds on PushT and Reacher. Three evaluation seeds, 50 episodes per task per seed. Time is mean evaluation wall time.

Updates J	5	10	20	30	40
Average success (%)	72.33	75.67	78.00	78.67	73.00
Time (s)	81.12	96.60	123.28	154.68	182.19

C.2 Optimization Rounds

Table 8 shows that performance first increases and then decreases as the number of updates grows, whereas planning time increases monotonically. We attribute this trend to additional optimization shifting the search mean toward candidates favored by an imperfect objective.

D Qualitative Rollout Sequences

As shown in Figures 5, 6, 7, and 8, we compare representative executions of BIPR and the baselines across four environments. All displayed episodes use LeWM and the final hard-gate BIPR configuration. We crop the live top-left region and the static-goal top-right region into 224×224 quadrants. The first-success snapshot is captured directly from the simulator when the evaluator first marks the episode successful. The results indicate that BIPR better captures goal-conditioned action relationships and model–prior consistency, enabling stable success across contact-rich manipulation, precise reaching, and visual navigation. We attribute this behavior to persistent prior anchoring and adaptive calibration, which keep candidate actions near reliable behavioral structure while permitting goal-improving deviations. MPPI exhibits search drift and unstable progress without a learned behavioral reference, whereas PRISM benefits from prior-guided initialization but can deviate from that guidance during optimization. Its lack of persistent candidate regulation and adaptive calibration makes sustained goal-directed execution more difficult. Figures 9 and 10 show examples in which BIPR and the comparison methods all fail across different environments. In PushT, the planners fail to establish effective contact and move the object toward the goal, while in Reacher they approach but do not attain the target configuration. These cases indicate that, under difficult initial–goal configurations and a limited execution budget, BIPR still requires more reliable action priors and stronger long-horizon goal guidance.



Figure 5: **Cube: matched execution sequences.** BIPR and PRISM succeed; MPPI fails. BIPR's final displayed state is its first recorded success at step 12; baseline outcomes refer to the full evaluation budget.

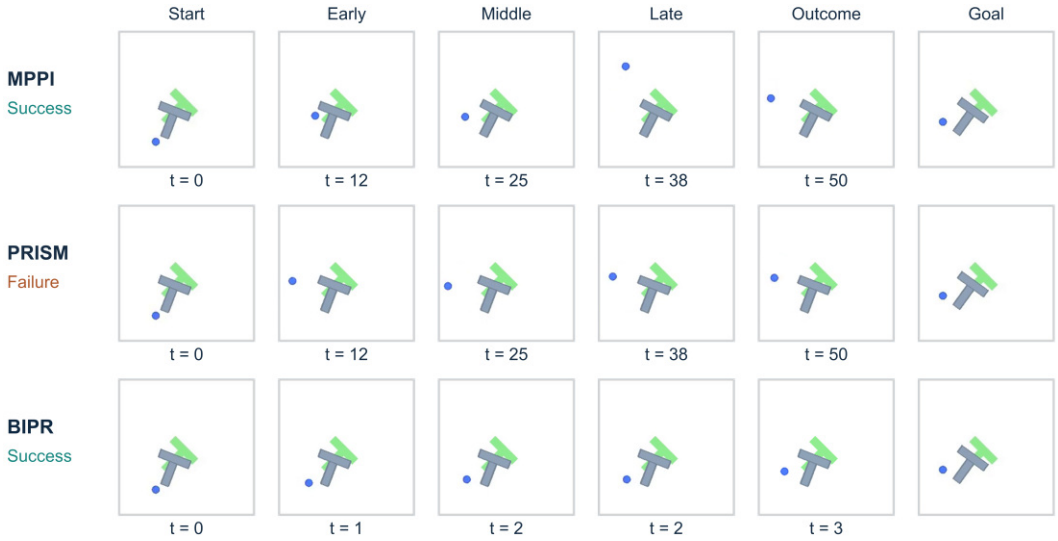

Figure 6: **PushT: matched execution sequences.** BIPR and MPPI succeed; PRISM fails. BIPR's final displayed state is its first recorded success at step 3; baseline outcomes refer to the full evaluation budget.

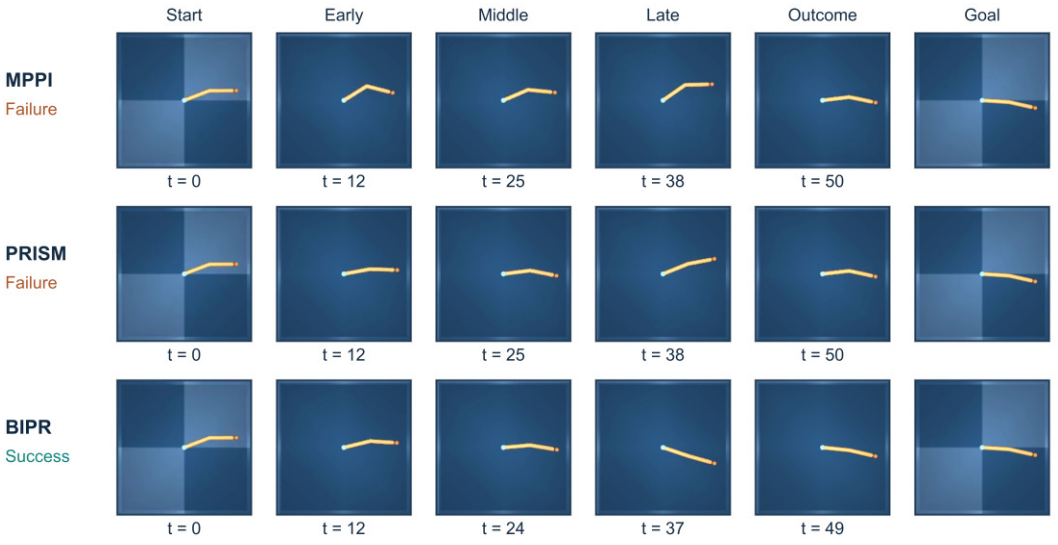

Figure 7: **Reacher: matched execution sequences.** BIPR succeeds while MPPI and PRISM fail. BIPR's final displayed state is its first recorded success at step 49; baseline outcomes refer to the full evaluation budget.

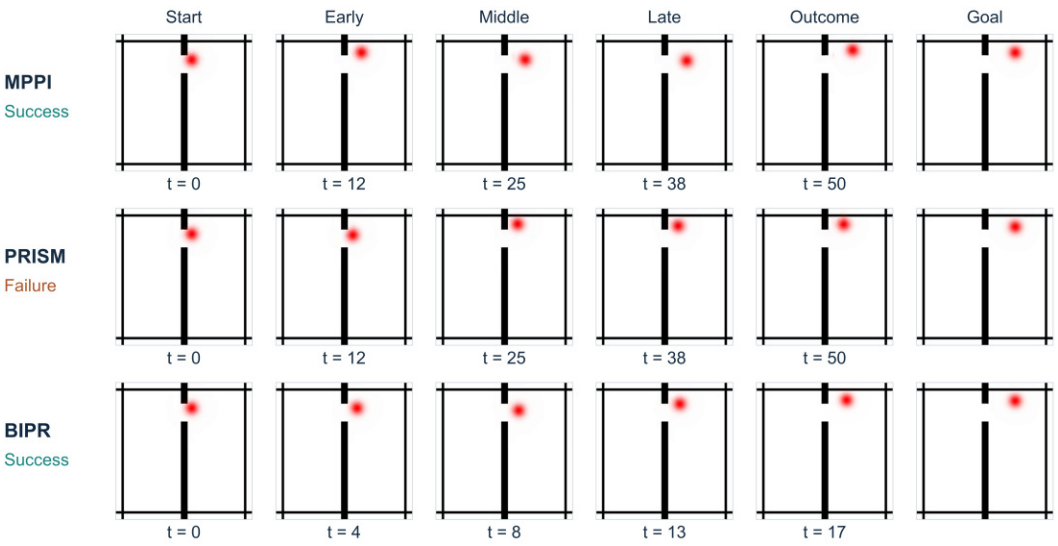

Figure 8: **TwoRoom: matched execution sequences.** BIPR and MPPI succeed; PRISM fails. BIPR's final displayed state is its first recorded success at step 17; baseline outcomes refer to the full evaluation budget.

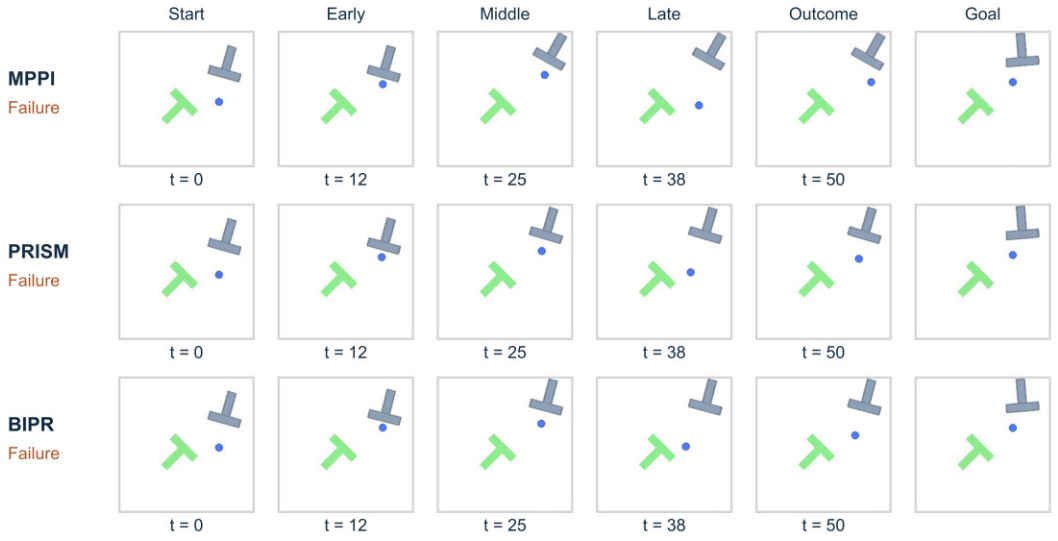

Figure 9: **PushT: shared failure case.** MPPI, PRISM, and BIPR start from the same state and pursue the same visual goal, but none completes the task within 50 execution steps.

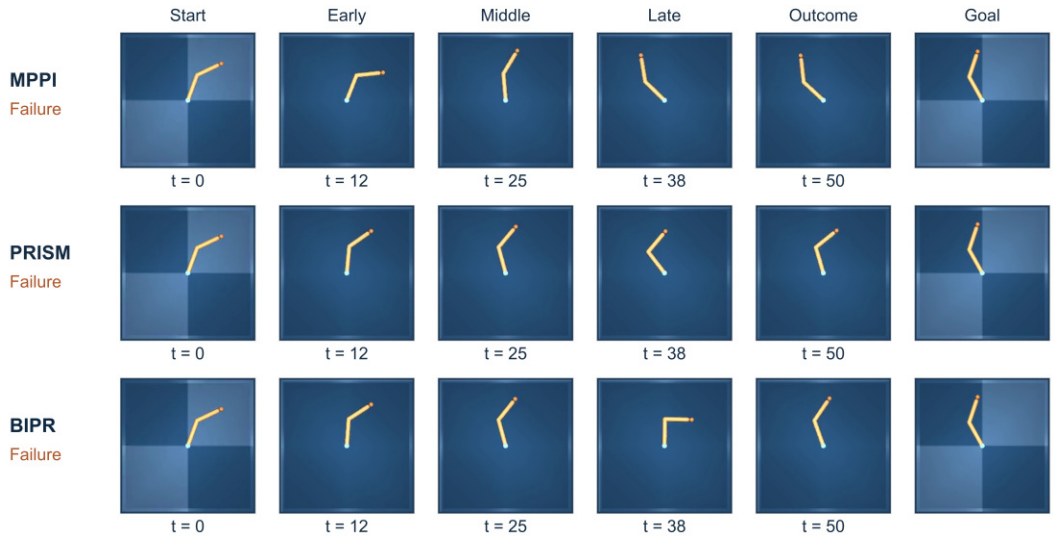

Figure 10: **Reacher: shared failure case.** MPPI, PRISM, and BIPR start from the same state and pursue the same visual goal, but none reaches the target configuration within 50 execution steps.