RIME: RELIABLE IMAGINATION FOR META-BLACK-BOX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Meta-black-box optimization (MetaBBO) learns optimizer configuration from optimization trajectories. Surrogate-assisted training reduces the cost of collecting optimization trajectories by replacing expensive objective evaluations with surrogate predictions. However, existing methods use the collected transitions only for direct policy updates and do not learn the optimizer dynamics needed to generate additional training experience. A learned dynamics model can reuse these trajectories through imagined transitions, provided its predictions are reliable. This is challenging because the collected trajectories already reflect surrogate approximation errors, while errors in predicted dynamics can accumulate over imagined rollouts. We propose RIME, a MetaBBO algorithm that learns optimizer configuration through task-conditioned, uncertainty-gated imagination. We first learn a task representation through objective-surrogate pretraining to provide shared context for optimizer control and dynamics prediction. We then learn a task-conditioned world model that reuses collected trajectories through imagined experience, with predictive uncertainty determining which transitions are trusted and the length of imagined rollouts. RIME achieves the best average rank on the evaluated BBOB benchmark across five dimensionalities and outperforms most MetaBBO baselines on UAV path-planning tasks.

1 INTRODUCTION

Evolutionary algorithms such as Differential Evolution (DE) (Storn & Price, 1997) and Particle Swarm Optimization (Kennedy & Eberhart, 1995) are robust, but their performance depends strongly on operator and hyperparameter choices. Classical parameter-control methods reduce manual tuning through adaptive or self-adaptive mechanisms (Eiben et al., 1999; Brest et al., 2006; Zhang & Sanderson, 2009; Tanabe & Fukunaga, 2013), whereas MetaBBO replaces part of this design process with a learned policy that observes an optimizer state and configures its next update (Li & Malik, 2017; Ma et al., 2023; Lange et al., 2023).

Surrogate models have long been used to reduce expensive black-box and evolutionary optimization evaluations (Jones et al., 1998; Jin, 2005; Forrester & Keane, 2009; Jin, 2011). Surrogate-assisted MetaBBO brings this idea into learned optimizer control by replacing expensive objective queries during policy training with learned approximations (Ma et al., 2025b). Model-free policy learning uses the resulting transitions to update the controller directly. These transitions also record how optimizer configurations affect search states and progress. Learning this relationship can provide additional training experience through predicted transitions, as world models do in reinforcement learning (Hafner et al., 2019a; Hansen et al., 2024).

The challenge is to turn predicted trajectories into useful training experience without amplifying model errors. Objective-surrogate errors can already alter the trajectories collected during training. A world model fitted to these trajectories introduces further prediction errors that can accumulate over imagined rollouts and bias controller updates (Janner et al., 2019). Its reliability also varies across tasks and regions of the search space, so a fixed rollout length can extend predictions beyond what the model has learned accurately. Learning from imagination therefore requires deciding which predicted transitions to use and how far to roll out the model (Chua et al., 2018; Pan et al., 2020).

These considerations suggest that imagination in MetaBBO should be used selectively: predicted optimizer transitions should be exploited only when the learned dynamics are sufficiently trustworthy,

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.

and the extent of imagination should depend on that reliability. Based on this principle, we propose RIME, which augments surrogate-assisted MetaBBO with task-conditioned optimizer dynamics. A task representation learned through objective-surrogate pretraining provides common context for both configuration decisions and dynamics modeling, allowing transitions to be interpreted relative to the underlying optimization task. Rather than treating model-generated trajectories as uniformly valid training data, RIME uses predictive uncertainty as a control signal: reliable predictions are used for imagined controller updates, increasing uncertainty progressively shortens the imagination horizon, and unreliable predictions trigger a fallback to executed replay. In this way, the world model serves as a selective mechanism for reusing collected optimization experience rather than an unrestricted simulator of future trajectories.

Our contributions are:

- We propose RIME, a reliability-aware MetaBBO algorithm with three properties. (1) *Task awareness*: a shared representation provides context for dynamics prediction and optimizer configuration. (2) *Trajectory reuse*: learned dynamics generate additional training experience from collected trajectories. (3) *Uncertainty-gated imagination*: prediction uncertainty controls which imagined transitions are used and how far rollouts extend.
- We develop a training mechanism that combines executed and imagined experience within a fixed controller-update budget. Global readiness checks, local disagreement filtering, and adaptive rollout lengths govern model use, with executed replay supplying updates when predictions are rejected.
- We derive a finite-horizon return-error bound for a fixed world model and policy in the surrogate environment. The bound relates accumulated rollout error to one-step prediction errors and rollout length.
- Experiments on BBOB benchmarks across five dimensionalities and on UAV path-planning tasks evaluate cross-dimensional generalization and transfer to practical optimization problems. RIME achieves the best average rank on the BBOB benchmarks and outperforms most MetaBBO baselines on UAV tasks.

2 Related Work

We review learned optimizer configuration and surrogate-assisted training, then discuss world models for experience reuse and task representations for prediction and control.

Meta-black-box optimization. Parameter and operator control has long been a central problem in evolutionary optimization (Eiben et al., 1999). For DE, jDE, JADE, SHADE, and L-SHADE adapt parameters using self-adaptation, past search successes, or population-size control (Brest et al., 2006; Zhang & Sanderson, 2009; Tanabe & Fukunaga, 2013; 2014). These approaches adapt an optimizer from statistics collected during a single run, whereas learning-based optimization seeks reusable control policies that transfer across problem instances. Early learning-to-optimize work formulated optimizer design itself as a reinforcement-learning problem (Li & Malik, 2017). MetaBBO learns optimizer behavior across distributions of black-box tasks. MetaBox provides a common interface for RL-based MetaBBO (Ma et al., 2023), and MetaBox-v2 extends benchmarking to multiple learning paradigms and realistic tasks (Ma et al., 2026). Learned evolution strategies parameterize optimizer updates directly (Lange et al., 2023). GLEET learns exploration–exploitation control (Ma et al., 2024), while DE-DDQN learns DE parameter and operator control (Sharma et al., 2019).

Surrogate modeling provides a complementary mechanism for reducing the cost of obtaining optimization experience. Approximate objective models have long been used to reduce expensive fitness evaluations in black-box and evolutionary optimization (Jones et al., 1998; Jin, 2005; Forrester & Keane, 2009; Jin, 2011). Within MetaBBO, Surr-RLDE incorporates this idea into policy learning by replacing expensive objective queries with learned surrogates during controller training (Ma et al., 2025b).

World models and model-based reinforcement learning. Dyna introduced the use of executed and model-generated experience for policy learning (Sutton, 1991). Neural world models extended this approach to learned simulated dynamics (Ha & Schmidhuber, 2018), and PlaNet and Dreamer

developed latent-space planning and imagination-based control (Hafner et al., 2019b;a). TD-MPC2 and DreamerV3 scale these ideas across continuous-control and multi-domain task suites (Hansen et al., 2024; Hafner et al., 2023). To address inaccurate predictions, ensembles estimate predictive uncertainty (Lakshminarayanan et al., 2017; Chua et al., 2018), short branch rollouts limit compounding model error (Janner et al., 2019), and confidence-based masking avoids unreliable model updates (Pan et al., 2020). Related model-based offline RL methods similarly restrict the influence of unreliable model predictions through uncertainty penalties or pessimistic dynamics (Yu et al., 2020; Kidambi et al., 2020).

Task and landscape representations. Characterizing an optimization task from sampled objective values has a long history in exploratory landscape analysis (ELA), where hand-designed landscape features are used to describe continuous black-box problems (Mersmann et al., 2011) and can support data-driven algorithm selection (Kerschke & Trautmann, 2019). Neural set encoders provide a complementary way to learn such representations directly from unordered observations. Deep Sets characterizes permutation-invariant set functions (Zaheer et al., 2017), while Set Transformer extends permutation-invariant modeling with attention between set elements (Lee et al., 2019). NeurELA learns attention-based exploratory landscape features and transfers dynamic landscape information across MetaBBO tasks (Ma et al., 2025a). Our encoder instead summarizes a fixed pre-optimization support set, enforces both sample- and coordinate-permutation invariance by construction, and remains frozen during controller training.

3 Proposed Method

3.1 Overview

RIME learns optimizer configuration from executed trajectories and predictions of how the search will evolve. Figure 1 shows the workflow. We first jointly train a task encoder and an objective surrogate. The encoder summarizes evaluated samples from each task, and the surrogate uses that context to predict candidate fitness.

The controller selects a configuration, whose execution produces a state transition and an improvement reward. These observations train a world-model ensemble that generates further experience for controller learning. Reliability gates select usable predictions and restrict rollout length. Section 3.2 describes task representation and optimizer interaction, and Section 3.3 explains dynamics learning and imagination. Appendix A gives the world-model background.

3.2 Task representation and optimizer interaction

Learning task context. A shared surrogate must adapt its predictions to the objective being approximated. We supply this information through a support set $\mathcal{D}_\tau = \{(x_i, y_i)\}_{i=1}^m$, where $y_i = f_\tau(x_i)$ and $m = 50$. Coordinates are normalized to $[0, 1]$, and objective values are standardized within each task. The encoder compresses these observations into a 64-dimensional task representation. Its parameters are shared across tasks and input dimensions.

The encoder aggregates information at two levels. A shared coordinate network first processes each coordinate together with its point's objective value. Mean and maximum pooling combine the coordinate features into a point representation; a second pooling step combines the evaluated points into a task representation. This construction accepts variable-size coordinate sets without dimension-specific encoder parameters. A scalar dimension token preserves information about problem size. The exact aggregation equations and permutation-invariance property are given in Appendices C.1 and B.1.

We train the encoder jointly with dimension-specific surrogate heads. For each task, the support set provides context and a query set $\mathcal{Q}_\tau$ supplies the prediction targets. The regression objective is

$$\min_{\psi,\omega} \mathbb{E}_{\tau,\mathcal{D}_\tau,\mathcal{Q}_\tau} \frac{1}{|\mathcal{Q}_\tau|} \sum_{(x,y)\in\mathcal{Q}_\tau} \left(S_{\omega,d}(x, E_\psi(\mathcal{D}_\tau, d)) - y\right)^2 . \tag{1}$$

Objective prediction trains the task representation before controller learning begins. We then fix the encoder and surrogate heads to keep the task representation and training objective stable as the

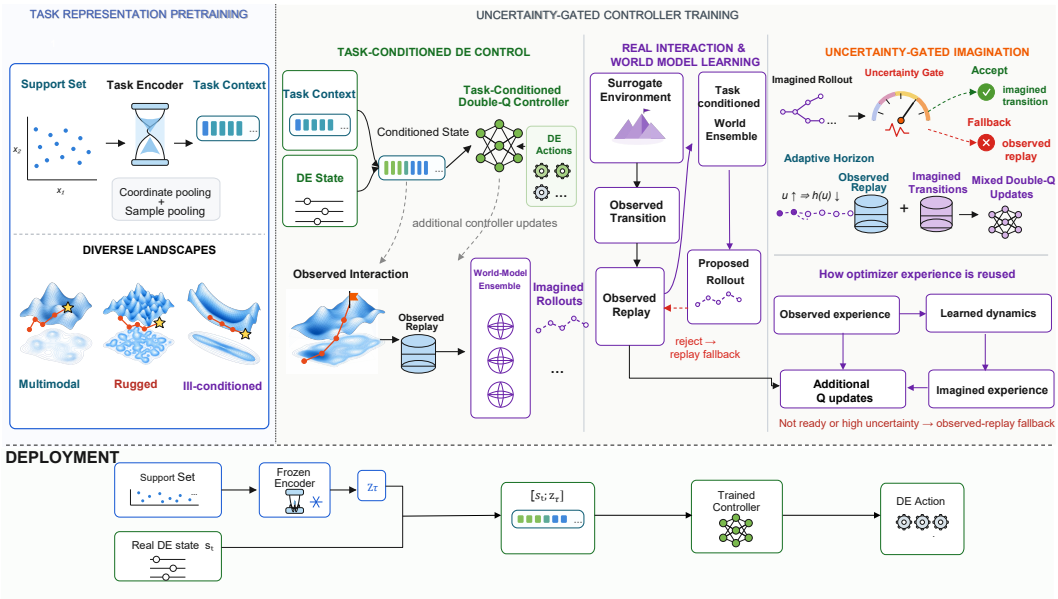

Figure 1: Overview of RIME training and deployment. Objective-surrogate pretraining learns a shared task representation. Executed optimizer transitions train the world model, and uncertainty-gated predictions supply imagined controller updates. Rejected predictions trigger a return to executed replay. Deployment retains the task encoder and controller.

controller changes. A support set defines the task context and surrogate objective:

$$z_\tau = E_\psi(\mathcal{D}_\tau, d_\tau), \qquad \widehat{f}_\tau(x) := S_{\omega,d_\tau}(x, z_\tau). \tag{2}$$

The task context conditions both the controller and world model. Section 5.1 specifies the training dimensions and function families.

Query-value prediction trains the encoder to retain objective information from a small evaluated sample. The same task description then supports configuration selection and dynamics prediction. Each replay transition carries its episode's context, which imagined branches inherit and retain as their predicted states evolve.

Collecting optimizer experience. We instantiate the configuration interface with DE and follow the action-space design of Surr-RLDE (Ma et al., 2025b). Each action selects a mutation operator and strength. The five operators are `rand/1`, `best/1`, `current-to-rand`, `current-to-pbest`, and `current-to-best`. Pairing each with $F \in \{0.1, 0.5, 0.9\}$ gives 15 actions. The crossover rate is fixed at $C_r = 0.7$.

At each generation, the controller observes nine dynamic features together with the task context. The dynamic features describe population diversity, distances and objective gaps relative to the best solutions, objective dispersion, and the progress of the search. They provide current optimization information, while the support-set representation remains fixed for the episode. The resulting 73-dimensional input is used to select the next configuration. Appendix C.5 lists the features and their normalization.

Executing the selected configuration generates candidate solutions, evaluates them with $\widehat{f}_\tau$, and applies DE selection. The reward is one when the best objective value improves and zero otherwise. The updated population supplies the next state, forming a replay record $(s_t, z_\tau, a_t, r_t, s_{t+1})$. We call these records *executed* or *real* experience to distinguish optimizer steps in the surrogate environment from world-model predictions. Appendix C.2 formalizes the transition.

3.3 WORLD-MODEL LEARNING AND UNCERTAINTY-GATED IMAGINATION

Predicting optimizer dynamics. The world model learns what follows a configuration decision from the replay records collected above. Each of its three ensemble members receives the current dynamic state, the task context, and a one-hot action encoding, and predicts a state change and an improvement-reward logit:

$$(\delta_t^{(k)}, \lambda_t^{(k)}) = W_k(s_t, z_\tau, \mathrm{onehot}(a_t)). \tag{3}$$

The state-change prediction specifies how the summarized search state evolves, while the reward head estimates the probability of finding an improved solution. Conditioning on z_τ allows the shared model to use task information when learning from trajectories pooled across objectives. During an imagined rollout, the dynamic state evolves and the task context stays fixed.

Each member is trained on an independently sampled minibatch of executed transitions. A Huber loss fits the observed state change, and binary cross-entropy fits the improvement reward. Bounded state-change predictions and reward probabilities are averaged across members to construct imagined transitions. Appendix C.3 gives the transformations and loss.

The objective surrogate and world model predict different quantities. The surrogate scores candidate solutions, after which DE executes selection and forms the next population. The world model instead predicts the resulting search-state change and reward directly from a state, action, and task context. An imagined transition therefore bypasses candidate generation and population selection. Its nine-dimensional prediction matches the controller's dynamic input even when the underlying decision-space dimension changes.

The ensemble also supplies the reliability signal used for imagination. We combine disagreement about state changes with disagreement about the improvement probability:

$$u(s, a, z) = \frac{1}{9}\sum_{q=1}^{9} \mathrm{Var}_k[0.25 \tanh \delta_q^{(k)}] + \mathrm{Var}_k[\sigma(\lambda^{(k)})]. \tag{4}$$

Here, σ is the sigmoid function, and variance is taken across ensemble members. The two terms measure disagreement in state changes and reward probability. We use their sum to filter predictions.

Selecting imagined experience. The global and local gates address different sources of unreliable predictions. The global gate delays imagination until the world model has received enough training and its moving-average state prediction loss falls below a threshold. The local gate then admits predictions when ensemble disagreement is at most τ_u, screening model use within the rollout.

These checks operate at different scales. Training loss summarizes how well the ensemble fits recently executed transitions, whereas disagreement is evaluated at the inputs used for imagination. A model can fit recent replay well while disagreeing on a particular branch. Conversely, agreement early in training can occur before the models have learned accurate transitions. Requiring both readiness and local agreement is intended to address these two situations.

Accepted predictions can be extended into a short rollout. To reduce reliance on uncertain multi-step predictions, the allowed length decreases with disagreement:

$$h(u) = \max\{1, \mathrm{round}(H_{\max}\,(1 - \min\{1, u/\tau_u\}))\}\,. \tag{5}$$

Low disagreement permits longer rollouts, while predictions close to the threshold receive shorter ones. The local gate is checked throughout each branch to limit continued use of uncertain predictions. Appendix C.3 lists the gate thresholds and warm-up requirements.

The controller uses Double-Q learning (Mnih et al., 2015; Van Hasselt et al., 2016). Each imagined branch starts from a replay state with its task context, and its predicted states and rewards supply controller targets. A fraction β of scheduled updates uses imagined experience. If the gates reject all predictions for an update, that update uses executed replay.

This scheduling rule keeps the number of controller updates fixed as acceptance and rollout length vary. Appendix C.4 gives the target equation and update accounting, and Algorithm 1 describes the training procedure.

Both experience sources use the same Double-Q target: the online network selects the next action and the target network evaluates it. Executed transitions use observed binary rewards, while imagined transitions use predicted improvement probabilities.

At deployment, 50 evaluations of the new objective construct the support set and count toward the evaluation budget. The encoder supplies task context to the controller, which configures DE using evaluations of the original objective.

Algorithm 1 Controller and world-model training for RIME

Input: frozen $E_\psi, S_{\omega,10}$; task distribution $\mathcal{T}_{\text{train}}$; G, β
initialize $Q_\theta, Q_{\bar{\theta}}$, ensemble $\{W_k\}_{k=1}^K$, and real replay $\mathcal{B}$
for each policy-training episode **do**
  sample $(\tau, \mathcal{D}_\tau) \sim \mathcal{T}_{\text{train}}$; set $z_\tau = E_\psi(\mathcal{D}_\tau, 10)$
  instantiate $\widehat{f}_\tau(x) = S_{\omega,10}(x, z_\tau)$; initialize P_0, s_0
  for each optimizer step **do**
    select a_t from $Q_\theta([s_t; z_\tau], \cdot)$ and apply its DE configuration
    evaluate candidates with $\widehat{f}_\tau$; select P_{t+1}; derive (s_{t+1}, r_t)
    store $(s_t, z_\tau, a_t, r_t, s_{t+1})$ in $\mathcal{B}$
    update each W_k on an independent minibatch from $\mathcal{B}$
    set $n_m \leftarrow \text{round}(\beta G)$ and $n_r \leftarrow G - n_m$
    for $g = 1, \ldots, G$ **do**
      if $g \leq n_r$ **then**
        update Q_θ from real replay
      else
        branch from $\mathcal{B}$ and propose up to $h(u)$ model transitions
        if the global gate is open and any proposed transition passes the local gate **then**
          perform one Q_θ update from the accepted transition(s)
        else
          perform one fallback Q_θ update from real replay
        end if
      end if
    end for
    periodically synchronize $Q_{\bar{\theta}}$; set $s_t \leftarrow s_{t+1}$
  end for
end for
return trained controller Q_θ

4 Analysis of Imagination Error

We bound the return error of imagined trajectories for a fixed world model and policy, using the executed surrogate-environment dynamics as the reference.

Proposition 1 (Finite-horizon imagination error)**.** *Fix a task context z, a policy π, and a horizon h. Let $V_h^\pi(s_0)$ and $\widehat{V}_h^\pi(s_0)$ be the expected h-step discounted returns under the executed surrogate-environment dynamics and the learned model, respectively. If the reward error plus the discounted transition error evaluated on the remaining true return is bounded by ϵ_i at every model-reachable state at depth i, then*

$$|\widehat{V}_h^\pi(s_0) - V_h^\pi(s_0)| \leq \sum_{i=0}^{h-1} \gamma^i \epsilon_i. \tag{6}$$

The bound connects one-step prediction errors with rollout length: smaller errors reduce each term, and shorter rollouts limit their accumulation. Relating ensemble disagreement to this bound requires a calibration assumption over all model-reachable states. Appendix B.3 states the condition and gives the proof. Appendix B.2 analyzes the bias of a fixed mixture of executed and model targets.

The proof accumulates discounted one-step discrepancies along the model rollout. Reward mismatch contributes directly, while transition mismatch changes the states from which future rewards and configuration decisions are evaluated. Shortening a branch limits the number of these contributions.

5 EXPERIMENTAL RESULTS

5.1 EXPERIMENTAL SETUP

Training Setting. Encoder and surrogate pretraining covers all 24 BBOB families at $d \in \{2, 5, 10\}$ with randomized shifts, rotations, and objective biases. Controller and world-model training uses the 10D surrogate head and F1–F15 and F20, matching the Surr-RLDE policy-training split (Ma et al., 2025b). The six (H, β) configurations are selected on these same 16 training families at $d = 10$ (Table 21). The selected setting is $H_{\max} = 5$ and $\beta = 0.5$.

Policy training uses a population size of 100 and three world-model ensemble members. After 64 warm-up transitions, each optimizer step triggers four Q updates. Thus, the 1.5M-update budget corresponds to 375,064 executed transitions per ablation (Equation 35).

Architectures and optimization. The Q network is a LayerNorm–MLP$(73, 64, 64, 15)$ with ReLU activations. Each world model uses an MLP trunk $(88, 256, 128)$ with separate nine-dimensional state-delta and scalar reward heads. Q-learning uses AdamW at 10^{-4}, Huber loss, batch size 64, target synchronization every 1,000 updates, gradient norm 10, and an epsilon schedule from 0.5 to 0.1. The world models use AdamW at 10^{-3}, gradient norm 10, and a replay capacity of 100,000 transitions.

Evaluation Setting. P-Test contains F16–F19 and F21–F24, which are held out from controller and world-model training, and is evaluated at $d \in \{2, 5, 10, 20, 30\}$. Each BBOB run uses a $2000d$ objective-evaluation budget, including the 50 support queries. We also evaluate 14 UAV tasks (Ma et al., 2026) under the benchmark's official budget. All comparisons use 10 random seeds.

Baselines and metrics. Surr-RLDE (Ma et al., 2025b) shares the surrogate training source and 16-family policy-training split. GLEET (Ma et al., 2024), DE-DDQN (Sharma et al., 2019), and DEDQN (Tan & Li, 2021) are MetaBBO baselines trained on original objectives. DE (Storn & Price, 1997), PSO (Kennedy & Eberhart, 1995), SAHLPSO (Tao et al., 2021), and random search are non-learning baselines. We average final objectives over the 10 evaluation seeds, rank methods within each task, and then average ranks over each reported task set. Appendix E reports the per-task values.

5.2 MAIN RESULTS

Table 1 summarizes two forms of generalization: cross-dimensional transfer on BBOB families and cross-domain transfer to UAV path planning.

Table 1: Cross-dimensional evaluation on P-Test and cross-domain transfer on 14 UAV tasks. Entries are average ranks over 10 seeds (lower is better); "Avg." averages the five P-Test dimensions.

	P-Test						Transfer
Method	2D	5D	10D	20D	30D	Avg.	UAV-14
RIME	**2.375**	**2.625**	**2.375**	2.875	**2.625**	**2.575**	2.714
DE	**2.375**	2.875	3.125	3.375	2.875	2.925	5.071
GLEET	3.500	3.375	2.750	**2.125**	2.875	2.925	**1.929**
Surr-RLDE	4.000	3.375	3.125	3.500	3.500	3.500	4.929
DE-DDQN	3.625	3.000	4.625	4.875	4.250	4.075	4.571
PSO	6.250	5.750	5.750	5.375	5.750	5.775	3.357
SAHLPSO	6.875	7.625	6.625	6.500	6.375	6.800	7.429
Random	7.000	7.375	7.625	7.375	7.750	7.425	6.143
DEDQN	9.000	9.000	9.000	9.000	9.000	9.000	8.857

Cross-dimensional generalization. The controller is trained only at 10D, yet RIME obtains the lowest average rank on P-Test across the five evaluated dimensions (2.575), compared with 2.925 for DE and GLEET and 3.500 for Surr-RLDE. More importantly, its advantage over the closely aligned Surr-RLDE baseline persists at every dimension. The 20D and 30D results are particularly informative because both dimensions lie outside controller training and encoder pretraining, showing

that the shared state–action interface can support extrapolation beyond the dimensionalities seen during learning.

This advantage is not universal across function distributions. On the complete 24-function BBOB suite in Appendix E, RIME has the lowest average rank at 2D, 5D, and 10D, whereas GLEET and DE lead at 20D and 30D, respectively. The results therefore support cross-dimensional transfer while also showing that high-dimensional performance depends on the evaluated function set.

Cross-domain transfer. The UAV benchmark (Ma et al., 2026) asks whether a policy learned from synthetic BBOB tasks remains useful on a structurally different optimization domain. RIME reaches an average rank of 2.714 and obtains lower final objectives than Surr-RLDE on 12 of 14 UAV tasks, although GLEET remains the strongest method on this benchmark. Importantly, the world model is discarded at deployment. The UAV improvement therefore reflects a controller shaped by imagination-augmented training rather than online model-based planning, suggesting that the additional predicted experience can improve the learned policy without introducing a world-model dependency at test time.

5.3 Ablation Studies

Table 2: Component ablation on the 24-function BBOB-10D suite. Entries are average ranks over functions (lower is better).

Variant	BBOB-10D average rank
Full	**1.833**
No world model	2.375
Imagined only	2.750
Real only	3.792
No task context	4.250

Table 3: Reliability-control ablation on the 24-function BBOB-10D suite. Lower average rank is better.

Variant	Global	Local	Horizon	Avg. rank	Firsts
Full	on	on	adaptive, ≤ 5	**1.792**	16
Fixed-H5	on	on	fixed 5	2.479	5
Ungated-WM	off	off	fixed 5	2.500	2
No-local	on	off	adaptive, ≤ 5	3.667	0
No-global	off	on	adaptive, ≤ 5	4.562	0

Task context and experience composition. Table 2 separates the roles of task awareness and experience reuse. Removing task context causes the largest degradation, increasing the average rank from 1.833 to 4.250. Because the same context conditions both the controller and the world model, this ablation measures its contribution to the coupled system rather than to either module in isolation.

The experience-source ablations reveal a second pattern: neither executed nor imagined experience alone matches the full method. Real-only gives up model-based reuse of collected trajectories, whereas Imagined-only removes the executed replay that anchors learning to observed optimizer transitions. Under the same executed-transition and Q-update budgets, adding the gated world-model pathway also improves over No world model, from 2.375 to 1.833. Together, these results indicate that imagination is most useful as a complement to, rather than a replacement for, executed optimizer experience.

Reliability-controlled imagination. Table 3 tests whether the gain comes merely from generating additional transitions or from controlling which predictions are trusted. The complete design performs best among the tested variants: fixing the horizon or removing uncertainty control both reduce performance. Removing the global or local gate individually causes an even larger degradation, despite retaining the adaptive horizon.

The single-gate results are non-monotonic: either individual gate removal performs worse than the ungated fixed-horizon variant. Because these variants alter interacting parts of the reliability mechanism, the comparison does not support attributing the gain independently to one gate. Instead, the evidence favors treating readiness checking, local disagreement filtering, and adaptive rollout depth as a coupled mechanism for deciding when and how far model predictions should influence controller learning.

Sensitivity to imagination budget. Sensitivity on the 16 controller-training families shows an interaction between rollout capacity and the scheduled imagined-update fraction. Increasing β from 0.25 to 0.50 improves average rank when $H = 3$ or $H = 5$, but degrades it when $H = 1$. Hence, the amount of imagined training cannot be selected independently of the rollout horizon. The selected $H = 5, \beta = 0.50$ setting has the lowest mean rank (2.250), while the uncertainty-dependent rule can still shorten individual branches below this maximum.

Does uncertainty identify unreliable predictions? Ensemble disagreement is moderately positively correlated with one-step state-prediction error (Pearson $r = 0.498$), indicating that it carries useful information about model reliability without implying calibrated uncertainty. The realized imagined-update fraction is 0.4339, below the scheduled value of 0.5, and the final moving-average rollout horizon is 1.425 despite a maximum of five. Thus, the gates actively suppress model use rather than simply permitting the maximum amount of imagination.

The rollout diagnostic provides a complementary view: state error changes little from depth one to five, while reward error increases, and both errors are larger at depth fifteen. Combined with the gate ablation, these observations connect the uncertainty signal to its intended role—limiting exposure to increasingly unreliable multi-step predictions. Appendix Table 24 and Figure 2 provide the diagnostic details.

6 Conclusion

We introduced RIME, a reliability-aware framework for augmenting MetaBBO training with task-conditioned imagined experience. A shared task representation provides context for both optimizer control and dynamics prediction, while a world-model ensemble reuses executed trajectories to generate additional training experience. Rather than trusting model predictions uniformly, RIME uses global readiness checking, local ensemble disagreement, and adaptive rollout horizons to determine when and how far imagined experience should influence controller learning, falling back to executed replay when predictions are unreliable. Our finite-horizon analysis further relates imagination error to one-step model errors and rollout length, providing a theoretical rationale for restricting unreliable long-horizon predictions.

Across BBOB functions spanning five dimensionalities, RIME achieves the lowest average rank among the evaluated methods and also transfers effectively to UAV path-planning tasks without retaining the world model at deployment. Ablation studies show that task context, executed experience, and imagined experience play complementary roles, while reliability control is important for realizing the benefit of model-generated transitions. These results suggest that optimizer trajectories can serve not only as direct training samples, but also as a reusable source of learned dynamics for MetaBBO. Extending this principle to broader optimizer families, structured search spaces, and better-calibrated uncertainty models offers a promising direction for future work.

AI use statement. Generative AI was used only to assist with language polishing. The authors verified all content and take full responsibility for the paper and its artifacts.

Ethics statement. The benchmarks use no personal data or human subjects; safety-critical applications require domain validation and human oversight.

Reproducibility statement. Appendices B–C give proofs and implementation details; Appendix E reports per-task outcomes, ablations, and model diagnostics. We also provide the core implementation code of the proposed algorithm to facilitate reproducibility.

REFERENCES

Janez Brest, Sao Greiner, Borko Boskovic, Marjan Mernik, and Viljem Zumer. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE transactions on evolutionary computation*, 10(6):646–657, 2006.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

Ágoston E Eiben, Robert Hinterding, and Zbigniew Michalewicz. Parameter control in evolutionary algorithms. *IEEE Transactions on evolutionary computation*, 3(2):124–141, 1999.

Alexander IJ Forrester and Andy J Keane. Recent advances in surrogate-based optimization. *Progress in aerospace sciences*, 45(1-3):50–79, 2009.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019b.

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

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Yaochu Jin. A comprehensive survey of fitness approximation in evolutionary computation. *Soft computing*, 9(1):3–12, 2005.

Yaochu Jin. Surrogate-assisted evolutionary computation: Recent advances and future challenges. *Swarm and Evolutionary Computation*, 1(2):61–70, 2011.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pp. 1942–1948. ieee, 1995.

Pascal Kerschke and Heike Trautmann. Automated algorithm selection on continuous black-box problems by combining exploratory landscape analysis and machine learning. *Evolutionary computation*, 27(1):99–127, 2019.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Robert Tjarko Lange, Tom Schaul, Yutian Chen, Tom Zahavy, Valentin Dalibard, Chris Lu, Satinder Singh, and Sebastian Flennerhag. Discovering evolution strategies via meta-black-box optimization. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=mFDU0fP3EQH`.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Ke Li and Jitendra Malik. Learning to optimize. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=ry4Vrt5gl.

Zeyuan Ma, Hongshu Guo, Jiacheng Chen, Zhenrui Li, Guojun Peng, Yue-Jiao Gong, Yining Ma, and Zhiguang Cao. Metabox: A benchmark platform for meta-black-box optimization with reinforcement learning. *Advances in Neural Information Processing Systems*, 36:10775–10795, 2023.

Zeyuan Ma, Jiacheng Chen, Hongshu Guo, Yining Ma, and Yue-Jiao Gong. Auto-configuring exploration-exploitation tradeoff in evolutionary computation via deep reinforcement learning. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1497–1505, 2024.

Zeyuan Ma, Jiacheng Chen, Hongshu Guo, and Yue-Jiao Gong. Neural exploratory landscape analysis for meta-black-box-optimization. In *International Conference on Learning Representations*, volume 2025, pp. 15373–15393, 2025a.

Zeyuan Ma, Zhiyang Huang, Jiacheng Chen, Zhiguang Cao, and Yue-Jiao Gong. Surrogate learning in meta-black-box optimization: A preliminary study. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1137–1145, 2025b.

Zeyuan Ma, Yue-Jiao Gong, Hongshu Guo, Wenjie Qiu, Sijie Ma, Hongqiao Lian, Jiajun Zhan, Kaixu Chen, Chen Wang, Zhiyang Huang, et al. Metabox-v2: A unified benchmark platform for meta-black-box optimization. *Advances in Neural Information Processing Systems*, 38, 2026.

Olaf Mersmann, Bernd Bischl, Heike Trautmann, Mike Preuss, Claus Weihs, and Günter Rudolph. Exploratory landscape analysis. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pp. 829–836, 2011.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Feiyang Pan, Jia He, Dandan Tu, and Qing He. Trust the model when it is confident: Masked model-based actor-critic. *Advances in neural information processing systems*, 33:10537–10546, 2020.

Mudita Sharma, Alexandros Komninos, Manuel López-Ibáñez, and Dimitar Kazakov. Deep reinforcement learning based parameter control in differential evolution. In *Proceedings of the genetic and evolutionary computation conference*, pp. 709–717, 2019.

Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

Zhiping Tan and Kangshun Li. Differential evolution with mixed mutation strategy based on deep reinforcement learning. *Applied Soft Computing*, 111:107678, 2021.

Ryoji Tanabe and Alex Fukunaga. Success-history based parameter adaptation for differential evolution. In *2013 IEEE congress on evolutionary computation*, pp. 71–78. IEEE, 2013.

Ryoji Tanabe and Alex S Fukunaga. Improving the search performance of shade using linear population size reduction. In *2014 IEEE congress on evolutionary computation (CEC)*, pp. 1658–1665. IEEE, 2014.

Xinmin Tao, Xiangke Li, Wei Chen, Tian Liang, Yetong Li, Jie Guo, and Lin Qi. Self-adaptive two roles hybrid learning strategies-based particle swarm optimization. *Information Sciences*, 578: 457–481, 2021.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in neural information processing systems*, 33:14129–14142, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Jingqiao Zhang and Arthur C Sanderson. Jade: adaptive differential evolution with optional external archive. *IEEE Transactions on evolutionary computation*, 13(5):945–958, 2009.

A World-model Background

In model-based reinforcement learning, a world model learns how an environment responds to an action from observed transitions (Sutton, 1991; Ha & Schmidhuber, 2018). Given a state s_t and action a_t, it predicts the next state and reward,

$$(\hat{s}_{t+1}, \hat{r}_t) = W_\phi(s_t, a_t). \tag{7}$$

Repeatedly applying this model produces an imagined trajectory that can supply training targets without executing each transition in the environment (Hafner et al., 2019a; Janner et al., 2019). Prediction errors can accumulate along the trajectory, making rollout length and model reliability important for policy learning.

For optimizer control, the state summarizes population and search progress, and the action specifies an optimizer configuration. The objective surrogate evaluates individual candidate solutions. The world model predicts the state and reward resulting from a configuration, as described in Section 3.3.

B Supporting Analysis and Proofs

B.1 Encoder permutation invariance

Proposition 2 (Sample- and coordinate-permutation invariance)**.** *For any fixed dimension d, let π_m permute support samples and π_d permute coordinates identically in every x_i. The encoder in Equation (27) satisfies*

$$E_\psi(\{(\pi_d x_{\pi_m(i)}, y_{\pi_m(i)})\}_{i=1}^m, d) = E_\psi(\mathcal{D}_\tau, d). \tag{8}$$

Proof. For a fixed sample i, let π_d permute the d coordinate tokens. Because ϕ_c is shared across coordinates,

$$\{\phi_c([x_{i,\pi_d(j)}, y_i])\}_{j=1}^d = \pi_d\{\phi_c([x_{ij}, y_i])\}_{j=1}^d. \tag{9}$$

Both coordinatewise mean and elementwise maximum are symmetric set functions, so c_i is unchanged. Consequently $g_i = \phi_s([c_i; y_i])$ is unchanged by π_d.

Now let π_m permute the support samples. Shared ϕ_s maps the permuted inputs to the same multiset $\{g_i\}_{i=1}^m$ in a different order. Mean and maximum over i are invariant to that order. Finally, $\ell(d)$ depends only on the cardinality of the coordinate set and is unchanged by either permutation. Applying ρ to identical pooled inputs gives identical z_τ, proving the result. □

B.2 Fixed-model mixed-target bias

The following analysis concerns an idealized operator on bounded action-value functions with bounded rewards and $0 \le \gamma < 1$. Here, the reference dynamics are those of the executed surrogate environment, not the original objective. Let $\mathcal{T}$ and $\widehat{\mathcal{T}}$ be the optimal Bellman operators of the real meta-MDP and a fixed world model. For a fixed Q-independent pointwise model weight $\alpha(s, a, z) \in [0, 1]$, define

$$\mathcal{T}_\alpha Q = (1-\alpha)\mathcal{T}Q + \alpha\widehat{\mathcal{T}}Q, \qquad \alpha_{\max} = \sup_{s,a,z} \alpha(s, a, z) \le 1. \tag{10}$$

Rejected inputs have $\alpha = 0$. Algorithm 1 uses β only as a *global* scheduling fraction; because real and imagined updates can occupy different state distributions, neither β nor β_{eff} is identified with $\alpha_{\max}$.

Proposition 3 (Contraction and state-dependent mixed-target bias)**.** *Assume that on inputs with $\alpha(s,a,z) > 0$, $|\widehat{\mathcal{T}}Q^*(s,a,z) - \mathcal{T}Q^*(s,a,z)| \le \epsilon_B$, where Q^* is the fixed point of $\mathcal{T}$. Then $\mathcal{T}_\alpha$ is a γ-contraction, and its fixed point Q_α satisfies*

$$\|Q_\alpha - Q^*\|_\infty \le \frac{\alpha_{\max}\epsilon_B}{1-\gamma}. \tag{11}$$

Rejected inputs use the reference operator pointwise. This does not imply that a fitted neural Q function is unaffected by model updates elsewhere.

Proof. For a fixed Q-independent map α, both $\mathcal{T}$ and $\widehat{\mathcal{T}}$ are γ-contractions in the sup norm. For arbitrary Q_1, Q_2 and every (s,a,z),

$$\begin{aligned} |\mathcal{T}_\alpha Q_1(s,a,z) - \mathcal{T}_\alpha Q_2(s,a,z)| &\le (1-\alpha(s,a,z))\gamma \|Q_1 - Q_2\|_\infty && (12)\\ &\quad + \alpha(s,a,z)\gamma \|Q_1 - Q_2\|_\infty && (13)\\ &= \gamma \|Q_1 - Q_2\|_\infty . && (14)\end{aligned}$$

Taking the supremum shows that $\mathcal{T}_\alpha$ is a γ-contraction and therefore has a unique fixed point. Using $Q^* = \mathcal{T}Q^*$ and $Q_\alpha = \mathcal{T}_\alpha Q_\alpha$,

$$\begin{aligned} \|Q_\alpha - Q^*\|_\infty &\le \|\mathcal{T}_\alpha Q_\alpha - \mathcal{T}_\alpha Q^*\|_\infty + \|\mathcal{T}_\alpha Q^* - \mathcal{T}Q^*\|_\infty && (15)\\ &\le \gamma \|Q_\alpha - Q^*\|_\infty + \sup_{s,a,z} \alpha(s,a,z) \left|\widehat{\mathcal{T}}Q^*(s,a,z) - \mathcal{T}Q^*(s,a,z)\right| && (16)\\ &\le \gamma \|Q_\alpha - Q^*\|_\infty + \alpha_{\max}\epsilon_B, && (17)\end{aligned}$$

where the last step uses the Bellman-error condition in Proposition 3 on inputs with $\alpha > 0$; inputs with $\alpha = 0$ contribute zero. Rearranging gives Equation (11). No relation between $\alpha_{\max}$ and the global scheduling fraction β is required. □

B.3 Finite-horizon imagination error

Fix a task context z, model parameters, a deterministic policy π, and a horizon h. Write P, R for the transition kernel and expected reward of the executed surrogate environment, and $\widehat{P}, \widehat{R}$ for the learned model. All expectations and returns below use this fixed policy and context. Let V_i^π and $\widehat{V}_i^\pi$ denote returns from depth i to h, with $V_h^\pi = \widehat{V}_h^\pi = 0$. Define the model-reachable sets by $\mathcal{C}_0 = \{s_0\}$ and

$$\mathcal{C}_{i+1} = \bigcup_{s \in \mathcal{C}_i} \operatorname{supp} \widehat{P}(\cdot \mid s, \pi(s,z), z). \tag{18}$$

Assume bounded rewards and, for every $s \in \mathcal{C}_i$ with $a = \pi(s,z)$,

$$|\widehat{R}(s,a,z) - R(s,a,z)| + \gamma \left|\mathbb{E}_{\widehat{P}} V_{i+1}^\pi - \mathbb{E}_P V_{i+1}^\pi\right| \le \epsilon_i. \tag{19}$$

The condition uses the actual remaining reference return V_{i+1}^π; it does not require a uniform bound over arbitrarily scaled continuation functions.

Proof of Proposition 1. Let $\Delta_i = \sup_{s \in \mathcal{C}_i} |\widehat{V}_i^\pi(s) - V_i^\pi(s)|$. Adding and subtracting $\gamma \mathbb{E}_{\widehat{P}} V_{i+1}^\pi$ in the Bellman recursions gives

$$\begin{aligned} |\widehat{V}_i^\pi(s) - V_i^\pi(s)| &\le |\widehat{R} - R| + \gamma |\mathbb{E}_{\widehat{P}} V_{i+1}^\pi - \mathbb{E}_P V_{i+1}^\pi| && (20)\\ &\quad + \gamma \mathbb{E}_{\widehat{P}} |\widehat{V}_{i+1}^\pi - V_{i+1}^\pi| && (21)\\ &\le \epsilon_i + \gamma \Delta_{i+1}. && (22)\end{aligned}$$

The successor distribution is supported on $\mathcal{C}_{i+1}$. Taking the supremum and using $\Delta_h = 0$ yields $\Delta_0 \le \sum_{i=0}^{h-1} \gamma^i \epsilon_i$, which proves Equation (6). □

Conditional connection to uncertainty gating. Suppose, additionally, that the left-hand side of Equation (19) is at most $c_0 + c_1\sqrt{u(s, \pi(s, z), z)}$ at every model-reachable state, where $c_0, c_1 \geq 0$. Set $u_i = \sup_{s \in \mathcal{C}_i} u(s, \pi(s, z), z)$. If $u_i \leq \tau_u$ for every $i < h$, then

$$|\widehat{V}_h^\pi(s_0) - V_h^\pi(s_0)| \leq (c_0 + c_1\sqrt{\tau_u})\frac{1-\gamma^h}{1-\gamma}. \tag{23}$$

Both calibration and the reachable-set threshold condition are assumptions. Acceptance of a single sampled branch does not establish them for a stochastic model. With our ensemble-mean transition and a fixed deterministic policy, each reachable set is a singleton, so checking every step of that branch checks the threshold condition for that branch; the calibration assumption is still required. The result concerns a fixed horizon and does not directly characterize the distribution induced by adaptive stopping or the evolving training process.

C Implementation Details

C.1 Task encoder construction

For normalized support samples (x_i, y_i), the shared coordinate network ϕ_c and point network ϕ_s compute

$$e_{ij} = \phi_c([x_{ij}, y_i]), \tag{24}$$

$$c_i = \left[\operatorname*{mean}_{j \leq d} e_{ij}; \max_{j \leq d} e_{ij}\right], \tag{25}$$

$$g_i = \phi_s([c_i; y_i]). \tag{26}$$

The task embedding is

$$z_\tau = \rho\left(\left[\operatorname*{mean}_{i \leq m} g_i; \max_{i \leq m} g_i; \ell(d_\tau)\right]\right), \quad \ell(d) = \frac{\log(1+d)}{\log(1001)}. \tag{27}$$

Here, ρ produces the task embedding and $\ell(d)$ supplies the dimension token. The encoder parameters are shared across dimensions; the surrogate heads remain dimension-specific.

C.2 Executed surrogate-environment transitions

Let $P_t \in \mathbb{R}^{N \times d_\tau}$ denote the DE population. The induced training environment is $\widehat{\mathcal{M}}_\tau = (\mathcal{S}, \mathcal{A}, \widehat{P}_\tau, \widehat{r}_\tau, \gamma)$. For population member n, random variation $\xi_{t,n}$ and the selected configuration a_t produce

$$x'_{t,n} = \mathrm{DE}(P_t, n; a_t, \xi_{t,n}), \qquad \widehat{y}'_{t,n} = \widehat{f}_\tau(x'_{t,n}), \tag{28}$$

$$P_{t+1} = \mathrm{Select}(P_t, \{(x'_{t,n}, \widehat{y}'_{t,n})\}_{n=1}^N), \qquad r_t = \mathbb{I}[\widehat{f}_\tau(P^\star_{t+1}) < \widehat{f}_\tau(P^\star_t)]. \tag{29}$$

The best-population member is denoted by $P^\star_t$. The transition is obtained by executing candidate generation and selection using surrogate fitness, rather than by querying the world model.

C.3 World-model output transformations and loss

Each member predicts the raw state-change vector $\delta_t^{(k)}$ and reward logit $\lambda_t^{(k)}$. We bound the former and apply a sigmoid to the latter before averaging:

$$\bar{\delta}_t = \frac{1}{K}\sum_{k=1}^K 0.25\tanh(\delta_t^{(k)}), \tag{30}$$

$$\hat{s}_{t+1} = \mathrm{clip}_{[0,1]}(s_t + \bar{\delta}_t), \tag{31}$$

$$\hat{p}_t = \frac{1}{K}\sum_{k=1}^K \sigma(\lambda_t^{(k)}). \tag{32}$$

Here, $\hat{p}_t$ is the expected binary reward used for imagined targets. For an executed transition with next state s' and reward r, the training loss is

$$\mathcal{L}_{\mathrm{WM}} = \frac{1}{K}\sum_{k=1}^{K}\left[\mathrm{Huber}(0.25\tanh\delta^{(k)}, s'-s) + \mathrm{BCE}(\lambda^{(k)}, r)\right]. \tag{33}$$

Independent replay minibatches train the ensemble members. The global gate opens after at least 1,000 world-model updates and an exponential moving average of the state-delta Huber loss of at most 0.02. The local disagreement threshold is $\tau_u = 0.01$. The selected maximum imagination horizon is $H_{\max} = 5$.

C.4 Controller targets and update accounting

Both executed and imagined updates use the task-conditioned Double-Q target

$$y = r + \gamma Q_{\bar{\theta}}\Big([s';z], \arg\max_{a'} Q_\theta([s';z],a')\Big), \qquad \gamma = 0.99. \tag{34}$$

The online network selects the next action and the target network evaluates it. With $G = 4$ updates per executed transition after $N_{\mathrm{warm}} = 64$ warm-up transitions,

$$N_{\mathrm{env}} = N_{\mathrm{warm}} + \frac{N_Q}{G} \tag{35}$$

At $N_Q = 1.5$M this yields $N_{\mathrm{env}} = 375{,}064$. The full model schedules βG imagined updates per environment step with $\beta = 0.5$. Each scheduled slot receives exactly one minibatch update from accepted model transitions, or one executed-replay update if none is admissible. Consequently $\beta_{\mathrm{eff}} \le \beta$. Rollout depth controls which model transitions are available, without changing the Q-update budget.

Algorithm 1 in Section 3.3 summarizes this update schedule.

C.5 Dynamic state

The nine dynamic features are: (1) mean pairwise population distance; (2) mean distance to the current best; (3) mean distance to the historical best; (4) mean objective gap to the historical best; (5) mean objective gap to the current best; (6) objective standard deviation; (7) remaining FE fraction; (8) normalized stagnation length; and (9) a binary improvement flag. Distance features are divided by the box diagonal. Objective gaps and dispersion use within-population scale normalization. All dynamic features are clipped or transformed to $[0,1]$, and the task embedding is passed through $\tanh$ before concatenation.

C.6 Training-task construction

Each controller-training episode samples a BBOB family, translation, orthogonal rotation, and objective bias. A support set determines z_τ and $\widehat{f}_\tau$ through Equation (2). Candidate generation, surrogate evaluation, and selection produce the executed transitions in Equation (29), from which the world ensemble and Q networks learn online.

D Additional Experimental Notes

The supplementary BBOB evaluation covers all 24 functions at five dimensions, giving 120 function–dimension conditions. Each condition uses the same 10 evaluation seeds: $\{33629, 190681, 200182, 241870, 289561, 335970, 760064, 949631, 962973, 971186\}$. Best counts denote the number of task-level first-place outcomes. The six-configuration sensitivity comparison uses the 16 controller-training families at $d = 10$, with a budget of 20,000 objective evaluations including 50 support queries. Figure 2 reports one-step prediction errors, update acceptance, and rollout-depth effects.

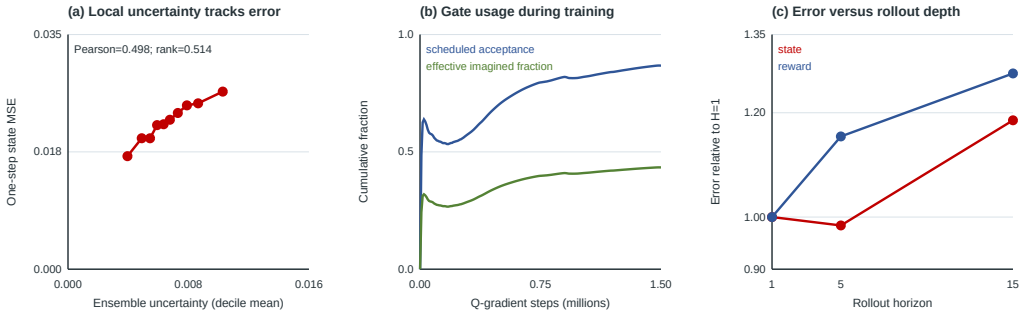

Figure 2: Descriptive diagnostics from the selected H5-F50 training run. (a) Decile-binned ensemble disagreement versus one-step state MSE. (b) Cumulative acceptance among scheduled imagined updates and the realized imagined-update fraction. (c) state and reward errors at each diagnostic rollout depth, normalized by their depth-1 values. Exact values are in Table 24.

E Complete Experimental Results

Unless stated otherwise, each cell reports the mean final objective error over 10 evaluation seeds, with the standard deviation in parentheses. Lower is better. Bold marks the lowest mean across all methods in a comparison, including when the table spans multiple panels. Task IDs follow the benchmark ordering. For P-Test, P1–P8 denote Weierstrass, Schaffers, high-conditioned Schaffers, composite Griewank–Rosenbrock, Gallagher-101, Gallagher-21, Katsuura, and Lunacek bi-Rastrigin, respectively.

E.1 Cross-dimensional BBOB results

Across the complete 24-function suite, RIME has the lowest average rank at 2D, 5D, and 10D. These dimensions are covered by encoder pretraining, with controller training confined to 10D. At the extrapolation dimensions, GLEET leads at 20D and DE at 30D. This pattern places the strongest aggregate performance within the dimensions represented during pretraining.

BBOB at $d = 2$. At 2D, RIME has average rank 1.938, compared with 2.646 for DE and 3.562 for GLEET.

Table 4: BBOB-2D mean final error over 10 evaluation seeds, panel A. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	GLEET	DE-DDQN	Surr-RLDE	DEDQN
F1	**0** (0)	8.66E-07 (1.09E-06)	4.34E-09 (3.14E-09)	**0** (0)	2.30E-01 (1.84E-01)
F2	**2.40E-12** (4.44E-12)	1.05E-01 (2.33E-01)	7.49E-09 (2.20E-09)	5.90E+00 (7.09E+00)	1.44E+03 (1.64E+03)
F3	**9.69E-12** (2.34E-11)	4.09E-03 (5.95E-03)	3.60E-03 (4.59E-03)	1.89E+00 (2.17E+00)	8.66E+00 (4.54E+00)
F4	**6.55E-04** (2.03E-03)	2.97E-03 (3.93E-03)	2.62E-02 (3.01E-02)	1.49E+00 (1.50E+00)	1.28E+01 (6.64E+00)
F5	6.16E-02 (1.95E-01)	**0** (0)	**0** (0)	**0** (0)	6.69E-01 (1.95E+00)
F6	5.59E-09 (7.37E-09)	1.19E-04 (1.09E-04)	2.90E-08 (2.83E-08)	**1.59E-12** (0)	2.66E+00 (2.98E+00)
F7	**4.41E-10** (4.51E-10)	4.78E-09 (2.93E-09)	4.08E-09 (2.63E-09)	4.02E-02 (1.27E-01)	1.02E+00 (4.57E-01)
F8	**2.46E-04** (3.14E-04)	1.67E-03 (4.40E-03)	2.60E-04 (8.19E-04)	1.36E-01 (2.48E-01)	3.42E+00 (2.69E+00)
F9	**0** (0)	6.38E-05 (6.37E-05)	1.30E-05 (4.10E-05)	1.68E-01 (2.75E-01)	4.36E+00 (5.58E+00)
F10	**1.62E-11** (4.13E-11)	4.74E-02 (1.29E-01)	2.31E-08 (3.36E-08)	6.84E+00 (7.66E+00)	6.28E+03 (1.73E+04)
F11	**1.22E-11** (2.14E-11)	1.01E+00 (1.15E+00)	1.72E-02 (5.42E-02)	9.46E+00 (1.93E+01)	6.60E+03 (1.42E+04)
F12	2.09E-04 (4.45E-04)	1.32E-01 (3.22E-01)	1.15E-02 (3.56E-02)	2.40E+00 (3.22E+00)	2.75E+03 (3.35E+03)
F13	4.68E-07 (6.26E-07)	1.87E-02 (1.37E-02)	6.41E-07 (1.18E-06)	5.62E+00 (6.85E+00)	1.17E+01 (1.09E+01)
F14	5.73E-07 (1.24E-06)	5.19E-05 (3.19E-05)	6.11E-09 (2.66E-09)	5.69E-05 (7.24E-05)	3.44E-01 (2.17E-01)
F15	**8.19E-11** (2.55E-10)	3.79E-03 (3.45E-03)	6.80E-03 (1.00E-02)	2.69E+00 (2.30E+00)	9.98E+00 (3.65E+00)
F16	8.51E-03 (6.60E-03)	8.92E-03 (1.59E-02)	2.66E-02 (2.79E-02)	1.99E-02 (3.95E-02)	2.25E+00 (1.74E+00)
F17	3.33E-05 (7.05E-05)	3.58E-03 (2.00E-03)	2.09E-04 (3.09E-04)	5.86E-03 (1.82E-02)	1.58E+00 (7.98E-01)
F18	1.10E-03 (2.58E-03)	1.38E-02 (1.69E-02)	3.32E-03 (5.34E-03)	1.01E+00 (1.66E+00)	5.67E+00 (4.32E+00)
F19	**1.11E-04** (3.44E-04)	1.57E-03 (4.97E-03)	3.58E-03 (6.25E-03)	9.43E-03 (8.11E-03)	2.71E-01 (3.51E-01)
F20	**4.39E-05** (1.27E-04)	2.34E-04 (2.92E-04)	6.00E-03 (1.87E-02)	2.86E-01 (5.47E-01)	2.08E+00 (8.59E-01)
F21	4.49E-09 (1.10E-08)	3.01E-09 (3.60E-09)	3.11E-09 (3.11E-09)	**0** (0)	6.94E-01 (7.37E-01)
F22	1.82E-11 (4.27E-11)	9.82E-08 (2.98E-07)	2.26E-09 (1.88E-09)	**0** (0)	4.06E-01 (7.96E-01)
F23	1.98E-01 (9.58E-02)	5.93E-01 (2.60E-01)	1.48E-01 (1.25E-01)	3.32E-01 (2.59E-01)	3.88E+00 (1.74E+00)
F24	9.77E-01 (8.70E-01)	**7.93E-01** (9.42E-01)	1.35E+00 (7.36E-01)	1.20E+00 (1.01E+00)	5.46E+00 (2.39E+00)

Table 5: BBOB-2D mean final error over 10 evaluation seeds, panel B. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	DE	PSO	SAHLPSO	Random
F1	**0** (0)	4.37E-09 (3.03E-09)	2.38E-04 (2.41E-04)	7.56E-03 (8.71E-03)	9.33E-03 (1.10E-02)
F2	**2.40E-12** (4.44E-12)	3.31E-09 (3.27E-09)	5.19E+00 (7.53E+00)	6.51E+00 (7.99E+00)	4.80E+00 (3.26E+00)
F3	**9.69E-12** (2.34E-11)	9.95E-02 (3.15E-01)	6.09E-01 (3.73E-01)	1.43E+00 (1.58E+00)	1.29E+00 (6.78E-01)
F4	**6.55E-04** (2.03E-03)	6.16E-01 (4.91E-01)	9.28E-01 (6.41E-01)	2.26E+00 (1.66E+00)	2.08E+00 (1.32E+00)
F5	6.16E-02 (1.95E-01)	**0** (0)	**0** (0)	**0** (0)	6.03E-01 (2.71E-01)
F6	5.59E-09 (7.37E-09)	6.28E-09 (2.66E-09)	9.99E-03 (6.48E-03)	2.63E-02 (5.92E-02)	1.64E-01 (1.89E-01)
F7	**4.41E-10** (4.51E-10)	6.75E-09 (2.95E-09)	3.72E-03 (5.11E-03)	6.49E-03 (1.04E-02)	1.97E-02 (1.92E-02)
F8	**2.46E-04** (3.14E-04)	8.67E-04 (2.74E-03)	6.87E-03 (2.74E-03)	4.61E-02 (4.34E-02)	6.13E-02 (2.97E-02)
F9	**0** (0)	7.03E-09 (2.24E-09)	7.04E-03 (3.94E-03)	3.36E-02 (6.88E-02)	1.11E-01 (9.21E-02)
F10	**1.62E-11** (4.13E-11)	5.63E-09 (3.06E-09)	5.32E+00 (5.21E+00)	9.19E+00 (1.41E+01)	1.40E+01 (1.32E+01)
F11	**1.22E-11** (2.14E-11)	1.10E-07 (3.30E-07)	1.27E+01 (2.15E+01)	2.37E+01 (2.80E+01)	8.41E+00 (7.20E+00)
F12	2.09E-04 (4.45E-04)	**2.28E-05** (7.21E-05)	6.65E-01 (7.58E-01)	2.25E+00 (2.15E+00)	1.13E+01 (1.09E+01)
F13	4.68E-07 (6.26E-07)	**7.19E-09** (1.74E-09)	5.03E-01 (3.68E-01)	1.09E+00 (1.05E+00)	2.01E+00 (1.75E+00)
F14	5.73E-07 (1.24E-06)	**5.14E-09** (1.95E-09)	4.10E-03 (2.75E-03)	7.46E-03 (7.90E-03)	1.91E-02 (9.24E-03)
F15	**8.19E-11** (2.55E-10)	9.95E-02 (3.15E-01)	4.13E-01 (3.20E-01)	1.18E+00 (8.48E-01)	1.35E+00 (3.99E-01)
F16	8.51E-03 (6.60E-03)	**7.57E-03** (1.23E-02)	8.10E-02 (1.01E-01)	1.65E-01 (2.70E-01)	7.90E-02 (3.66E-02)
F17	3.33E-05 (7.05E-05)	**4.84E-09** (2.22E-09)	3.60E-02 (1.84E-02)	8.03E-02 (9.28E-02)	2.51E-01 (1.30E-01)
F18	1.10E-03 (2.58E-03)	**6.62E-09** (2.55E-09)	2.17E-01 (1.30E-01)	1.07E+00 (9.25E-01)	5.82E-01 (2.88E-01)
F19	**1.11E-04** (3.44E-04)	4.98E-03 (8.82E-03)	2.50E-03 (5.50E-03)	7.92E-03 (1.20E-02)	2.32E-02 (1.24E-02)
F20	**4.39E-05** (1.27E-04)	5.92E-02 (1.87E-01)	1.08E-01 (1.90E-01)	3.00E-01 (3.10E-01)	4.70E-01 (2.57E-01)
F21	4.49E-09 (1.10E-08)	1.78E-09 (2.69E-09)	6.92E-02 (2.19E-01)	6.92E-02 (2.19E-01)	7.79E-04 (1.11E-03)
F22	1.82E-11 (4.27E-11)	3.36E-09 (3.27E-09)	7.02E-06 (1.21E-05)	1.51E-05 (3.09E-05)	4.19E-04 (4.05E-04)
F23	1.98E-01 (9.58E-02)	**7.74E-02** (1.68E-01)	6.33E-01 (2.74E-01)	4.47E-01 (2.61E-01)	7.71E-01 (2.53E-01)
F24	9.77E-01 (8.70E-01)	1.32E+00 (1.02E+00)	1.96E+00 (7.67E-01)	1.76E+00 (8.39E-01)	1.58E+00 (8.08E-01)

BBOB at $d = 5$. At 5D, RIME leads with average rank 2.396, followed by DE-DDQN at 2.562.

Table 6: BBOB-5D mean final error over 10 evaluation seeds, panel A. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	GLEET	DE-DDQN	Surr-RLDE	DEDQN
F1	8.11E-08 (2.26E-07)	1.11E-06 (6.86E-07)	**6.34E-09** (1.80E-09)	1.04E-01 (1.41E-01)	6.35E+00 (4.07E+00)
F2	6.87E-03 (1.21E-02)	1.12E+01 (1.05E+01)	**7.67E-09** (2.68E-09)	5.56E+02 (1.54E+03)	4.83E+04 (4.41E+04)
F3	**1.89E+00** (5.65E-01)	2.39E+00 (1.46E+00)	5.07E+00 (1.57E+00)	8.88E+00 (4.47E+00)	5.22E+01 (1.67E+01)
F4	**3.90E+00** (1.50E+00)	4.49E+00 (1.77E+00)	7.05E+00 (2.26E+00)	1.38E+01 (5.72E+00)	8.49E+01 (2.98E+01)
F5	**0** (0)	**0** (0)	**0** (0)	1.36E-01 (2.94E-01)	6.30E+00 (6.82E+00)
F6	**4.27E-04** (1.35E-03)	3.63E-03 (3.85E-03)	1.49E+00 (1.63E+00)	6.01E+00 (4.07E+00)	5.88E+01 (2.51E+01)
F7	2.28E-01 (3.04E-01)	4.84E-02 (1.53E-01)	**5.67E-09** (2.66E-09)	1.37E+00 (7.67E-01)	1.79E+01 (1.03E+01)
F8	3.32E-01 (4.00E-01)	**1.05E-01** (1.12E-01)	1.08E+00 (7.68E-01)	1.15E+01 (2.09E+01)	1.14E+03 (8.73E+02)
F9	1.32E+00 (1.54E+00)	**7.78E-01** (1.13E+00)	1.22E+00 (9.07E-01)	1.51E+01 (1.08E+01)	8.09E+02 (4.66E+02)
F10	2.43E-04 (6.24E-04)	9.32E+00 (1.35E+01)	**3.76E-08** (9.67E-08)	4.71E+02 (8.70E+02)	5.06E+04 (4.40E+04)
F11	3.51E-04 (9.44E-04)	1.59E+00 (1.33E+00)	**6.32E-09** (2.90E-09)	5.76E+00 (5.65E+00)	4.56E+02 (5.65E+02)
F12	7.74E-01 (1.91E+00)	3.05E+00 (4.17E+00)	**6.28E-03** (8.07E-03)	5.31E+03 (1.14E+04)	4.70E+06 (3.22E+06)
F13	3.59E+00 (4.23E+00)	1.34E+01 (2.27E+01)	**5.60E-07** (2.13E-07)	6.68E+00 (5.44E+00)	3.39E+02 (9.86E+01)
F14	8.64E-07 (1.66E-06)	1.88E-04 (1.43E-04)	**5.84E-07** (1.62E-06)	4.82E-02 (7.64E-02)	4.19E+00 (1.59E+00)
F15	4.06E+00 (3.46E+00)	2.68E+00 (1.11E+00)	5.99E+00 (1.52E+00)	1.00E+01 (5.94E+00)	6.13E+01 (1.09E+01)
F16	1.88E+00 (1.24E+00)	**1.45E-01** (1.83E-01)	2.50E+00 (1.22E+00)	1.80E+00 (9.84E-01)	8.78E+00 (4.02E+00)
F17	1.68E-02 (1.91E-02)	1.79E-02 (1.92E-02)	1.03E-04 (4.42E-05)	3.72E-01 (2.68E-01)	4.67E+00 (9.19E-01)
F18	7.24E-02 (1.15E-01)	1.02E-01 (7.47E-02)	**2.47E-04** (1.31E-04)	1.55E+00 (1.31E+00)	1.63E+01 (4.46E+00)
F19	5.10E-01 (1.10E-01)	4.64E-01 (1.94E-01)	5.19E-01 (2.03E-01)	**3.98E-01** (1.36E-01)	6.85E+00 (2.07E+00)
F20	2.49E-01 (2.09E-01)	5.90E-01 (2.61E-01)	1.04E+00 (4.63E-01)	1.51E+00 (5.86E-01)	3.06E+02 (8.76E+02)
F21	3.64E-01 (1.15E+00)	7.84E-01 (7.83E-01)	6.05E-01 (5.79E-01)	**4.77E-02** (1.02E-01)	9.12E+00 (3.46E+00)
F22	**7.40E-06** (1.82E-05)	8.06E-02 (2.18E-01)	1.38E-01 (2.92E-01)	2.00E-01 (6.17E-01)	1.37E+01 (6.11E+00)
F23	**8.99E-01** (2.55E-01)	1.04E+00 (1.51E-01)	9.08E-01 (1.87E-01)	9.84E-01 (2.60E-01)	2.77E+00 (1.07E+00)
F24	9.26E+00 (3.08E+00)	9.64E+00 (2.16E+00)	9.03E+00 (1.37E+00)	9.90E+00 (4.25E+00)	5.09E+01 (6.80E+00)

Table 7: BBOB-5D mean final error over 10 evaluation seeds, panel B. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	DE	PSO	SAHLPSO	Random
F1	8.11E-08 (2.26E-07)	7.52E-09 (2.22E-09)	1.22E-01 (3.30E-02)	5.81E-01 (3.17E-01)	1.04E+00 (3.99E-01)
F2	6.87E-03 (1.21E-02)	2.15E+00 (5.32E+00)	8.17E+02 (5.36E+02)	3.73E+03 (3.78E+03)	3.04E+03 (2.11E+03)
F3	**1.89E+00** (5.65E-01)	2.51E+00 (1.10E+00)	1.25E+01 (4.28E+00)	2.39E+01 (6.84E+00)	2.60E+01 (8.51E+00)
F4	**3.90E+00** (1.50E+00)	4.80E+00 (1.52E+00)	1.87E+01 (5.63E+00)	3.00E+01 (1.07E+01)	3.38E+01 (9.31E+00)
F5	**0** (0)	**0** (0)	**0** (0)	**0** (0)	1.15E+01 (2.75E+00)
F6	**4.27E-04** (1.35E-03)	5.16E-01 (1.23E+00)	1.65E+00 (7.45E-01)	1.40E+01 (7.77E+00)	1.20E+01 (4.90E+00)
F7	2.28E-01 (3.04E-01)	4.97E-04 (1.57E-03)	5.36E-01 (2.16E-01)	2.54E+00 (1.72E+00)	3.88E+00 (1.39E+00)
F8	3.32E-01 (4.00E-01)	6.03E-01 (4.26E-01)	9.01E+00 (3.37E+00)	7.50E+01 (6.23E+01)	7.67E+01 (4.82E+01)
F9	1.32E+00 (1.54E+00)	1.53E+00 (1.03E+00)	8.32E+00 (3.09E+00)	6.51E+01 (5.00E+01)	8.99E+01 (4.50E+01)
F10	2.43E-04 (6.24E-04)	1.24E+01 (2.68E+01)	5.75E+02 (2.87E+02)	2.05E+03 (2.38E+03)	2.02E+03 (1.18E+03)
F11	3.51E-04 (9.44E-04)	5.79E-01 (7.99E-01)	8.15E+00 (6.80E+00)	2.91E+01 (1.96E+01)	2.10E+01 (1.06E+01)
F12	7.74E-01 (1.91E+00)	8.09E-01 (1.12E+00)	3.82E+04 (2.77E+04)	3.61E+05 (3.40E+05)	5.73E+05 (2.33E+05)
F13	3.59E+00 (4.23E+00)	1.48E-02 (2.87E-02)	4.54E+01 (1.38E+01)	1.10E+02 (7.04E+01)	1.17E+02 (4.07E+01)
F14	8.64E-07 (1.66E-06)	2.29E-05 (3.69E-05)	1.25E-01 (6.82E-02)	5.45E-01 (5.38E-01)	7.41E-01 (2.61E-01)
F15	4.06E+00 (3.46E+00)	**2.21E+00** (6.24E-01)	1.24E+01 (2.96E+00)	3.12E+01 (9.30E+00)	2.88E+01 (7.72E+00)
F16	1.88E+00 (1.24E+00)	6.88E-01 (5.67E-01)	1.80E+00 (1.15E+00)	3.64E+00 (2.64E+00)	3.06E+00 (8.90E-01)
F17	1.68E-02 (1.91E-02)	**9.93E-07** (1.28E-06)	6.24E-01 (2.04E-01)	9.18E-01 (7.41E-01)	1.75E+00 (6.15E-01)
F18	7.24E-02 (1.15E-01)	8.44E-03 (1.33E-02)	2.07E+00 (4.62E-01)	5.02E+00 (4.09E+00)	5.74E+00 (1.62E+00)
F19	5.10E-01 (1.10E-01)	5.68E-01 (1.27E-01)	6.99E-01 (3.27E-01)	2.39E+00 (7.14E-01)	1.98E+00 (5.78E-01)
F20	2.49E-01 (2.09E-01)	**6.80E-02** (1.50E-01)	1.51E+00 (4.13E-01)	2.40E+00 (7.52E-01)	2.63E+00 (2.67E-01)
F21	3.64E-01 (1.15E+00)	5.81E-01 (1.10E+00)	9.63E-01 (1.38E+00)	1.79E+00 (1.48E+00)	1.19E+00 (7.23E-01)
F22	**7.40E-06** (1.82E-05)	5.33E-01 (8.03E-01)	6.32E-01 (8.35E-01)	2.53E+00 (2.79E+00)	1.13E+00 (9.26E-01)
F23	**8.99E-01** (2.55E-01)	1.02E+00 (2.28E-01)	1.05E+00 (2.94E-01)	1.42E+00 (3.24E-01)	1.11E+00 (2.23E-01)
F24	9.26E+00 (3.08E+00)	**8.99E+00** (2.53E+00)	1.40E+01 (3.26E+00)	2.31E+01 (8.31E+00)	2.67E+01 (3.08E+00)

BBOB at $d = 10$. At the controller-training dimension, RIME ranks 2.458 on the full 24-function suite, ahead of GLEET at 2.729.

Table 8: BBOB-10D mean final error over 10 evaluation seeds, panel A. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	GLEET	DE-DDQN	Surr-RLDE	DEDQN
F1	2.14E-02 (5.02E-02)	4.34E-06 (4.10E-06)	**7.81E-09** (1.11E-09)	8.93E-01 (5.40E-01)	2.59E+01 (8.97E+00)
F2	**2.13E+01** (2.53E+01)	3.55E+02 (3.39E+02)	1.46E+02 (3.08E+02)	1.43E+03 (2.57E+03)	2.57E+05 (1.57E+05)
F3	**1.33E+01** (8.57E+00)	1.95E+01 (8.32E+00)	2.89E+01 (4.52E+00)	2.54E+01 (9.83E+00)	1.80E+02 (2.43E+01)
F4	**2.00E+01** (1.07E+01)	2.43E+01 (9.46E+00)	3.27E+01 (7.95E+00)	4.62E+01 (2.84E+01)	3.14E+02 (2.65E+01)
F5	8.43E+00 (1.03E+01)	**0** (0)	9.05E-09 (1.04E-09)	4.91E+00 (2.97E+00)	4.77E+01 (1.15E+01)
F6	1.17E-01 (1.15E-01)	1.20E+01 (9.11E+00)	2.14E+00 (6.78E+00)	3.58E+01 (2.26E+01)	7.93E+04 (4.18E+04)
F7	1.25E-01 (3.63E-01)	7.51E-01 (5.24E-01)	**5.60E-09** (3.05E-09)	7.84E+00 (8.98E+00)	1.18E+02 (3.77E+01)
F8	1.99E+01 (3.27E+01)	1.98E+01 (2.91E+01)	**8.11E-01** (2.56E+00)	2.12E+02 (9.14E+01)	1.18E+04 (5.99E+03)
F9	7.47E+00 (2.13E+00)	6.75E+00 (1.72E+00)	**3.52E+00** (4.62E+00)	1.33E+02 (7.89E+01)	1.44E+04 (7.33E+03)
F10	**3.94E+01** (2.92E+01)	5.44E+02 (7.86E+02)	6.99E+01 (1.55E+02)	3.34E+02 (4.32E+02)	4.51E+05 (2.27E+05)
F11	**1.03E+00** (2.08E+00)	1.01E+01 (6.91E+00)	2.81E+00 (4.34E+00)	3.41E+01 (2.49E+01)	1.14E+03 (1.54E+03)
F12	1.22E+04 (3.85E+04)	1.31E+04 (4.13E+04)	1.54E+05 (3.63E+05)	1.33E+06 (1.72E+06)	3.37E+07 (1.40E+07)
F13	2.99E+00 (8.38E+00)	1.02E+01 (1.36E+01)	**5.68E-07** (2.63E-07)	8.08E+01 (4.83E+01)	8.93E+02 (1.16E+02)
F14	1.07E-02 (3.35E-02)	6.54E-04 (3.75E-04)	8.04E-03 (1.96E-02)	3.21E-01 (3.17E-01)	1.27E+01 (4.47E+00)
F15	1.67E+01 (1.21E+01)	**1.49E+01** (7.86E+00)	2.78E+01 (4.30E+00)	2.25E+01 (9.45E+00)	1.92E+02 (2.82E+01)
F16	6.41E+00 (3.79E+00)	**1.39E+00** (1.13E+00)	1.00E+01 (1.61E+00)	6.80E+00 (1.33E+00)	2.26E+01 (6.02E+00)
F17	7.24E-02 (1.53E-01)	4.35E-01 (6.26E-01)	1.19E-02 (3.34E-02)	1.04E+00 (3.80E-01)	9.19E+00 (1.98E+00)
F18	6.42E-02 (5.08E-02)	1.49E+00 (7.55E-01)	**5.70E-03** (1.49E-02)	3.28E+00 (8.40E-01)	3.05E+01 (7.86E+00)
F19	2.07E+00 (3.91E-01)	1.79E+00 (3.29E-01)	2.15E+00 (6.60E-01)	**1.41E+00** (1.03E+00)	1.18E+01 (2.88E+00)
F20	1.00E+00 (5.46E-01)	1.41E+00 (3.58E-01)	1.64E+00 (4.00E-01)	1.70E+00 (2.78E-01)	5.48E+03 (2.80E+03)
F21	**1.70E+00** (2.52E+00)	3.19E+00 (2.64E+00)	3.41E+00 (3.26E+00)	2.64E+00 (2.95E+00)	4.27E+01 (8.22E+00)
F22	**2.08E-01** (6.25E-01)	2.19E+00 (4.17E+00)	3.23E+00 (3.86E+00)	4.23E-01 (8.03E-01)	5.62E+01 (1.74E+01)
F23	1.40E+00 (1.97E-01)	**1.29E+00** (2.97E-01)	1.44E+00 (2.37E-01)	1.41E+00 (2.50E-01)	3.61E+00 (5.79E-01)
F24	3.38E+01 (7.03E+00)	3.40E+01 (7.41E+00)	4.21E+01 (3.14E+00)	**3.02E+01** (1.11E+01)	1.46E+02 (2.35E+01)

Table 9: BBOB-10D mean final error over 10 evaluation seeds, panel B. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	DE	PSO	SAHLPSO	Random
F1	2.14E-02 (5.02E-02)	8.66E-09 (1.51E-09)	1.66E+00 (6.74E-01)	4.98E+00 (2.67E+00)	1.16E+01 (2.17E+00)
F2	**2.13E+01** (2.53E+01)	1.12E+03 (8.81E+02)	1.05E+04 (5.21E+03)	4.17E+04 (5.02E+04)	4.82E+04 (1.10E+04)
F3	**1.33E+01** (8.57E+00)	2.62E+01 (5.18E+00)	6.09E+01 (1.06E+01)	1.00E+02 (1.78E+01)	1.11E+02 (1.38E+01)
F4	**2.00E+01** (1.07E+01)	3.17E+01 (7.50E+00)	8.94E+01 (9.10E+00)	1.29E+02 (2.68E+01)	1.64E+02 (1.97E+01)
F5	8.43E+00 (1.03E+01)	**0** (0)	1.00E+00 (3.16E+00)	1.28E+00 (4.05E+00)	5.31E+01 (5.15E+00)
F6	1.17E-01 (1.15E-01)	**1.01E-02** (2.14E-02)	3.42E+01 (1.83E+01)	4.86E+03 (1.28E+04)	1.19E+04 (6.69E+03)
F7	1.25E-01 (3.63E-01)	7.58E-01 (8.02E-01)	1.04E+01 (7.81E+00)	1.86E+01 (1.22E+01)	3.88E+01 (8.29E+00)
F8	1.99E+01 (3.27E+01)	6.40E+00 (1.36E+00)	2.93E+02 (1.66E+02)	5.02E+02 (4.44E+02)	2.12E+03 (8.39E+02)
F9	7.47E+00 (2.13E+00)	7.27E+00 (1.03E+00)	1.55E+02 (7.08E+01)	4.01E+02 (2.86E+02)	1.92E+03 (8.74E+02)
F10	**3.94E+01** (2.92E+01)	1.61E+03 (6.47E+02)	9.05E+03 (5.58E+03)	7.36E+04 (6.56E+04)	5.09E+04 (1.86E+04)
F11	**1.03E+00** (2.08E+00)	3.69E+01 (1.59E+01)	2.91E+01 (1.05E+01)	9.14E+01 (3.15E+01)	5.28E+01 (1.28E+01)
F12	1.22E+04 (3.85E+04)	**5.93E+00** (5.86E+00)	1.02E+06 (1.42E+06)	3.23E+06 (3.21E+06)	1.02E+07 (3.17E+06)
F13	2.99E+00 (8.38E+00)	4.28E+00 (5.94E+00)	2.05E+02 (2.80E+01)	3.04E+02 (1.28E+02)	5.35E+02 (5.47E+01)
F14	1.07E-02 (3.35E-02)	**3.93E-04** (2.22E-04)	7.67E-01 (1.42E-01)	2.63E+00 (2.18E+00)	4.79E+00 (9.36E-01)
F15	1.67E+01 (1.21E+01)	2.67E+01 (6.37E+00)	5.95E+01 (5.65E+00)	8.64E+01 (1.22E+01)	1.16E+02 (1.51E+01)
F16	6.41E+00 (3.79E+00)	8.73E+00 (2.61E+00)	9.85E+00 (2.69E+00)	8.13E+00 (3.93E+00)	9.66E+00 (1.34E+00)
F17	7.24E-02 (1.53E-01)	**3.29E-03** (7.39E-03)	1.92E+00 (2.63E-01)	3.34E+00 (1.96E+00)	4.58E+00 (4.05E-01)
F18	6.42E-02 (5.08E-02)	2.59E-02 (2.79E-02)	6.44E+00 (1.42E+00)	1.24E+01 (5.58E+00)	1.77E+01 (3.17E+00)
F19	2.07E+00 (3.91E-01)	2.14E+00 (5.90E-01)	3.13E+00 (5.66E-01)	5.10E+00 (9.84E-01)	6.77E+00 (1.33E+00)
F20	1.00E+00 (5.46E-01)	**1.64E-01** (1.57E-01)	2.49E+00 (1.72E-01)	1.42E+02 (1.83E+02)	3.78E+02 (4.26E+02)
F21	**1.70E+00** (2.52E+00)	1.96E+00 (8.49E-01)	8.23E+00 (6.69E+00)	1.64E+01 (1.42E+01)	1.50E+01 (3.77E+00)
F22	**2.08E-01** (6.25E-01)	4.18E+00 (4.02E+00)	7.44E+00 (8.79E+00)	1.11E+01 (1.37E+01)	1.72E+01 (5.86E+00)
F23	1.40E+00 (1.97E-01)	1.32E+00 (3.33E-01)	1.39E+00 (2.49E-01)	1.43E+00 (3.58E-01)	1.53E+00 (1.91E-01)
F24	3.38E+01 (7.03E+00)	3.86E+01 (4.10E+00)	5.85E+01 (7.74E+00)	1.00E+02 (1.29E+01)	1.05E+02 (9.74E+00)

BBOB at $d = 20$. At 20D, GLEET has average rank 2.458 and RIME 2.667.

Table 10: BBOB-20D mean final error over 10 evaluation seeds, panel A. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	GLEET	DE-DDQN	Surr-RLDE	DEDQN
F1	2.84E-01 (4.11E-01)	1.11E-04 (1.48E-04)	4.06E-01 (4.81E-01)	3.64E+00 (1.24E+00)	8.65E+01 (1.74E+01)
F2	1.29E+03 (1.54E+03)	2.74E+03 (1.51E+03)	**3.65E+02** (2.54E+02)	1.42E+04 (2.72E+04)	1.39E+06 (8.62E+05)
F3	**2.73E+01** (1.91E+01)	7.31E+01 (1.90E+01)	1.02E+02 (1.21E+01)	6.32E+01 (2.32E+01)	5.39E+02 (9.46E+01)
F4	**5.89E+01** (3.23E+01)	1.10E+02 (4.28E+01)	1.23E+02 (1.44E+01)	8.98E+01 (2.68E+01)	9.42E+02 (1.85E+02)
F5	3.69E+00 (1.04E+01)	1.44E+01 (1.55E+01)	9.45E-09 (3.98E-10)	7.96E+01 (2.94E+01)	1.20E+02 (2.87E+01)
F6	1.65E+01 (1.92E+01)	6.46E+00 (6.03E+00)	1.04E+01 (1.33E+01)	1.48E+02 (3.80E+01)	1.20E+04 (1.04E+04)
F7	5.08E+00 (4.56E+00)	8.79E+00 (2.83E+00)	3.29E+00 (3.16E+00)	2.95E+01 (2.56E+01)	4.75E+02 (1.00E+02)
F8	8.26E+01 (5.15E+01)	1.61E+01 (3.45E+00)	4.84E+01 (3.91E+01)	1.73E+03 (1.01E+03)	1.06E+05 (1.05E+04)
F9	2.72E+01 (2.59E+01)	2.88E+01 (2.65E+01)	4.50E+01 (4.25E+01)	1.04E+03 (4.18E+02)	7.02E+04 (1.33E+04)
F10	1.62E+03 (7.25E+02)	3.85E+03 (4.04E+03)	**4.86E+02** (4.86E+02)	2.52E+04 (3.32E+04)	1.23E+06 (5.03E+05)
F11	3.42E+01 (1.33E+01)	**2.80E+01** (7.86E+00)	2.96E+01 (2.05E+01)	6.03E+01 (2.14E+01)	5.42E+02 (6.63E+02)
F12	3.60E+05 (7.83E+05)	3.24E+04 (9.22E+04)	6.53E+06 (1.83E+07)	4.91E+06 (3.11E+06)	1.20E+08 (3.71E+07)
F13	1.47E+02 (1.88E+02)	8.76E+00 (7.09E+00)	9.54E+01 (6.08E+01)	2.74E+02 (6.22E+01)	1.70E+03 (1.58E+02)
F14	2.53E-01 (5.79E-01)	2.41E-03 (1.21E-03)	4.12E-01 (5.98E-01)	1.98E+00 (7.62E-01)	2.78E+01 (5.14E+00)
F15	**4.35E+01** (3.87E+01)	6.05E+01 (2.07E+01)	1.10E+02 (1.65E+01)	6.59E+01 (1.57E+01)	4.71E+02 (5.98E+01)
F16	1.76E+01 (5.29E+00)	**4.67E+00** (2.12E+00)	1.94E+01 (1.13E+00)	1.42E+01 (1.82E+00)	3.24E+01 (4.53E+00)
F17	6.58E-01 (7.24E-01)	1.30E+00 (6.95E-01)	7.00E-01 (5.37E-01)	1.44E+00 (5.16E-01)	1.15E+01 (1.99E+00)
F18	9.04E-01 (9.74E-01)	3.95E+00 (2.41E+00)	1.43E+00 (1.18E+00)	6.00E+00 (2.96E+00)	4.47E+01 (4.24E+00)
F19	3.69E+00 (5.54E-01)	**3.45E+00** (7.92E-01)	4.50E+00 (6.07E-01)	4.10E+00 (1.02E+00)	1.92E+01 (3.17E+00)
F20	1.73E+00 (4.05E-01)	1.80E+00 (3.53E-01)	2.11E+00 (7.30E-01)	2.29E+01 (4.99E+01)	3.03E+04 (7.04E+03)
F21	8.57E+00 (7.34E+00)	**5.92E+00** (7.18E+00)	1.05E+01 (1.06E+01)	7.89E+00 (5.29E+00)	7.52E+01 (1.88E+00)
F22	**9.09E+00** (3.44E+00)	1.02E+01 (4.36E+00)	1.17E+01 (5.02E+00)	1.38E+01 (8.42E+00)	7.48E+01 (2.90E+00)
F23	2.42E+00 (1.50E-01)	**1.87E+00** (3.44E-01)	2.44E+00 (2.61E-01)	2.38E+00 (2.13E-01)	3.99E+00 (4.94E-01)
F24	1.05E+02 (1.92E+01)	8.61E+01 (2.32E+01)	1.21E+02 (1.15E+01)	**8.59E+01** (3.05E+01)	4.43E+02 (5.07E+01)

Table 11: BBOB-20D mean final error over 10 evaluation seeds, panel B. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	DE	PSO	SAHLPSO	Random
F1	2.84E-01 (4.11E-01)	**9.22E-09** (7.05E-10)	9.53E+00 (1.15E+00)	2.82E+01 (1.76E+01)	6.12E+01 (7.25E+00)
F2	1.29E+03 (1.54E+03)	3.47E+04 (1.37E+04)	6.49E+04 (1.29E+04)	4.06E+05 (2.36E+05)	4.14E+05 (9.98E+04)
F3	**2.73E+01** (1.91E+01)	9.83E+01 (8.54E+00)	1.78E+02 (1.47E+01)	3.08E+02 (4.99E+01)	3.82E+02 (3.97E+01)
F4	**5.89E+01** (3.23E+01)	1.14E+02 (1.23E+01)	2.58E+02 (2.60E+01)	5.55E+02 (2.60E+02)	6.05E+02 (7.44E+01)
F5	3.69E+00 (1.04E+01)	**5.33E-09** (4.10E-09)	1.98E+01 (1.54E+01)	9.46E+00 (1.56E+01)	1.71E+02 (8.83E+00)
F6	1.65E+01 (1.92E+01)	**1.23E-03** (3.79E-03)	1.58E+02 (1.12E+02)	6.78E+02 (8.54E+02)	9.56E+02 (5.51E+02)
F7	5.08E+00 (4.56E+00)	**2.19E+00** (9.97E-01)	5.96E+01 (1.26E+01)	1.58E+02 (4.88E+01)	2.28E+02 (3.03E+01)
F8	8.26E+01 (5.15E+01)	**1.57E+01** (2.02E+00)	2.04E+03 (9.40E+02)	5.91E+03 (4.86E+03)	2.97E+04 (1.02E+04)
F9	2.72E+01 (2.59E+01)	**1.69E+01** (1.19E+00)	1.29E+03 (2.86E+02)	5.44E+03 (2.39E+03)	2.47E+04 (6.50E+03)
F10	1.62E+03 (7.25E+02)	4.36E+04 (1.84E+04)	9.76E+04 (3.20E+04)	3.82E+05 (3.53E+05)	3.56E+05 (5.86E+04)
F11	3.42E+01 (1.33E+01)	1.14E+02 (2.69E+01)	8.23E+01 (3.30E+01)	1.63E+02 (5.94E+01)	1.39E+02 (1.10E+01)
F12	3.60E+05 (7.83E+05)	**2.55E+00** (3.14E+00)	9.04E+06 (2.31E+06)	2.28E+07 (1.43E+07)	7.62E+07 (1.26E+07)
F13	1.47E+02 (1.88E+02)	**1.37E+00** (1.84E+00)	5.73E+02 (5.65E+01)	9.85E+02 (2.74E+02)	1.25E+03 (9.42E+01)
F14	2.53E-01 (5.79E-01)	**1.36E-03** (1.41E-04)	3.70E+00 (8.47E-01)	9.58E+00 (4.40E+00)	1.57E+01 (2.62E+00)
F15	**4.35E+01** (3.87E+01)	1.01E+02 (9.58E+00)	1.91E+02 (2.03E+01)	3.16E+02 (4.16E+01)	3.81E+02 (4.67E+01)
F16	1.76E+01 (5.29E+00)	1.92E+01 (1.73E+00)	1.79E+01 (3.97E+00)	1.99E+01 (6.28E+00)	1.90E+01 (1.89E+00)
F17	6.58E-01 (7.24E-01)	**1.96E-02** (2.94E-02)	3.36E+00 (5.70E-01)	7.54E+00 (2.41E+00)	8.16E+00 (7.52E-01)
F18	9.04E-01 (9.74E-01)	**4.73E-02** (2.30E-02)	1.38E+01 (1.65E+00)	2.78E+01 (7.58E+00)	2.96E+01 (3.02E+00)
F19	3.69E+00 (5.54E-01)	4.40E+00 (3.33E-01)	6.12E+00 (5.98E-01)	7.68E+00 (1.44E+00)	1.16E+01 (1.08E+00)
F20	1.73E+00 (4.05E-01)	**4.06E-01** (1.36E-01)	1.40E+01 (1.44E+01)	1.07E+03 (1.45E+03)	1.04E+04 (1.75E+03)
F21	8.57E+00 (7.34E+00)	1.43E+01 (1.01E+01)	2.36E+01 (1.21E+01)	3.70E+01 (2.44E+01)	5.99E+01 (6.09E+00)
F22	**9.09E+00** (3.44E+00)	1.01E+01 (4.45E+00)	2.13E+01 (8.43E+00)	2.90E+01 (1.48E+01)	6.43E+01 (4.53E+00)
F23	2.42E+00 (1.50E-01)	2.25E+00 (2.29E-01)	2.21E+00 (3.74E-01)	2.03E+00 (6.09E-01)	2.40E+00 (2.94E-01)
F24	1.05E+02 (1.92E+01)	1.15E+02 (1.36E+01)	1.88E+02 (1.31E+01)	2.62E+02 (3.79E+01)	3.37E+02 (3.48E+01)

BBOB at $d = 30$. At 30D, DE leads with average rank 2.500, followed by RIME at 2.625 and GLEET at 2.667.

Table 12: BBOB-30D mean final error over 10 evaluation seeds, panel A. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	GLEET	DE-DDQN	Surr-RLDE	DEDQN
F1	4.00E+00 (6.33E+00)	1.51E-01 (4.77E-01)	1.84E+01 (8.41E+00)	1.05E+01 (3.05E+00)	1.38E+02 (1.70E+01)
F2	**9.42E+03** (1.78E+04)	1.27E+04 (1.44E+04)	1.00E+04 (7.49E+03)	8.25E+04 (8.63E+04)	2.57E+06 (7.90E+05)
F3	9.01E+01 (5.92E+01)	1.08E+02 (2.86E+01)	2.48E+02 (2.77E+01)	**8.92E+01** (1.56E+01)	8.82E+02 (1.26E+02)
F4	**1.54E+02** (9.10E+01)	2.11E+02 (5.66E+01)	3.17E+02 (9.16E+01)	1.69E+02 (1.13E+02)	1.66E+03 (3.90E+02)
F5	3.48E+01 (5.34E+01)	4.58E+01 (2.34E+01)	8.72E-09 (3.08E-09)	1.70E+02 (2.83E+01)	2.23E+02 (3.09E+01)
F6	1.34E+02 (7.73E+01)	7.24E+01 (4.13E+01)	1.64E+02 (4.55E+01)	5.72E+02 (5.26E+02)	1.80E+05 (7.59E+04)
F7	3.71E+01 (1.60E+01)	2.06E+01 (8.38E+00)	4.60E+01 (1.78E+01)	5.39E+01 (2.30E+01)	7.74E+02 (1.30E+02)
F8	5.02E+02 (6.05E+02)	7.41E+01 (4.46E+01)	1.25E+03 (9.48E+02)	3.65E+03 (2.09E+03)	1.85E+05 (5.69E+04)
F9	5.44E+01 (4.30E+01)	3.90E+01 (1.87E+01)	3.14E+02 (1.15E+02)	1.80E+03 (6.97E+02)	1.56E+05 (3.56E+04)
F10	2.96E+04 (6.45E+04)	7.88E+03 (4.96E+03)	**5.72E+03** (3.67E+03)	1.02E+05 (1.02E+05)	2.94E+06 (7.10E+05)
F11	4.93E+01 (9.77E+00)	**3.43E+01** (1.08E+01)	7.13E+01 (3.56E+01)	1.03E+02 (3.22E+01)	4.38E+02 (8.57E+01)
F12	3.09E+06 (3.31E+06)	2.64E+05 (6.37E+05)	2.81E+07 (2.71E+07)	1.05E+07 (4.42E+06)	2.59E+08 (6.57E+07)
F13	2.80E+02 (1.87E+02)	3.13E+01 (5.58E+01)	7.29E+02 (1.88E+02)	3.44E+02 (2.37E+01)	2.27E+03 (1.75E+02)
F14	2.98E-01 (8.34E-01)	4.96E-02 (1.38E-01)	4.98E+00 (2.79E+00)	2.80E+00 (5.03E-01)	4.77E+01 (5.30E+00)
F15	**7.57E+01** (5.82E+01)	1.36E+02 (3.56E+01)	2.52E+02 (3.01E+01)	1.10E+02 (2.81E+01)	9.82E+02 (1.05E+02)
F16	2.29E+01 (6.74E+00)	**8.34E+00** (2.92E+00)	2.56E+01 (1.66E+00)	1.93E+01 (1.16E+00)	4.00E+01 (2.60E+00)
F17	4.31E-01 (3.46E-01)	2.74E+00 (7.64E-01)	2.64E+00 (9.62E-01)	1.87E+00 (6.12E-01)	1.36E+01 (1.56E+00)
F18	1.63E+00 (1.67E+00)	8.95E+00 (2.03E+00)	5.12E+00 (2.14E+00)	7.66E+00 (2.90E+00)	5.11E+01 (5.84E+00)
F19	**4.75E+00** (4.58E-01)	4.86E+00 (5.42E-01)	5.21E+00 (4.90E-01)	5.24E+00 (3.61E-01)	2.16E+01 (3.43E+00)
F20	2.54E+00 (1.32E+00)	1.89E+00 (1.71E-01)	1.80E+00 (1.13E-01)	7.71E+01 (1.32E+02)	6.23E+04 (1.10E+04)
F21	**1.90E+00** (1.45E+00)	2.90E+00 (3.50E+00)	1.19E+01 (6.93E+00)	1.18E+01 (4.58E+00)	7.89E+01 (2.04E+00)
F22	2.17E+00 (3.55E-01)	8.17E+00 (8.09E+00)	1.92E+01 (1.36E+01)	1.17E+01 (2.83E+00)	7.93E+01 (2.07E+00)
F23	3.02E+00 (3.28E-01)	2.60E+00 (2.73E-01)	2.69E+00 (3.62E-01)	2.90E+00 (2.74E-01)	5.47E+00 (4.82E-01)
F24	1.65E+02 (5.09E+01)	1.70E+02 (3.23E+01)	2.27E+02 (2.67E+01)	**1.40E+02** (6.58E+01)	7.59E+02 (6.23E+01)

Table 13: BBOB-30D mean final error over 10 evaluation seeds, panel B. Lower is better. Bold marks the lowest mean across all methods in both panels.

Task	RIME	DE	PSO	SAHLPSO	Random
F1	4.00E+00 (6.33E+00)	**8.82E-09** (1.30E-09)	2.33E+01 (7.61E+00)	6.20E+01 (2.00E+01)	1.16E+02 (9.01E+00)
F2	**9.42E+03** (1.78E+04)	1.23E+05 (3.84E+04)	2.33E+05 (4.24E+04)	9.65E+05 (5.60E+05)	1.08E+06 (2.79E+05)
F3	9.01E+01 (5.92E+01)	1.90E+02 (1.41E+01)	3.47E+02 (3.41E+01)	5.16E+02 (8.12E+01)	7.04E+02 (4.40E+01)
F4	**1.54E+02** (9.10E+01)	2.01E+02 (1.45E+01)	5.13E+02 (7.36E+01)	1.05E+03 (3.80E+02)	1.36E+03 (1.42E+02)
F5	3.48E+01 (5.34E+01)	**8.03E-09** (8.91E-10)	4.23E+01 (3.36E+01)	2.19E+01 (1.87E+01)	3.01E+02 (9.52E+00)
F6	1.34E+02 (7.73E+01)	**5.43E-04** (7.33E-04)	4.56E+02 (1.50E+02)	5.58E+04 (9.59E+04)	1.74E+05 (3.77E+04)
F7	3.71E+01 (1.60E+01)	**1.11E+01** (5.48E+00)	1.22E+02 (2.84E+01)	3.58E+02 (1.24E+02)	4.93E+02 (6.28E+01)
F8	5.02E+02 (6.05E+02)	**2.68E+01** (1.12E+00)	6.01E+03 (1.21E+03)	5.15E+04 (2.57E+04)	1.04E+05 (2.18E+04)
F9	5.44E+01 (4.30E+01)	**3.24E+01** (1.84E+01)	5.03E+03 (9.74E+02)	8.92E+03 (2.63E+03)	7.11E+04 (8.09E+03)
F10	2.96E+04 (6.45E+04)	1.06E+05 (4.26E+04)	2.29E+05 (5.26E+04)	9.15E+05 (5.79E+05)	1.15E+06 (3.15E+05)
F11	4.93E+01 (9.77E+00)	1.56E+02 (3.12E+01)	1.24E+02 (2.85E+01)	2.49E+02 (3.89E+01)	2.34E+02 (2.90E+01)
F12	3.09E+06 (3.31E+06)	**2.45E+00** (3.45E+00)	3.01E+07 (9.20E+06)	9.47E+07 (5.38E+07)	1.88E+08 (1.66E+07)
F13	2.80E+02 (1.87E+02)	**1.72E+00** (1.68E+00)	9.22E+02 (1.20E+02)	1.38E+03 (3.64E+02)	1.97E+03 (1.21E+02)
F14	2.98E-01 (8.34E-01)	**1.58E-03** (2.35E-04)	7.28E+00 (1.65E+00)	2.04E+01 (9.11E+00)	3.00E+01 (2.70E+00)
F15	**7.57E+01** (5.82E+01)	1.92E+02 (8.95E+00)	3.34E+02 (1.77E+01)	5.77E+02 (8.29E+01)	7.35E+02 (7.85E+01)
F16	2.29E+01 (6.74E+00)	2.55E+01 (1.90E+00)	2.56E+01 (2.00E+00)	2.61E+01 (6.16E+00)	2.58E+01 (2.63E+00)
F17	4.31E-01 (3.46E-01)	**7.42E-03** (7.85E-03)	4.50E+00 (8.85E-01)	1.04E+01 (1.36E+00)	1.05E+01 (6.97E-01)
F18	1.63E+00 (1.67E+00)	**2.32E-01** (1.49E-01)	1.58E+01 (2.02E+00)	2.86E+01 (7.29E+00)	3.74E+01 (3.55E+00)
F19	**4.75E+00** (4.58E-01)	5.50E+00 (3.61E-01)	7.41E+00 (4.38E-01)	9.24E+00 (9.29E-01)	1.47E+01 (1.24E+00)
F20	2.54E+00 (1.32E+00)	**7.13E-01** (2.44E-01)	1.60E+03 (2.00E+03)	1.01E+04 (6.74E+03)	2.61E+04 (5.99E+03)
F21	**1.90E+00** (1.45E+00)	5.23E+00 (4.99E+00)	1.75E+01 (6.33E+00)	4.32E+01 (2.35E+01)	7.01E+01 (2.02E+00)
F22	2.17E+00 (3.55E-01)	**1.95E+00** (0)	2.03E+01 (9.00E+00)	5.22E+01 (1.44E+01)	7.31E+01 (2.38E+00)
F23	3.02E+00 (3.28E-01)	2.84E+00 (3.06E-01)	2.89E+00 (3.78E-01)	**2.60E+00** (1.15E+00)	3.00E+00 (2.77E-01)
F24	1.65E+02 (5.09E+01)	2.09E+02 (1.34E+01)	3.33E+02 (1.50E+01)	4.57E+02 (3.70E+01)	6.16E+02 (3.61E+01)

E.2 P-TEST COMPARISON

Table 14 reports the separate eight-function P-Test comparison. RIME achieves lower mean errors than Surr-RLDE on the two Schaffers tasks and Gallagher-101. Surr-RLDE performs better on composite Griewank–Rosenbrock and Gallagher-21. GLEET and DE-DDQN also lead on individual functions.

Table 14: P-Test mean final error over 10 evaluation seeds, panel A. Function-family exposure and the transformed-instance protocol are specified in Section 5.1. Bold marks the lowest mean across both panels.

Task	RIME	GLEET	DE-DDQN	Surr-RLDE	DEDQN
P1	5.22E+00 (2.77E+00)	**5.12E-01** (3.82E-01)	4.76E+00 (2.90E+00)	5.29E+00 (1.27E+00)	2.09E+01 (3.16E+00)
P2	3.11E-02 (3.05E-02)	1.25E-01 (1.14E-01)	**2.20E-04** (1.67E-04)	7.80E-01 (2.56E-01)	7.42E+00 (1.44E+00)
P3	9.19E-02 (4.10E-02)	4.43E-01 (2.74E-01)	**7.29E-04** (2.97E-04)	2.77E+00 (1.27E+00)	2.69E+01 (4.66E+00)
P4	1.52E+00 (3.76E-01)	1.77E+00 (5.22E-01)	2.41E+00 (5.29E-01)	1.27E+00 (8.34E-01)	1.23E+01 (3.38E+00)
P5	6.81E-07 (2.15E-06)	**7.02E-09** (2.13E-09)	6.96E-01 (9.64E-01)	3.83E-02 (7.27E-02)	4.67E+01 (7.71E+00)
P6	1.06E+00 (1.10E+00)	**1.82E-07** (5.10E-07)	3.12E+00 (3.03E+00)	6.22E-01 (9.66E-01)	5.55E+01 (9.80E+00)
P7	1.31E+00 (5.96E-01)	1.48E+00 (2.18E-01)	1.31E+00 (2.10E-01)	1.61E+00 (2.29E-01)	3.23E+00 (7.46E-01)
P8	3.35E+01 (1.27E+01)	**3.06E+01** (7.98E+00)	4.26E+01 (7.27E+00)	3.14E+01 (1.66E+01)	1.64E+02 (2.13E+01)

Table 15: P-Test mean final error over 10 evaluation seeds, panel B. Function-family exposure and the transformed-instance protocol are specified in Section 5.1. Bold marks the lowest mean across both panels.

Task	RIME	DE	PSO	SAHLPSO	Random
P1	5.22E+00 (2.77E+00)	8.01E+00 (1.27E+00)	6.39E+00 (1.51E+00)	6.38E+00 (4.55E+00)	9.79E+00 (1.82E+00)
P2	3.11E-02 (3.05E-02)	2.44E-04 (5.30E-04)	1.60E+00 (2.94E-01)	2.19E+00 (7.09E-01)	4.38E+00 (6.67E-01)
P3	9.19E-02 (4.10E-02)	1.55E-02 (1.39E-02)	5.67E+00 (9.47E-01)	6.74E+00 (4.02E+00)	1.48E+01 (2.25E+00)
P4	1.52E+00 (3.76E-01)	**8.60E-01** (1.54E-01)	2.85E+00 (6.20E-01)	5.43E+00 (9.57E-01)	7.09E+00 (4.08E-01)
P5	6.81E-07 (2.15E-06)	2.87E-01 (6.50E-01)	1.47E+00 (1.95E+00)	2.26E+00 (3.78E+00)	1.28E+01 (4.96E+00)
P6	1.06E+00 (1.10E+00)	3.90E+00 (4.32E+00)	4.59E+00 (6.21E+00)	6.42E+00 (7.90E+00)	1.68E+01 (3.65E+00)
P7	1.31E+00 (5.96E-01)	1.70E+00 (2.31E-01)	1.60E+00 (2.00E-01)	**1.27E+00** (6.53E-01)	1.40E+00 (2.96E-01)
P8	3.35E+01 (1.27E+01)	3.96E+01 (3.72E+00)	6.40E+01 (5.49E+00)	9.15E+01 (1.03E+01)	1.07E+02 (9.65E+00)

E.3 UAV TRANSFER

Table 16 reports transfer to 14 UAV path-planning tasks. RIME improves on Surr-RLDE on 12 tasks and has average rank 2.714, while GLEET leads at 1.929.

Table 16: Mean final objective on UAV path-planning tasks over 10 evaluation seeds, panel A. Bold marks the lowest mean across all methods in both panels.

Task	RIME	GLEET	DE-DDQN	Surr-RLDE	DEDQN
U1	**1.07E+04** (3.77E+03)	1.18E+04 (4.91E+03)	1.18E+04 (4.02E+03)	1.23E+04 (4.22E+03)	2.58E+04 (1.14E+04)
U2	8.02E+03 (2.78E+03)	**6.24E+03** (3.30E+02)	1.05E+04 (3.69E+03)	7.91E+03 (3.08E+02)	2.05E+04 (3.09E+03)
U3	6.80E+03 (3.55E+02)	**5.99E+03** (1.26E+02)	6.78E+03 (4.79E+02)	7.14E+03 (2.58E+02)	9.31E+03 (4.43E+02)
U4	7.09E+03 (3.92E+02)	6.66E+03 (5.14E+02)	**6.63E+03** (7.46E+02)	7.67E+03 (7.99E+02)	1.26E+04 (3.22E+03)
U5	1.24E+04 (3.43E+03)	**9.64E+03** (3.83E+03)	1.35E+04 (3.88E+03)	1.25E+04 (4.49E+03)	2.63E+04 (1.14E+04)
U6	1.61E+04 (4.71E+02)	1.73E+04 (3.42E+03)	1.60E+04 (2.38E+02)	1.73E+04 (7.15E+02)	3.34E+04 (6.86E+03)
U7	1.56E+04 (3.88E+03)	1.43E+04 (3.50E+03)	1.56E+04 (4.71E+03)	1.71E+04 (2.19E+03)	3.68E+04 (9.57E+03)
U8	7.20E+03 (7.96E+02)	**6.17E+03** (3.36E+02)	1.05E+04 (3.73E+03)	8.03E+03 (3.32E+02)	2.06E+04 (3.10E+03)
U9	8.39E+03 (4.15E+02)	**7.99E+03** (4.44E+02)	8.52E+03 (3.19E+02)	9.22E+03 (9.15E+02)	2.32E+04 (7.71E+03)
U10	6.32E+03 (3.26E+02)	**5.53E+03** (1.62E+02)	6.58E+03 (2.89E+02)	6.72E+03 (1.98E+02)	8.85E+03 (5.71E+02)
U11	1.11E+04 (3.55E+03)	1.16E+04 (4.58E+03)	1.35E+04 (3.77E+03)	1.19E+04 (3.76E+03)	2.62E+04 (1.13E+04)
U12	**1.10E+04** (4.22E+03)	1.11E+04 (3.98E+03)	1.43E+04 (3.50E+03)	1.19E+04 (3.91E+03)	2.01E+04 (3.33E+03)
U13	8.06E+03 (3.34E+02)	**6.71E+03** (4.15E+02)	7.97E+03 (6.72E+02)	7.90E+03 (3.81E+02)	1.63E+04 (4.80E+03)
U14	7.43E+03 (3.77E+02)	**6.11E+03** (3.44E+02)	7.56E+03 (5.70E+02)	7.67E+03 (3.04E+02)	1.37E+04 (5.34E+03)

Table 17: Mean final objective on UAV path-planning tasks over 10 evaluation seeds, panel B. Bold marks the lowest mean across all methods in both panels.

Task	RIME	DE	PSO	SAHLPSO	Random
U1	**1.07E+04** (3.77E+03)	1.31E+04 (3.14E+03)	1.14E+04 (4.47E+03)	1.51E+04 (3.76E+03)	1.29E+04 (3.83E+03)
U2	8.02E+03 (2.78E+03)	9.00E+03 (4.97E+02)	8.49E+03 (2.46E+03)	1.47E+04 (4.97E+03)	8.85E+03 (5.84E+02)
U3	6.80E+03 (3.55E+02)	6.98E+03 (5.24E+02)	6.81E+03 (6.37E+02)	1.12E+04 (1.47E+03)	8.09E+03 (4.50E+02)
U4	7.09E+03 (3.92E+02)	6.78E+03 (8.05E+02)	7.18E+03 (4.07E+02)	1.08E+04 (1.39E+03)	9.53E+03 (6.96E+02)
U5	1.24E+04 (3.43E+03)	1.56E+04 (1.88E+03)	1.25E+04 (4.34E+03)	1.40E+04 (3.51E+03)	1.34E+04 (3.81E+03)
U6	1.61E+04 (4.71E+02)	**1.53E+04** (1.59E+02)	1.67E+04 (4.09E+03)	1.89E+04 (3.99E+03)	2.04E+04 (3.02E+03)
U7	1.56E+04 (3.88E+03)	**1.22E+04** (5.10E+03)	1.40E+04 (4.93E+03)	1.23E+04 (1.63E+03)	1.65E+04 (4.18E+03)
U8	7.20E+03 (7.96E+02)	1.00E+04 (2.07E+03)	9.33E+03 (3.63E+03)	1.49E+04 (4.68E+03)	8.97E+03 (5.20E+02)
U9	8.39E+03 (4.15E+02)	8.97E+03 (3.21E+02)	8.50E+03 (7.16E+02)	1.48E+04 (2.96E+03)	1.10E+04 (6.44E+02)
U10	6.32E+03 (3.26E+02)	6.92E+03 (4.19E+02)	6.70E+03 (2.75E+02)	1.15E+04 (1.94E+03)	7.97E+03 (2.46E+02)
U11	1.11E+04 (3.55E+03)	1.38E+04 (3.08E+03)	**1.09E+04** (4.28E+03)	1.53E+04 (3.42E+03)	1.34E+04 (3.84E+03)
U12	**1.10E+04** (4.22E+03)	1.35E+04 (3.49E+03)	1.15E+04 (4.37E+03)	1.35E+04 (3.58E+03)	1.18E+04 (4.07E+03)
U13	8.06E+03 (3.34E+02)	7.76E+03 (3.67E+02)	7.70E+03 (4.85E+02)	1.52E+04 (4.21E+03)	9.00E+03 (5.51E+02)
U14	7.43E+03 (3.77E+02)	7.93E+03 (6.48E+02)	7.64E+03 (5.09E+02)	1.22E+04 (1.64E+03)	8.66E+03 (4.16E+02)

E.4 Component and reliability ablations

Table 18 shows how component effects vary across the 24 BBOB functions. Full has the best average rank, and removing task context gives the worst. Because context enters both the controller and world model, this ablation measures its joint contribution. The per-function results also show variation: No-WM has a lower final error on Sphere.

Table 18: Full 24-family BBOB-10D component ablation: mean final error over 10 test seeds under matched Q-update counts.

Task	Full	No-WM	Imag.-only	Real-only	No-context
F1	2.14E-02 (5.02E-02)	**3.13E-14** (8.99E-15)	3.22E-01 (1.02E+00)	4.32E-02 (1.20E-01)	1.05E+00 (4.77E-01)
F2	**2.13E+01** (2.53E+01)	1.20E+02 (1.34E+02)	5.25E+01 (3.70E+01)	4.53E+02 (5.70E+02)	9.29E+02 (9.36E+02)
F3	**1.33E+01** (8.57E+00)	3.59E+01 (9.33E+00)	2.96E+01 (1.13E+01)	4.65E+01 (1.72E+01)	3.78E+01 (1.23E+01)
F4	**2.00E+01** (1.07E+01)	5.89E+01 (2.87E+01)	2.86E+01 (3.07E+01)	7.62E+01 (4.26E+01)	6.22E+01 (1.70E+01)
F5	8.43E+00 (1.03E+01)	**2.21E-01** (6.99E-01)	2.41E+01 (1.36E+01)	9.58E+00 (1.44E+01)	1.51E+01 (7.15E+00)
F6	1.17E-01 (1.15E-01)	**4.95E-03** (1.28E-02)	1.17E+03 (2.11E+03)	1.97E+00 (3.83E+00)	1.13E+02 (1.82E+02)
F7	**1.25E-01** (3.63E-01)	3.20E+00 (2.61E+00)	3.85E-01 (5.88E-01)	7.76E+00 (6.70E+00)	9.61E+00 (6.40E+00)
F8	1.99E+01 (3.27E+01)	3.19E+00 (2.95E+00)	**1.93E+00** (1.58E+00)	8.17E+01 (2.37E+02)	3.06E+02 (3.72E+02)
F9	7.47E+00 (2.13E+00)	**2.69E+00** (2.69E+00)	3.18E+00 (1.39E+00)	5.50E+01 (9.84E+01)	1.21E+02 (7.50E+01)
F10	**3.94E+01** (2.92E+01)	2.26E+02 (1.14E+02)	1.75E+02 (2.07E+02)	3.10E+02 (2.59E+02)	5.39E+03 (8.54E+03)
F11	**1.03E+00** (2.08E+00)	2.92E+00 (2.18E+00)	2.24E+00 (2.87E+00)	1.67E+01 (1.09E+01)	1.41E+00 (2.53E+00)
F12	1.22E+04 (3.85E+04)	**2.33E+00** (3.48E+00)	4.54E+01 (1.30E+02)	1.26E+01 (1.92E+01)	1.31E+06 (9.33E+05)
F13	**2.99E+00** (8.38E+00)	5.20E+00 (9.01E+00)	4.61E+01 (1.42E+02)	3.61E+01 (6.01E+01)	1.86E+02 (6.57E+01)
F14	1.07E-02 (3.35E-02)	**4.36E-05** (2.44E-05)	7.96E-05 (5.86E-05)	4.34E-01 (1.32E+00)	8.47E-01 (8.49E-01)
F15	**1.67E+01** (1.21E+01)	3.63E+01 (1.12E+01)	2.56E+01 (1.95E+01)	5.56E+01 (1.43E+01)	4.09E+01 (1.31E+01)
F16	6.41E+00 (3.79E+00)	**1.30E+00** (1.40E+00)	2.55E+00 (2.77E+00)	4.67E+00 (2.79E+00)	1.79E+00 (2.29E+00)
F17	**7.24E-02** (1.53E-01)	1.74E+00 (1.29E+00)	2.29E-01 (3.48E-01)	2.29E+00 (1.49E+00)	1.86E+00 (6.30E-01)
F18	**6.42E-02** (5.08E-02)	6.86E+00 (5.84E+00)	3.08E+00 (4.86E+00)	1.24E+01 (6.29E+00)	8.05E+00 (6.08E+00)
F19	**2.07E+00** (3.91E-01)	2.09E+00 (5.57E-01)	2.59E+00 (4.81E-01)	2.40E+00 (1.55E+00)	2.53E+00 (6.84E-01)
F20	1.00E+00 (5.46E-01)	1.21E+00 (3.87E-01)	**6.95E-01** (1.75E-01)	1.62E+00 (6.58E-01)	1.60E+00 (5.69E-01)
F21	**1.70E+00** (2.52E+00)	2.70E+00 (2.78E+00)	2.81E+00 (2.34E+00)	2.03E+00 (9.84E-01)	6.54E+00 (6.43E+00)
F22	**2.08E-01** (6.25E-01)	3.81E+00 (5.68E+00)	2.85E-01 (5.04E-01)	2.73E+00 (4.46E+00)	1.07E+01 (1.83E+01)
F23	1.40E+00 (1.97E-01)	1.49E+00 (1.40E-01)	1.07E+00 (4.94E-01)	**5.96E-01** (2.85E-01)	1.49E+00 (3.03E-01)
F24	3.38E+01 (7.03E+00)	**3.35E+01** (1.20E+01)	4.07E+01 (9.55E+00)	4.22E+01 (1.63E+01)	3.85E+01 (8.41E+00)

Table 19 compares gate and horizon variants. Full wins 21 of 24 pairwise comparisons against No-global and 22 against No-local. It also has a lower average rank than Fixed-H5 (2.479) and Ungated-WM (2.500). The larger degradation from removing an individual gate indicates that gate configuration matters alongside rollout length.

Table 19: Per-function reliability-control ablation on the 24-function BBOB-10D suite. Cells report mean final error over 10 paired evaluation seeds; all variants use the same instances, training seed, and 1.5M Q-update budget.

Task	Full	Fixed-H5	Ungated-WM	No-local	No-global
F1	9.20E-03 (2.55E-02)	5.24E-03 (1.10E-02)	**4.26E-14** (4.49E-14)	6.37E-01 (1.21E+00)	9.26E-01 (1.73E+00)
F2	**2.80E+01** (3.12E+01)	3.08E+02 (4.48E+02)	1.33E+02 (1.19E+02)	1.16E+03 (1.71E+03)	8.13E+03 (1.29E+04)
F3	**1.16E+01** (8.26E+00)	3.12E+01 (1.56E+01)	2.87E+01 (1.35E+01)	1.84E+01 (9.82E+00)	5.93E+01 (3.00E+01)
F4	**1.56E+01** (8.42E+00)	4.02E+01 (1.57E+01)	5.27E+01 (1.96E+01)	4.69E+01 (2.96E+01)	9.06E+01 (3.54E+01)
F5	9.54E+00 (1.24E+01)	**0** (0)	9.20E-05 (2.85E-04)	9.92E+00 (1.28E+01)	**0** (0)
F6	**3.16E-01** (3.42E-01)	1.91E+00 (4.23E+00)	2.37E+00 (7.50E+00)	6.66E+00 (1.46E+01)	3.48E+03 (1.09E+04)
F7	**1.37E-01** (3.65E-01)	2.40E+00 (2.12E+00)	2.09E+00 (1.56E+00)	7.94E+00 (1.08E+01)	2.11E+01 (1.95E+01)
F8	1.81E+01 (2.93E+01)	**9.09E-01** (1.64E+00)	4.49E+00 (5.47E+00)	9.50E+01 (2.52E+02)	4.15E+01 (5.38E+01)
F9	**7.07E+00** (2.01E+00)	2.59E+01 (4.89E+01)	1.71E+01 (3.57E+01)	8.62E+01 (9.55E+01)	2.28E+02 (2.53E+02)
F10	**1.91E+01** (1.27E+01)	3.06E+02 (3.42E+02)	1.65E+02 (1.91E+02)	1.17E+03 (2.59E+03)	7.16E+03 (9.79E+03)
F11	**6.28E-01** (6.48E-01)	9.62E-01 (5.93E-01)	2.37E+00 (2.07E+00)	5.26E+00 (3.69E+00)	2.81E+00 (4.50E+00)
F12	**2.71E+00** (2.91E+00)	1.24E+04 (3.19E+04)	3.53E+01 (8.83E+01)	1.91E+03 (5.77E+03)	4.19E+05 (6.53E+05)
F13	**5.53E+00** (1.56E+01)	1.25E+02 (1.12E+02)	1.10E+01 (1.44E+01)	2.34E+01 (3.93E+01)	1.74E+02 (1.51E+02)
F14	2.64E-04 (4.21E-04)	4.61E-02 (1.26E-01)	**6.20E-05** (3.21E-05)	3.26E-01 (7.99E-01)	2.01E+00 (2.71E+00)
F15	**1.55E+01** (1.22E+01)	2.80E+01 (1.58E+01)	3.09E+01 (1.28E+01)	2.22E+01 (2.21E+01)	6.19E+01 (1.95E+01)
F16	9.07E+00 (2.96E+00)	**1.48E+00** (1.57E+00)	1.80E+00 (1.16E+00)	2.26E+00 (2.09E+00)	1.82E+00 (1.20E+00)
F17	**7.02E-02** (1.20E-01)	8.31E-01 (6.11E-01)	1.05E+00 (7.73E-01)	1.53E+00 (1.69E+00)	2.73E+00 (2.14E+00)
F18	**6.41E-02** (9.85E-02)	3.84E+00 (1.78E+00)	3.04E+00 (2.13E+00)	2.80E+00 (3.88E+00)	8.48E+00 (5.87E+00)
F19	2.14E+00 (2.41E-01)	**1.47E+00** (8.80E-01)	1.73E+00 (1.17E+00)	2.18E+00 (1.16E+00)	2.78E+00 (8.54E-01)
F20	**9.79E-01** (4.73E-01)	1.10E+00 (3.58E-01)	1.17E+00 (2.06E-01)	4.58E+00 (1.04E+01)	2.49E+01 (7.34E+01)
F21	**9.54E-01** (1.05E+00)	5.77E+00 (6.48E+00)	4.28E+00 (4.33E+00)	5.76E+00 (6.51E+00)	6.99E+00 (6.34E+00)
F22	**2.35E-01** (3.12E-01)	5.20E+00 (8.16E+00)	4.06E+00 (7.08E+00)	9.99E+00 (1.07E+01)	1.29E+01 (1.07E+01)
F23	1.47E+00 (2.01E-01)	**4.74E-01** (2.14E-01)	5.75E-01 (4.64E-01)	1.17E+00 (5.36E-01)	9.56E-01 (6.18E-01)
F24	3.55E+01 (5.39E+00)	**3.34E+01** (1.35E+01)	4.39E+01 (2.02E+01)	4.17E+01 (1.30E+01)	4.33E+01 (1.14E+01)

E.5 HORIZON AND IMAGINED-FRACTION SENSITIVITY

Table 21 compares six configurations on the 16 controller-training BBOB families at $d = 10$. H5-F50 has the lowest mean rank (2.250) and eight first-place finishes, followed by H3-F50 (2.375) and H3-F25 (2.500). Increasing the scheduled imagined fraction from 0.25 to 0.50 improves average ranks at $H = 3$ and $H = 5$, but worsens the rank at $H = 1$. These results show why horizon and imagined fraction are selected jointly. Table 22 reports the per-function values.

Table 20 provides an additional comparison of H5-F50 with H3-F25 across BBOB, P-Test, and UAV tasks. H5-F50 wins 61 versus 56 of the 120 BBOB conditions, with three ties, and eight versus six UAV tasks. Hyperparameter selection uses the training-family results in Table 21.

Table 20: Additional comparison of H5-F50 and H3-F25 on BBOB (120 function–dimension conditions), P-Test (8 functions), and UAV (14 tasks). The last three columns count lower-mean wins and ties. Configuration selection uses the training-family results in Table 21.

Dataset	H5-F50 rank	H3-F25 rank	H5 wins	H3 wins	Ties
BBOB	2.417	2.983	61	56	3
P-Test	2.875	3.500	5	3	0
UAV	2.714	3.286	8	6	0

Table 21: Sensitivity to the maximum imagination horizon H and scheduled imagined-update fraction β on the 16 BBOB function families used for controller training (F1–F15 and F20, $d = 10$). Each rank is computed from mean final errors over 10 evaluation seeds. Lower is better; the first-place count is over the 16 functions.

Configuration	Mean rank	First-place count
$H = 5,\ \beta = 0.50$	**2.250**	**8**
$H = 3,\ \beta = 0.50$	2.375	4
$H = 3,\ \beta = 0.25$	2.500	4
$H = 5,\ \beta = 0.25$	3.875	0
$H = 1,\ \beta = 0.25$	4.188	0
$H = 1,\ \beta = 0.50$	5.813	0

Table 22: Per-function sensitivity on the 16 BBOB families used for controller training (F1–F15 and F20, $d = 10$), panel A. Each cell reports mean final error with population standard deviation in parentheses over 10 evaluation seeds. Bold marks the lowest mean among all six configurations; these means determine the ranks in Table 21.

Task	H5-F50	H3-F50	H3-F25
F1	2.14E-02 (4.76E-02)	1.78E-01 (5.32E-01)	**3.41E-14** (1.14E-14)
F2	**2.13E+01** (2.40E+01)	2.62E+02 (2.48E+02)	9.62E+01 (8.08E+01)
F3	**1.33E+01** (8.13E+00)	2.06E+01 (9.70E+00)	4.59E+01 (8.37E+00)
F4	**2.00E+01** (1.01E+01)	4.55E+01 (1.98E+01)	4.25E+01 (2.44E+01)
F5	8.43E+00 (9.74E+00)	7.36E+00 (1.24E+01)	**0.00E+00** (0.00E+00)
F6	1.17E-01 (1.09E-01)	5.45E+00 (1.54E+01)	**7.49E-04** (1.51E-03)
F7	**1.25E-01** (3.44E-01)	3.46E+00 (3.31E+00)	4.09E+00 (2.28E+00)
F8	1.99E+01 (3.11E+01)	1.80E+01 (3.92E+01)	**1.69E+00** (1.96E+00)
F9	7.47E+00 (2.02E+00)	**1.93E+00** (2.17E+00)	3.46E+00 (2.78E+00)
F10	**3.94E+01** (2.77E+01)	1.90E+02 (3.11E+02)	2.46E+02 (1.34E+02)
F11	**1.03E+00** (1.97E+00)	2.78E+00 (3.00E+00)	5.26E+00 (5.64E+00)
F12	1.22E+04 (3.65E+04)	**1.95E+00** (2.33E+00)	7.59E+00 (6.98E+00)
F13	2.99E+00 (7.95E+00)	**1.62E+00** (1.77E+00)	3.34E+01 (5.64E+01)
F14	1.07E-02 (3.18E-02)	**4.19E-05** (2.49E-05)	2.71E-03 (5.78E-03)
F15	**1.67E+01** (1.14E+01)	2.50E+01 (1.00E+01)	4.75E+01 (1.65E+01)
F20	**1.00E+00** (5.18E-01)	1.08E+00 (3.57E-01)	1.20E+00 (2.78E-01)

Table 23: Per-function sensitivity on the 16 BBOB families used for controller training (F1–F15 and F20, $d = 10$), panel B. Each cell reports mean final error with population standard deviation in parentheses over 10 evaluation seeds. Bold marks the lowest mean among all six configurations; these means determine the ranks in Table 21.

Task	H5-F25	H1-F25	H1-F50
F1	2.70E-01 (5.81E-01)	2.17E-01 (6.52E-01)	4.71E+00 (2.49E+00)
F2	3.00E+02 (2.45E+02)	1.45E+02 (1.43E+02)	1.02E+04 (1.55E+04)
F3	4.69E+01 (2.16E+01)	4.16E+01 (1.16E+01)	6.59E+01 (2.28E+01)
F4	6.42E+01 (2.54E+01)	8.00E+01 (2.26E+01)	9.48E+01 (5.31E+01)
F5	7.22E-03 (1.81E-02)	6.42E+00 (1.06E+01)	4.77E+00 (6.38E+00)
F6	1.52E+00 (2.33E+00)	1.46E+03 (2.88E+03)	4.87E+03 (5.77E+03)
F7	1.00E+01 (4.95E+00)	1.09E+01 (5.38E+00)	2.09E+01 (1.13E+01)
F8	1.05E+01 (1.03E+01)	3.13E+01 (6.87E+01)	8.60E+02 (1.39E+03)
F9	4.12E+01 (7.17E+01)	5.22E+00 (2.87E+00)	4.32E+02 (2.32E+02)
F10	4.57E+02 (3.86E+02)	1.01E+03 (2.09E+03)	1.78E+04 (1.30E+04)
F11	5.45E+00 (6.40E+00)	5.52E+00 (3.28E+00)	2.87E+01 (1.25E+01)
F12	8.79E+00 (7.73E+00)	3.03E+06 (5.34E+06)	4.96E+06 (4.76E+06)
F13	2.04E+01 (2.30E+01)	6.93E+00 (1.17E+01)	3.28E+02 (1.16E+02)
F14	5.79E-01 (9.81E-01)	1.71E-04 (1.39E-04)	3.85E+00 (1.44E+00)
F15	4.03E+01 (2.10E+01)	5.04E+01 (1.75E+01)	6.00E+01 (1.76E+01)
F20	1.24E+00 (3.57E-01)	1.49E+00 (2.68E-01)	6.73E+01 (1.16E+02)

E.6 World-model and gate diagnostics

Table 24 and Figure 2 describe how imagination was used. Disagreement correlates positively with state MSE (Pearson 0.498) and reward Brier error (0.408). Of 750,000 scheduled imagined updates, 99,213 are rejected, giving an effective imagined fraction of 0.433858. State MSE at depth five is slightly below its depth-one value, while both state and reward errors are higher at depth fifteen. These observations motivate limiting rollout length even when short predictions remain useful.

Table 24: World-model and gate diagnostics from the selected H5-F50 training run. Correlations use periodic replay-buffer samples. Rollout errors average the recorded diagnostic episodes at each depth.

Diagnostic	Value
One-step diagnostic records	1,500
Pearson corr.: uncertainty / state MSE	0.498
Rank corr.: uncertainty / state MSE	0.514
Pearson corr.: uncertainty / reward Brier	0.408
Rank corr.: uncertainty / reward Brier	0.402
First accepted imagination Q step	4,520
Scheduled / accepted / rejected updates	750,000 / 650,787 / 99,213
Cumulative scheduled acceptance	86.772%
Effective imagined-update fraction	0.433858
Final adaptive-horizon EMA	1.425
Rollout H=1: state MSE / reward Brier / uncertainty	0.02169 / 0.15124 / 0.00939
Rollout H=5: state MSE / reward Brier / uncertainty	0.02134 / 0.17461 / 0.01093
Rollout H=15: state MSE / reward Brier / uncertainty	0.02571 / 0.19288 / 0.01174