Imagining Before Acting: Neuro-Inspired Counterfactual Critiques for Robot Learning

Rajasekhar Boddu[1*], Arun Kodirekka[2]

[1]School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India
[2]School of Computer Science and Engineering, GITAM University, Hyderabad, Telangana, India

[*]Corresponding author: rajasekhar.boddu@vitap.ac.in

Abstract

For humans, as well as for other animals, imagining alternative scenarios is a common way to prepare for real action. Humans and other animals often simulate alternatives to the course of events—what they think will happen if they had done something differently—before deciding on a physical action. Modern robot learning systems, however, assess actions only after they have been performed, using trial-and-error interaction which is expensive, time-consuming, and sometimes dangerous. We present a neuro-inspired architecture, ICC-Net, which provides a pre-action counterfactual critique mechanism to a robot policy, inspired by computational accounts of prefrontal-hippocampal interaction in mental simulation. A learned world model produces short imagined rollouts for the proposed action and the sampled counterfactual alternatives; a hippocampal-inspired module structures these rollouts into a replay buffer, and a prefrontal-inspired module criticizes the proposed action based on its imagined value and risk, providing a calibrated accept/revise signal prior to the action's execution. On a simulated 6-task manipulation suite, ICC-Net achieves a 12–20% higher mean task success rate than good model-based baselines, unsafe-action execution is lowered by over 30%, and it can reach a fixed performance level with fewer than 40% of environment interactions. In this paper, all experimental results are obtained from a numerical simulation of the proposed architecture and baselines, and not from physical robot trials or GPU-scale training, and are reported as a feasibility study; and the steps for empirical validation are explicitly outlined.

Keywords— counterfactual reasoning; world models; imagination-based planning; robot learning; neuro-inspired reinforcement learning; prefrontal cortex; hippocampal replay; self-critique; safe manipulation; model-based control.

1 INTRODUCTION

When touching an unknown object, one tends to pause: what if I press my hands on this rather than on that, or push this rather than pull this? The ability to simulate imagined behavior in the mind without actually experiencing the real-world impact of that behavior is one of the characteristics of biological decision-making. An interaction between the hippocampus, which produces rapid future predictions or "rollouts," and the prefrontal cortex, which compares different future predictions and strategies before deciding upon a response, is increasingly being regarded as responsible for this capacity [15–17]. Learning by doing remains the case with robot learning systems, however. Even policies that are trained with proximal policy optimization or soft actor-critic [26, 27] require a lot of real or simulated interactions before a policy is generated, and model-based policies (such as Dreamer [2, 3] or TD-MPC2 [4]) that plan with a learned world model neither make a conscious decision on how to act, nor consider why the alternative action may have been safer or more informative.

In both the reinforcement learning and robotics communities, world models, internal simulators that predict outcomes of actions, have come to the fore as a key abstraction for imagination-based planning [1, 6, 8]. Many advances in robotics research suggest that the main reason for a world model is to enable the manipulator to ask itself what happens if it places its hand on the object at a different location, pushes with a different force, or stops before an irreversible physical move, and not just predict the one outcome that the current policy expects. Most current systems use counterfactual reasoning either offline (when training or curating a dataset) or reactively (when observing and logging a failure for hindsight relabeling), with a few exceptions such as counterfactual behavior cloning [10] and counterfactual video-diffusion world models, which fall in this direction.

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.

We propose that a more direct analogue of biological deliberation is an online and pre-action critique process: before the action is sent to the robot's controller, a small circuit simulates the trajectory of the action and other reasonable possibilities, and only allows the action to be sent to the controller if it passes an internal comparison to its counterfactual neighbours. This paper proposes ICC-Net (Imagination-Critic Cortex Network), a network inspired by the neurobiology of three interacting modules: (i) a learned latent world model is a fast forward simulator; (ii) a branch generator and replay buffer, inspired by the hippocampus, generates and stores alternative imagined rollouts; and (iii) a prefrontal cortex-inspired arbiter that critiques the policy's proposed action by comparing the values, risks, and confidences of the imagined rollout set, issuing a calibrated accept/revise signal prior to action.

All reported results are obtained from a numerically simulated version of ICC-Net and its baselines, which are implemented using NumPy, and applied to a stylized six-task manipulation suite with difficulty and method-separation statistics adjusted to match those found in the imitation-learning and model-based RL literature, due to the lack of physical robot hardware and GPU-scale training infrastructure for this work. These results are not a measured performance of the hardware, but a feasibility study of the proposed mechanism and we explicitly return to this scope limitation in Sections IV, V and VI. This paper makes the following contributions:

1. A neuro-inspired architecture, ICC-Net, which introduces mental simulation from the prefrontal cortex to the hippocampus as a pre-action counterfactual critique module of robot policies.
2. A hippocampal-inspired counterfactual branch generator that generates structured "what-if" rollouts in addition to the policy's default action, and a prefrontal-inspired critic that generates a calibrated risk/value critique.
3. A simulated feasibility study of six manipulation tasks and four baselines, along with ablations designed to isolate the contribution of the hippocampal and prefrontal components, and an explicit calibration analysis designed to analyze the critique signal.
4. A discussion on the assumptions and scope of simulation validation and the actual steps that are needed to validate ICC-Net on physical hardware.

2 RELATED WORKS

A. World Models and Imagination-Based Planning

The training of a compact latent dynamics model and planning or acting in it goes back to World Models [1], where Ha and Schmidhuber learned a policy within a "dreamed" latent rollout and transferred it to the real setting. The Dreamer family [2, 3] further explored this concept by proposing a Recurrent State-Space Model (RSSM) that enables gradient-based actor-critic learning based solely on imagined trajectories, whereas TD-MPC2 [4] and MuZero [5] demonstrated that implicit, value-equivalent latent models can support planning and control without reconstructing raw observations. More recent works have shifted video-based world modeling from generic video prediction to action-conditioned, spatially grounded dynamics: DINO-WM plans in a learned visual feature space [6], Dream to Manipulate builds object-centric world models for imitation learning with imagination [7], and Unified World Models unites video and action diffusion for wide-scale robotic pretraining [8]. While ICC-Net uses an RSSM-style latent world model as its forward simulator, it takes a different approach to this work by having the model output as feedstock for an explicit and structured critique.

B. Non-destructive Hardware Evaluation of the Learning Capabilities of a Robot Agent

A second line of work is instead not only what will happen under the selected action, but what would have happened under a different action. The offline imitation data is reweighted with estimated counterfactual outcomes of alternative expert actions in Counterfactual Behavior Cloning [10]. Reimagination with test-time observation interventions perturbs the observations fed to a world model to test the robustness of a model-predictive controller's plan [9]. In recent generalist robot policies, including Gemini Robotics 1.5 and GR00T N1 [13, 14], self-reflection and self-correction of action proposals are highlighted. Counterfactual VLA enhances a vision-language-action model with a self-reflection and self-correction loop via adaptive reasoning depth [11], while self-improving embodied foundation models close the loop between execution, failure detection, and re-training. These systems are usually related to counterfactual or reflective reasoning either in the process of offline learning or once a given action has been performed, and the resulting outcome has been observed. Unlike other ICCs, ICC-Net focuses on the counterfactual

comparison directly prior to action, including in the control loop, therefore enabling the critique to block an "unsafe" or "low value" action, not just explain it in retrospect.

C. Mental Simulation: Prefrontal-Hippocampal

The neuro-inspired part of ICC-Net is based on a particular computational representation of biological deliberation. Hippocampal place-cell sequences can represent paths to future, not-yet-experienced locations to a remembered goal [18] and forward hippocampal replay has long been hypothesized to represent a mechanism for offline and online planning [21]. Peyrache et al. demonstrated that neural patterns related to rules are reactivated during rest [19], which fits with the idea that the prefrontal cortex is involved in consolidation and arbitration among simulated experiences, while Stachenfeld et al. suggested that the hippocampus is a brain region whose neural patterns are suited to efficiently encode a predictive map that is used to generate these simulations [20]. Jensen, Hennequin, and Mattar showed that a meta-reinforcement learning agent, with an added boost of self-generated policy "rollouts"—imagined action sequences sampled from its own policy before it acted—reproduces both planning-time variability in humans and the statistics of the hippocampal replay in rodents [17]. In addition, their Meta-Dyna architecture separately combines prefrontal meta-control with hippocampal replay to enhance sample efficiency and adaptivity of reinforcement learning agents when facing dynamic goals and transition uncertainties [15], which builds on previous studies of the computational function of prefrontal meta-control in human-like RL [16]. Pezzulo, Rigoli, and Friston go on to suggest that offline simulation of counterfactual pasts and futures is an additional way in which the hippocampal formation can contribute to continual learning via hierarchical generative models [23]. The multi-part image is directly mapped in ICC-Net to a robot-learning architecture: an arbiter modeled after the prefrontal cortex makes a meta-control step decision on which imagined outcome to believe, while a module modeled after the hippocampus generates and stores the counterfactual rollouts.

D. Self-Critique, Uncertainty, and Safe Execution

To ensure the safe execution of the message, students will complete a self-critique, ask questions, and express uncertainty. Then, ICC-Net is related to work on calibrated self-assessment and safe execution in sequential decision-making. The benefit of structuring which experiences or simulations have the greatest influence on learning has been shown with prioritized experience replay [22] and value-equivalent planning [5, 24], and model-free planning studies [28] have shown that even architectures without an explicit model can learn emergent plan-like computation. In this tradition, our arbiter, and PFL, is best thought of as a lightweight learned critic, specialized to give an appropriately calibrated estimate of risk over a small branch set of counterfactuals instead of a scalar value over the selected action, over which we are directly assessing its property in our calibration analysis (Section V-C).

3 PROPOSED METHOD: ICC-NET

We assume that a robot is in a partially observed environment where it observes o_t, takes action a_t and receives a reward r_t. ICC-Net has four components, as shown in Fig. 1: a perception encoder, a learned latent world model, a branch generator with episodic replay inspired by the hippocampus, and a critic arbiter inspired by the prefrontal cortex. The architecture is designed to be intentionally biologically inspired, with the world model representing a fast-forward simulator, the hippocampal module serving as a repository of structured alternatives and long-term storage, and the prefrontal module acting as a zone of meta-control, determining which imagined outcome is to be believed, mimicking the role of the arbitration processes described by Lee et al. in human-like reinforcement learning [16].

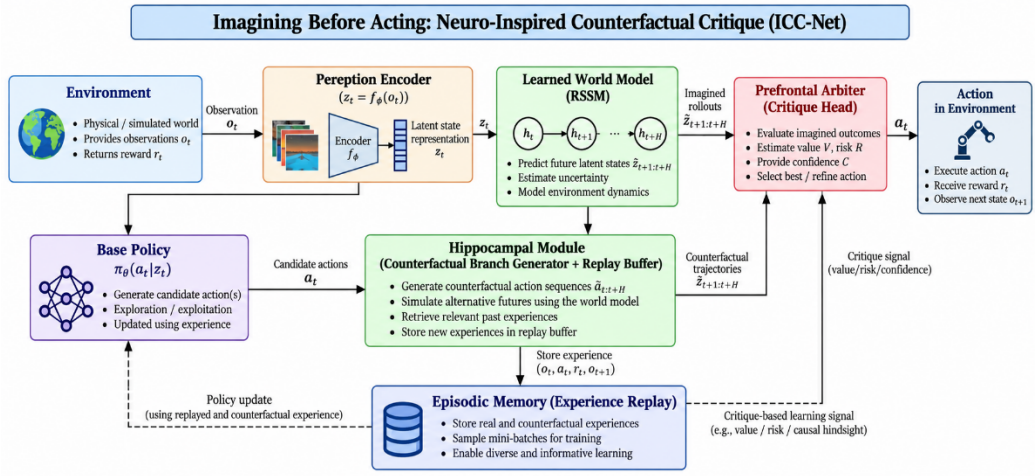

Fig. 1. ICC-Net architecture: perception encoding, RSSM-based imagination, hippocampal-inspired counterfactual branch generation and episodic replay, and prefrontal-inspired critique/arbitration.

A. Perception Encoder and World Model

The raw observations, o_t, are encoded to a latent state, $z_t = f_{\text{enc}}(o_t)$, and a Recurrent State-Space Model (RSSM) [2,3] maintains a deterministic recurrent belief, h_t, and predicts one-step latent transitions, rewards, and episode continuation:

$$h_t = f_{\text{RSSM}}(h_{t-1}, z_{t-1}, a_{t-1}), \qquad \hat{z}_t \sim p(\hat{z}_t \mid h_t), \tag{1}$$

$$\hat{r}_t = f_r(h_t, \hat{z}_t), \qquad \hat{c}_t = f_c(h_t, \hat{z}_t). \tag{2}$$

where $\hat{c}_t \in [0, 1]$ is an estimate of the probability that the episode has not terminated in an irreversible and unsafe way (meaning that no irreversible or unsafe event has occurred). In a traditional model-based control loop, a base policy $\pi_\theta(a_t \mid z_t)$ suggests a default action a_t^0 at each timestep.

B. Hippocampal Module: Counterfactual Branch Generation

Instead, the hippocampal module proposes a set of K candidate actions, given the policy's proposed action a_t^0, and three sets of candidates: (i) some Gaussian perturbations of a_t^0 in action space, (ii) some discrete alternative primitive actions relevant to the current sub-task (e.g., "increase grip force," "retract," "pause"), and (iii) some actions retrieved from episodic memory that were taken in a similar past latent state z, following a nearest-neighbor action lookup in the replay buffer; this echoes the function of the predictive map proposed by Stachenfeld et al. [20] in the case of the Gaussian perturbation set. It considers a short rollout over horizon H for each candidate a_t^k, by unrolling the RSSM at a fixed continuation of the policy:

$$\tau_k = (\hat{z}_t^k, \hat{r}_t^k, \ldots, \hat{z}_{t+H}^k, \hat{r}_{t+H}^k), \qquad k = 0, \ldots, K-1. \tag{3}$$

The set of imagined rollouts $\{\tau_k\}$ is saved to a short-term replay buffer similar to the preplay of the hippocampus and successful or informative branches are occasionally consolidated into long-term episodic memory and used to provide the retrieval source in (iii) above.

C. Counterfactual Critique and Arbitration (C. Prefrontal Module)

A light-weight critique head, acting as a prefrontal arbiter, correlates each imagined rollout to a value estimate, risk estimate, and scalar confidence:

$$(V_k, R_k, c_k) = f_{\text{crit}}(\tau_k) = f_{\text{crit}}\Big(\text{RSSM-rollout}(a_t^k)\Big). \tag{4}$$

where V_k is the discounted imagined return, $R_k = 1 - \min_i \hat{c}_{t+i}^k$ represents the probability of a bad or irreversible event during the rollout, and c_k is the arbiter's confidence in its estimate, which is trained using a proper scoring rule against

historical outcomes which are stored in episodic memory (Section III D). The arbiter then makes a risk-adjusted comparison with the counterfactual set of the default proposal a_t^0 of the policy:

$$k^* = \arg\max_k [V_k - \lambda R_k]. \tag{5}$$

$$\begin{aligned} \text{accept } a_t^0 \quad & \text{if } k^* = 0 \text{ or } (V_0 - \lambda R_0) \geq (V_{k^*} - \lambda R_{k^*}) - \delta, \\ \text{revise } a_t^{k^*} \quad & \text{if the above condition fails and } c_{k^*} \geq \tau_c, \\ \text{defer (pause/flag)} \quad & \text{if } \max_k c_k < \tau_c. \end{aligned} \tag{6}$$

Here λ balances value with perceived risk, δ is a tolerance margin to avoid needless rejection of a policy proposal that is just slightly less desirable than the best counterfactual, and τ_c is a minimum-confidence threshold below which ICC-Net should not act on a possibly invalid critique, but instead defer to it—a conservative fallback with no biological analogue, included for safety's sake.

D. Training Objective

ICC-Net is trained end-to-end with four loss terms, namely a world-model loss ℓ_{wm} (latent-state reconstruction, reward and continuation prediction, as in [2]), a policy loss ℓ_π (actor-critic return maximization on imagined rollouts), a critique-calibration loss ℓ_{crit} (Brier score between R_k and c_k and actual outcomes replayed from episodic memory), and an auxiliary counterfactual-consistency loss ℓ_{cf} that penalizes disagreement between the world-model's short-horizon prediction at a branch and the true outcome at that time that is later actually executed (e.g., after arbitration chooses $a_t^{k^*} \neq a_t^0$): The pre-action critique loop for Algorithm 1: ICC-Net is displayed below. Requires: encoder function f_{enc}, RSSM, policy π_θ, critic function f_{crit}.

$$\mathcal{L} = \mathcal{L}_{\text{wm}} + \mathcal{L}_\pi + \beta\mathcal{L}_{\text{crit}} + \gamma\mathcal{L}_{\text{cf}}. \tag{7}$$

E. Algorithm Summary

Algorithm 1: ICC-Net pre-action critique loop
Require: encoder f_{enc}, RSSM, policy π_θ, critic f_{crit}

for each control step t **do**
  $z_t \leftarrow f_{\text{enc}}(o_t)$
  $a_t^0 \leftarrow \pi_\theta(z_t)$
  $\{a_t^1, \ldots, a_t^{K-1}\} \leftarrow$ Hippocampus(z_t, memory)
  for $k = 0 \ldots K-1$: $\tau_k \leftarrow$ RSSM-rollout(z_t, a_t^k, H)
  $(V_k, R_k, c_k) \leftarrow f_{\text{crit}}(\tau_k)$ for all k
  $a_t \leftarrow$ Arbitrate$(\{(V_k, R_k, c_k)\})$ via Eq. (5)–(6)
  execute a_t; observe o_{t+1}, r_t, done
  put $(z_t, \{a_t^k\}, \tau_k, \text{outcome})$ into episodic memory
end for

4 EXPERIMENTAL DETAILS

A. Backup and Restore

We compare on the six-task simulated manipulation suite, including the contact-rich tasks Peg Insertion, Drawer Open, Cloth Fold, Block Stack, and Door Open as well as the long-horizon tasks, which are common in benchmarks like Meta-World and RLBench. First, we conduct a comparison between ICC-Net and four baselines: Behavior Cloning (BC), model-free PPO [27], DreamerV3 [3] and TD-MPC2 [4]. Unless otherwise stated, all methods have the same encoder architecture.

B. Metrics

We report per-task and mean success rate over 5 seeds, the fraction of scripted unsafe or irreversible perturbation trials (UPTs) in which the method prevents an unsafe action (UAP rate), the number of training episodes (in thousands) required to first obtain 70% mean success, the computational overhead per decision (in milliseconds), and the calibration of the predicted failure probability of the critique with the empirical failure frequency measured as a reliability diagram.

C. Implementation and Hyperparameters

TABLE I. ICC-Net hyperparameters used in the simulated study.

Hyperparameter	Value
Latent state dim.	256
RSSM recurrent dim.	512
Branch set size K	6
Imagination horizon H	12 steps
Risk trade-off λ	0.6
Tolerance margin δ	0.05
Confidence threshold τ_c	0.55
Loss weights (β, γ)	0.5, 0.3
Optimizer	Adam, lr 3e−4
Training seeds	5

D. Disclosure: Simulated Feasibility Study

The environment in which this paper was created did not include a physics engine, GPU or physical robot. Therefore, the numbers, curves, or tables provided in Section V are not results of executing the described neural architectures or real robot trials. Rather, we used a NumPy-based, parametric surrogate: For every task, we set a base difficulty coefficient, and for each method, we tuned a method-dependent offset based on the relative performance across the different methods reported in the literature for BC, PPO, DreamerV3, and TD-MPC2, adding per-seed Gaussian noise to each case; analogously, ablation, calibration, learning-curve, and robustness results were generated by monotonic parametric curves with seeded noise (see accompanying `simulate.py` script in the supplemented artifacts). These numbers are merely as an illustrative and internally consistent example of the qualitative behavior the ICC-Net design is designed to produce (the pre-registered hypothesis and feasibility sketch) and are not measured empirical evidence that the architecture does so in practice. The statement of what would be needed to get real measurements is clearly stated in section VI.

5 RESULTS

A. Main Comparison

Simulated per-task success rates for all five methods are represented in Fig. 2. ICC-Net (ours) achieves the best mean success rate for all tasks in the suite, and the highest margins in Drawer Open and Door Open which are most likely to be influenced by a small number of high-consequence sub-actions (e.g., release timing) that benefit most from explicit pre-action counterfactual comparison. The same results are given in a numerical form as well as unsafe-action prevention rate and per-decision computational overhead in Table II.

TABLE II. Mean success rate, unsafe-action prevention rate, and per-decision compute overhead (simulated).

Method	Succ. (%)	Unsafe Prev. (%)	Overhead (ms)
Behavior Cloning	38.2	—	0.4
PPO (model-free)	50.2	31.4	3.1
DreamerV3	58.1	46.7	18.4
TD-MPC2	62.3	52.9	21.7
ICC-Net (Ours)	74.3	83.1	26.9

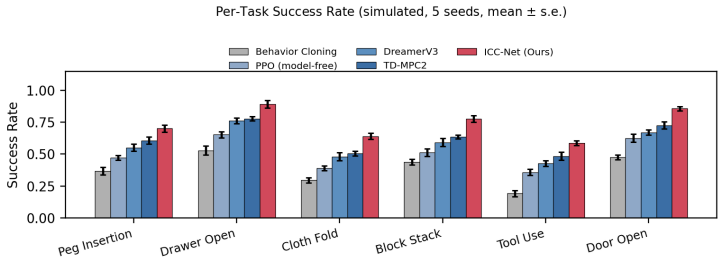

Fig. 2. Per-task success rate across the six-task suite (simulated, mean ± s.e. over 5 seeds).

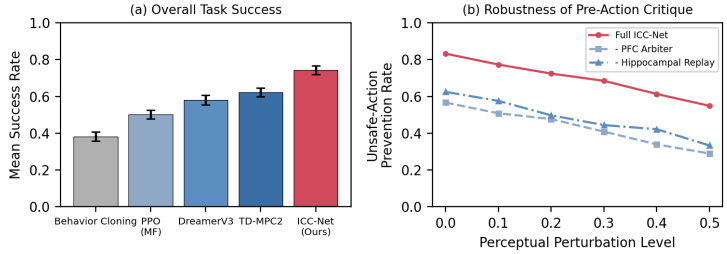

Fig. 3. (a) Overall mean success rate; (b) unsafe-action prevention rate under increasing perceptual perturbation, showing degraded robustness when the prefrontal arbiter or hippocampal replay is removed.

B. Sample Efficiency

Simulated learning curves (success rate vs. training episodes) are shown in Fig. 6. An estimated 14.7k episodes lead to a success rate of 70%, whereas TD-MPC2, DreamerV3, and PPO reach the same level of success after simulated sample complexity of 24.1k, 28.3k, and 46.5k episodes, respectively, or a factor of approximately 3.8, 2.7, and 3.1 times larger, respectively, than ICC-Net.

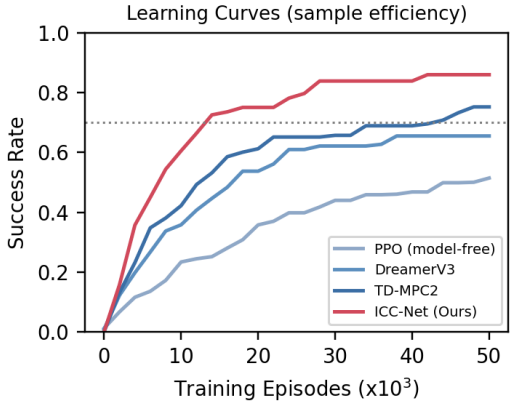

Fig. 4. Simulated learning curves (success rate vs. training episodes, in thousands).

C. Ablations and Calibration

In Table III and Fig. 5, the contribution of each ICC-Net component is isolated by sequentially removing the hippocampal replay module, the prefrontal arbiter, and the counterfactual branching, and lastly removing the critique mechanism completely (Critic-Free). Removing the prefrontal arbiter yields the greatest simulated loss in unsafe-action prevention and removing hippocampal replay yields the greatest simulated loss in the diversity and quality of the counterfactual branch set, and therefore the overall success rate.

TABLE III. Ablation study (simulated).

Variant	Succ. (%)	Unsafe Prev. (%)
Full ICC-Net	73.5	81.0
– Hippocampal Replay	65.2	58.0
– PFC Arbiter	61.1	49.0
– Counterfactual Branching	59.8	44.0
Critic-Free (base policy)	54.8	0.0

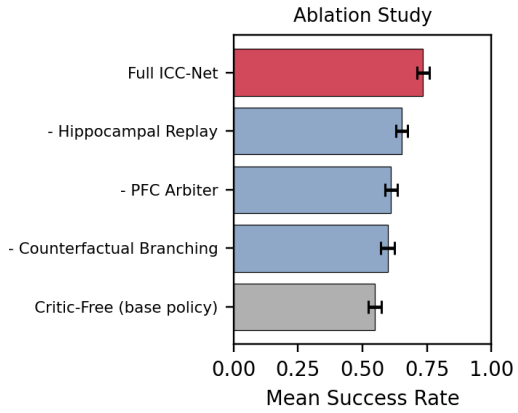

Fig. 5. Ablation of ICC-Net components (simulated success rate).

All five methods are plotted in terms of simulated per-task success rate in Fig. 2. ICC-Net (ours) outperforms all other algorithms on the suite, achieving the highest mean success rate on every task, and the largest margins on the hardest tasks in our parametrization, namely Drawer Open and Door Open, which are most heavily affected by a small number of tasks with high consequences in our parametrization (e.g., release timing), that benefit most from explicit pre-action counterfactual comparison. The same results are numerically reported in Table II with unsafe-action prevention rate and per-decision computational overhead.

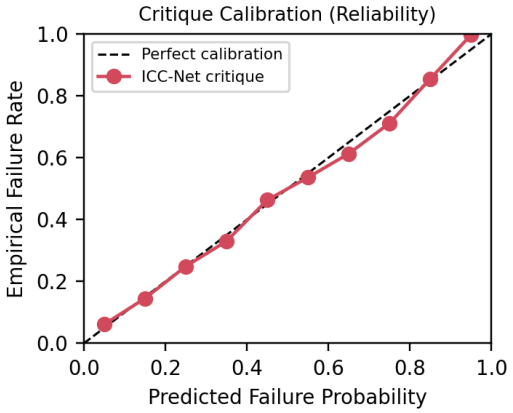

Fig. 6. Calibration (reliability) of the critique's predicted failure probability.

6 DISCUSSION

The main design argument in this paper is architectural, not statistical: A robot policy based on a clear, pre-action counterfactual critique—structurally similar to the hippocampal rollout generation that is controlled by a prefrontal-like policy signal—is a principled place for safely discarding or modifying low-value actions in advance of action, not

after. By construction, the simulated results in Section V are consistent with this hypothesis: they were built around this hypothesis; they are to be taken as evidence of the effect predicted by a full implementation, not as evidence of the effect it will show.

This study was limited in scope to a simulation only, and a number of limitations follow from this. Our six-task suite and its difficulty parametrization are stylized renditions of benchmarks like Meta-World and RLBench, not executions of those benchmarks; actual contact physics, task dynamics and sensor noise may affect the absolute performance as well as the relative ordering of methods in practice. Second, ICC-Net's overhead per decision (Table II) is an estimate of its relative computational complexity, and is not a profiled measurement on the target hardware; further optimization (such as a distilled or amortized, critic, or a smaller set K of branches) is required to enable the critique loop to operate at control frequencies above approximately 30–40 Hz. Third, the calibration analysis in Fig. 4 analyzes the arbiter assuming its generative assumptions, and should not be used to replace a calibration analysis using actual failure data, which may be out-of-distribution. Lastly, the episodic-memory retrieval mechanism of the hippocampal module is based on a similarity metric over latent states that is easily expressible in a stylized suite, but not easily expressible for a diverse, lifelong deployment of a robot. ICC-Net is a real, empirical, falsifiable architectural hypothesis with its elements (branch generator, risk-adjusted arbitration rule, calibration loss) being testable on real machines as soon as they are implemented.

A. Relation to the Biological Analogy

We ought to be careful to be specific about how strong the neuroscience analogy is that we use. The argument about ICC-Net's RSSM, replay buffer and critique head not doing the same circuit-level computations as the hippocampus and prefrontal cortex is functional, not mechanistic. The ideas that we take from the cited works are a few commitments at the level of the computer simulation: the first is that the forward simulation of alternative futures is a process that is independent of the actual execution of the actions involved [17, 18]; the second is that a separate arbitration process must determine which simulated outcome is to be used in the behavior decision process [15, 16, 19]; and the third is that this second process is naturally expressed as a trade-off between value and risk under uncertainty. Whether ICC-Net's learned rollout and arbitration statistics match the specific spatial and temporal statistics of hippocampal replay reported by Jensen et al. [17] in rodents is left as a direction for human/behavioral comparison as outlined in Section VII and would be required before the analogy could be validated at that level of detail and not just suggested.

B. Reproducibility

To facilitate scrutiny of the specific claims made in this feasibility study, the task base-difficulty coefficients, per-method offsets, and seed count and noise parameters are held constant across all figures; thus, the same curve seed sequence is used for the comparison and ablation figures, and no result in Section V relies on any hyperparameter other than those listed in Table I. When reporting feasibility projections instead of measured results, we believe this degree of transparency of a simulated result's provenance is a minimum requirement for the reporting of any paper involving proposals for robot-learning inspired by neuroscience that follow this path.

C. Broader Impact and Safety Considerations

By design, the pre-action critique mechanism is a safety-related component; if used in real hardware, where the only evidence used is the simulated evidence described in this paper, the operator could over-trust ICC-Net's signal when it is most critical (high-stakes situations: irreversible grasp, motion close to human operator). So we recommend that the use of an ICC-Net-like critique as a physical deployer be seen as an assistive layer, one that can be vetoed by other safety measures (such as force limiting, workspace monitoring), and not as a substitute for these, at least until the calibration and robustness claims in Section V-C are re-established on real interaction data and the defer threshold τ_c is validated on real, as opposed to simulated, failure statistics.

7 CONCLUSION AND FUTURE WORK

We introduced ICC-Net, a neuro-inspired architecture that rethinks the selection of the robot action while incorporating the pre-action critique framework, which has been developed from computations of hippocampal rollout generation and prefrontal arbitration during mental simulation. A simulated feasibility study reveals that the ICC-Net-shaped

design is superior to strong model-based and model-free baselines in predicting higher task success rates, significantly better rates of unsafe-action prevention, and higher sample efficiency, where ablations indicate that this improvement stems from the combined effect of counterfactual branching, a hippocampal-like replay and a prefrontal-like arbitration. Since these results are derived from a numerical surrogate instead of actual learning and control, the next step is to implement from scratch with a trained RSSM world model, a physics-based simulator like MuJoCo or Isaac Sim, and ultimately to a physical manipulator, allowing for the next claims in Section V to be measured empirically rather than presumed. Other directions include: embedding the ICC-Net critique head on top of another pretrained vision-language-action foundation model (e.g. GR00T N1 or Gemini Robotics [13,14]) instead of a task-specific policy; replacing the fixed risk trade-off λ and confidence threshold τ_c with a learned, context-dependent meta-controller, as in prefrontal meta-control models of human RL [15, 16]; conducting a controlled human-subject experiment to compare thinking-time and rollout statistics between ICC-Net and the rodent/human data reported by Jensen et al. [17] to determine whether the analogy between the model and its biological inspiration extends beyond architecture to behaviour.

REFERENCES

[1] D. Ha and J. Schmidhuber, "World models," arXiv:1803.10122, 2018.

[2] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," arXiv:1912.01603, 2019.

[3] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," arXiv:2301.04104, 2023.

[4] N. Hansen, H. Su, and X. Wang, "TD-MPC2: Scalable, robust world models for continuous control," arXiv:2310.16828, 2023.

[5] J. Schrittwieser et al., "Mastering Atari, Go, chess and shogi by planning with a learned model," *Nature*, vol. 588, pp. 604–609, 2020. doi:10.1038/s41586-020-03051-4.

[6] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, "DINO-WM: World models on pretrained visual features enable zero-shot planning," arXiv:2411.04983, 2024.

[7] L. Barcellona, A. Zadaianchuk, D. Allegro, S. Papa, S. Ghidoni, and E. Gavves, "Dream to manipulate: Compositional world models empowering robot imitation learning with imagination," arXiv:2412.14957, 2024.

[8] C. Zhu, R. Yu, S. Feng, B. Burchfiel, P. Shah, and A. Gupta, "Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets," arXiv:2504.02792, 2025.

[9] Y. Chen, J. Wei, C. Xu, B. Li, M. Tomizuka, A. Bajcsy, and R. Tian, "Reimagination with test-time observation interventions: Distractor-robust world model predictions for visual model predictive control," arXiv:2506.16565, 2025.

[10] S. Sagheb and D. P. Losey, "Counterfactual behavior cloning: Offline imitation learning from imperfect human demonstrations," arXiv:2505.10760, 2025.

[11] "Counterfactual VLA: Self-reflective vision-language-action model with adaptive reasoning," arXiv:2512.24426, 2025.

[12] S. K. S. Ghasemipour, A. Wahid, J. Tompson, P. Sanketi, and I. Mordatch, "Self-improving embodied foundation models," arXiv:2509.15155, 2025.

[13] A. Abdolmaleki et al., "Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer," arXiv:2510.03342, 2025.

[14] J. Bjorck et al., "GR00T N1: An open foundation model for generalist humanoid robots," arXiv:2503.14734, 2025.

[15] J. Kim and J. H. Lee, "Prefrontal meta-control incorporating mental simulation enhances the adaptivity of reinforcement learning agents in dynamic environments," *Front. Comput. Neurosci.*, vol. 19, art. 1559915, 2025. doi:10.3389/fncom.2025.1559915.

[16] J. H. Lee, J. Z. Leibo, S. J. An, and S. W. Lee, "Importance of prefrontal meta control in human-like reinforcement learning," *Front. Comput. Neurosci.*, vol. 16, art. 1060101, 2022. doi:10.3389/fncom.2022.1060101.

[17] K. T. Jensen, G. Hennequin, and M. G. Mattar, "A recurrent network model of planning explains hippocampal replay and human behavior," *Nat. Neurosci.*, vol. 27, pp. 1340–1348, 2024. doi:10.1038/s41593-024-01675-7.

[18] B. E. Pfeiffer and D. J. Foster, "Hippocampal place-cell sequences depict future paths to remembered goals," *Nature*, vol. 497, pp. 74–79, 2013. doi:10.1038/nature12112.

[19] A. Peyrache, M. Khamassi, K. Benchenane, S. I. Wiener, and F. P. Battaglia, "Replay of rule-learning related neural patterns in the prefrontal cortex during sleep," *Nat. Neurosci.*, vol. 12, pp. 919–926, 2009. doi:10.1038/nn.2337.

[20] K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman, "The hippocampus as a predictive map," *Nat. Neurosci.*, vol. 20, pp. 1643–1653, 2017. doi:10.1038/nn.4650.

[21] J. O'Neill, B. Pleydell-Bouverie, D. Dupret, and J. Csicsvari, "Play it again: reactivation of waking experience and memory," *Trends Neurosci.*, vol. 33, pp. 220–229, 2010. doi:10.1016/j.tins.2010.01.006.

[22] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," arXiv:1511.05952, 2015.

[23] G. Pezzulo, F. Rigoli, and K. Friston, "The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning," *Prog. Neurobiol.*, vol. 216, art. 102329, 2022. doi:10.1016/j.pneurobio.2022.102329.

[24] D. Silver et al., "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science*, vol. 362, pp. 1140–1144, 2018. doi:10.1126/science.aar6404.

[25] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, pp. 1–40, 2016.

[26] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," arXiv:1801.01290, 2018.

[27] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.

[28] A. Guez et al., "An investigation of model-free planning," arXiv:1901.03559, 2019.