ICPRL: Acquiring Physical Intuition from Interactive Control

Anonymous authors
Paper under double-blind review

Abstract

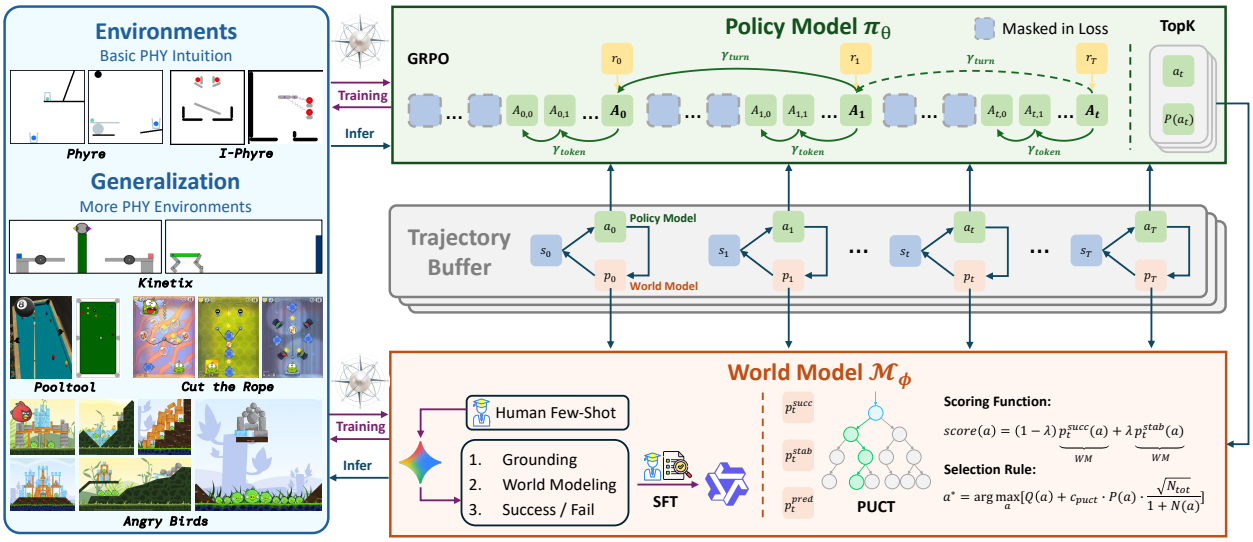

Figure 1: Overview of the **ICPRL** Framework, which decouples policy learning from world modeling for robust in-context planning. **Training Stage:** We separately train a **Policy Model** (π_θ) via **Turn-Aware GRPO** to generate context-aware actions—leveraging γ_{turn} and γ_{token} for precise multi-turn credit assignment—and a **World Model** ($\mathcal{M}_\phi$) to predict physical outcomes. **Inference Stage:** π_θ proposes candidate actions, which $\mathcal{M}_\phi$ evaluates by acting as an *in-context physical simulator*. These evaluations guide a PUCT search to select the optimal action, enabling effective zero-shot planning.

VLMs excel at static perception but falter in interactive reasoning in dynamic physical environments, which demands planning and adaptation to dynamic outcomes. Existing physical reasoning methods often depend on abstract symbolic inputs or lack the ability to learn and adapt from direct, pixel-based visual interaction in novel scenarios. We introduce **ICPRL** (In-Context Physical Reinforcement Learning), a framework inspired by In-Context Reinforcement Learning (ICRL) that empowers VLMs to acquire physical intuition and adapt their policies in-context. Our approach trains a vision-grounded policy model via **multi-turn** Group Relative Policy Optimization (GRPO) over diverse multi-episode interaction histories. This enables the agent to adapt strategies by conditioning on past trial-and-error sequences, without requiring any weight updates. This adaptive policy works in concert with a separately trained world model that provides explicit physical reasoning by predicting the results of potential actions. At inference, the policy proposes candidate actions, while the world model predicts outcomes to guide a root-node PUCT search to select the most promising action. Evaluated on the diverse physics-based puzzle-solving tasks in the DeepPHY benchmark, ICPRL demonstrates significant improvements across both its: **I.** policy-only, and **II.** world-model-augmented stages. Notably, these gains are retained in unseen physical environments, demonstrating that our framework facilitates genuine in-context acquisition of the environment's physical dynamics from interactive experience.

1 Introduction

Despite impressive perception, Vision-Language Models (VLMs) struggle with interactive physical reasoning that requires acting, observing consequences, and revising plans. Existing benchmarks largely test static understanding, leaving a gap in closed-loop, physics-grounded competence (Wang et al., 2023; Agarwal et al., 2025). Prior work either distills policy improvement from symbolic histories, scales meta-RL (Duan et al., 2016) with test-time weight updates, or focuses on state-based off-policy in-context RL (Grigsby et al., 2023); however, none directly equips VLMs to learn physics from raw pixels through interaction, to adapt to new tasks in a zero-shot setting (Keplinger et al., 2025).

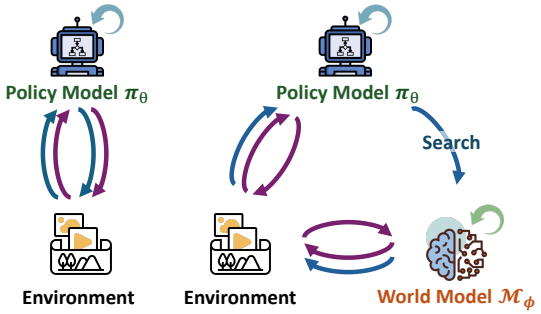

Figure 2: The ICPRL framework integrates world model and adaptive policy.

We introduce **ICPRL** (In-Context Physical Reinforcement Learning), an ICRL (Bauer et al., 2023) -inspired VLM framework that enables in-context policy adaptation for interactive physical reasoning. Our approach reinterprets the concept of distilling a policy improvement operator. Instead of training a model to *imitate* the learning trajectory of a separate RL algorithm, our vision-grounded policy model (π_θ) directly becomes the subject of policy improvement. We train this VLM using **Turn-Aware GRPO** (Shao et al., 2024)—a multi-turn adaptation of Group Relative Policy Optimization designed for precise temporal credit assignment—over a vast collection of diverse, multi-episode interaction histories. This process teaches the model *how* to adjust its strategy based on historical context, including previous successes and failures. Consequently, the ability to adapt to new unseen task instances is not an emergent phenomenon at test time, but rather a robust capability learned and encoded into the model's weights during training. At inference, π_θ leverages this ability to improve its performance purely by conditioning on recent interaction history within its forward pass, requiring no further gradient updates. Crucially, we augment this adaptive policy with an independently trained world model (M_ϕ), which acquires and provides explicit physical intuition by predicting action outcomes and dynamics. This world model serves as a powerful in-context planning component, guiding the policy's exploration and enabling the agent to discover task-specific physics and refine its plans more efficiently (Zhang et al., 2025; Chung et al., 2025).

This two-component architecture is illustrated in Figure 2. Our ICPRL agent (**Right**) consists of: (1) an adaptive **Policy Model** that learns to improve from its interaction history, and (2) an independently trained **World Model** that provides physical intuition. Training interactions for our models are denoted by **purple arrows**. At inference time (**blue arrows**), the policy performs an in-context Search over plans simulated by the world model, enabling more efficient reasoning. This approach stands in contrast to conventional agents (**Left**), which train a policy to react directly to environmental feedback without an explicit planning module.

We evaluate ICPRL on DeepPHY (Xu et al., 2025), a physical benchmark suite spanning single-shot placement to multi-step control, with standardized annotated observations and structured action spaces that preserve physics while enabling reliable VLM control (Wu et al., 2024). On the complex multi-step *I-PHYRE* task, our full model achieves a 93.3% success rate. More critically, ICPRL shows remarkable generalization to unseen environments; *e.g.,* on the *Pooltool* task, it achieves a 71.0% success rate, more than doubling the performance of strong baselines like GPT-o3. These gains on tasks the model was never trained on demonstrate that our framework facilitates genuine in-context acquisition of the environment's physical dynamics from interactive experience.

Building on this evidence, our contributions are twofold: First, we introduce a novel vision-grounded ICRL-inspired framework for VLMs that adapts at test time without weight updates, enabled by a Turn-Aware GRPO mechanism that effectively distills physical intuition from interaction histories. Second, we establish state-of-the-art zero-shot performance on diverse interactive physical reasoning tasks, along with analyses on world-model prediction, action discretization, and history length.

2 Related Works

2.1 In-Context Reinforcement Learning (ICRL)

ICRL represents a paradigm shift in how autonomous agents acquire new skills, moving from slow, gradient-based adaptation to rapid ICL within the forward pass of neural networks. Typically, Algorithm Distillation (AD) (Laskin et al., 2023) distills a policy improvement operator into a Transformer by training on numerous learning histories from a source RL algorithm (*e.g.,* A3C (Mnih et al., 2016)). While methods like AD focus on distilling an external RL algorithm—training a Transformer to *imitate* the learning process itself—our **ICPRL** internalizes this capability directly within the policy model. To circumvent high computational cost, AD$^\epsilon$ (Zisman et al., 2024) creates synthetic learning histories from a single demonstrator policy by simulating a learning process with a decaying noise curriculum. An alternative approach is memory-based meta-RL (Bauer et al., 2023), in which an agent with persistent memory is trained to improve its strategy over repeated trials on the same task. Such techniques have been extended to LLMs by PAPRIKA (Tajwar et al., 2025), fine-tuning on self-generated interaction data using RPO (Pang et al., 2024). SCoRe (Kumar et al., 2025) uses on-policy RL to fine-tune an LLM, teaching it a generalizable ICL skill for multi-turn self-correction. ICAL (Sarch et al., 2024) leverages a VLM to autonomously refine noisy interaction data into high-quality examples, enhancing retrieval-augmented ICL planning.

The challenge of credit assignment in multi-turn LLM interactions has attracted growing attention. ArCHer (Zhou et al., 2024) proposes an off-policy actor-critic framework that assigns per-turn credit through a hierarchical value function, enabling stable RL training over long interaction horizons. DAPO (Yu et al., 2025) addresses reward hacking and gradient instability in long-chain policy optimization at scale. Most directly relevant to our work, VAGEN (Wang et al., 2025a) introduces selective token masking and turn-level advantage estimation for multi-turn VLM agent training—mechanisms we adopt and extend in our Turn-Aware GRPO. Crucially, none of these works targets the generalization of a learned improvement operator across *physically diverse* environments; they are designed for single-task or domain-specific settings, whereas ICPRL is explicitly trained to transfer physical intuition zero-shot to unseen domains.

ICRL can also leverage Unsupervised Environment Design (UED) and automated curriculum learning to generate diverse and high-quality training tasks. Building on XLand (Stooke et al., 2021), which advanced the training of generalist agents through an open-ended process of dynamic task generation, Adaptive Agent (AdA) (Bauer et al., 2023) introduces an automated curriculum method to navigate the immense XLand 2.0 task space. Prioritized Level Replay (PLR) (Jiang et al., 2021b) is a heuristic that focuses training on tasks at the edge of an agent's competence, and has been theoretically formalized and improved (Dennis et al., 2020; Jiang et al., 2021a; Parker-Holder et al., 2022). Recent curriculum strategies have advanced from refining task selection at the learning frontier (Rutherford et al., 2024) to augmenting the training space with generative models (Garcin et al., 2024), and toward creating truly open-ended domains by evolving the game mechanics themselves (Earle & Togelius, 2024; Frans & Isola, 2023).

Instead of mimicking an optimization algorithm, our VLM policy is trained via Turn-Aware GRPO(Shao et al., 2024) on multi-episode histories to learn *how to execute* a policy improvement step in-context. By observing diverse histories of trial and error, the VLM learns to condition its future actions on past outcomes, effectively embedding an adaptive, in-context learning skill into its own parameters. This is achieved through gradient-based optimization over multi-episode histories, enabling the VLM to perform in-context adaptation at inference time.

2.2 Physical Reasoning in VLMs

Physical reasoning serves as the foundation for world model construction (Wu et al., 2024; Agarwal et al., 2025) and embodied intelligence tasks (Luo et al., 2023; Lin et al., 2025; Yuan et al., 2025). However, most evaluations of LLMs and VLMs focus on static problem-solving benchmarks. These evaluations—often large-scale QA tasks on object properties (Wang et al., 2023; Chow et al., 2025) or text-based physics exams (Wang et al., 2025b; Chung et al., 2025; Zhang et al., 2025)—assess primarily the agents' ability to recall scientific knowledge or infer logical outcomes from fixed contexts. While useful for evaluating declarative

knowledge, these approaches largely bypass the challenges of real-time visual perception and continuous interaction in dynamic environments.

To evaluate interactive physical reasoning, another line of research investigates agents in simulated environments. However, these works often abstract away perceptual grounding by relying on pre-processed symbolic inputs, such as object property matrices (Bakhtin et al., 2019; Li et al., 2024; Matthews et al., 2025), or by enabling interaction via code generation (Cherian et al., 2024). Within benchmarks like PHYRE specifically, specialized RL methods have achieved substantially higher performance than general-purpose models by leveraging Deep Q-Learning, relational neural physics models, and improved exploration strategies. While effective for isolating specific planning tasks, these approaches are tightly coupled to the specific simulator, action space, and state representation, providing no path to zero-shot generalization and bypassing raw sensory data understanding. Similarly, while digital game agents (Tan et al., 2025; Chen et al., 2025) process raw observations, they typically require only a shallow understanding of common-sense physics.

A parallel line of research employs learned world models to enable planning in physical environments through model-based RL. For instance, DreamerV3 (Hafner et al., 2025) learns a compact latent-space world model to optimize behavior entirely within imagination, while TD-MPC2 (Hansen et al., 2024) combines model predictive control with learned temporal difference models for continuous control. While these methods demonstrate the power of world models for physical reasoning, they operate on low-level continuous state representations, require environment-specific training from scratch, and provide no mechanism for zero-shot generalization to novel tasks from raw visual observations.

Our work addresses these limitations by focusing on interactive physical reasoning directly from visual inputs, where agents must plan and execute a sequence of actions guided by learned intuition. ICPRL trains a world model on only two source environments and applies it zero-shot as an in-context physical simulator across distinct target domains. Unlike recent candidate-filtering approaches (Qi et al., 2025)—where a world model serves merely as a one-shot, static discriminator to score a fixed set of actions in a single pass—ICPRL enables genuine test-time adaptation through a world-model-guided lookahead search. Specifically, ICPRL's PUCT search *iteratively* queries the world model over B planning iterations, maintaining per-action visit counts $N(a)$ and value estimates $Q(a)$. This progressive search dynamically balances the exploitation of high-scoring candidates with the exploration of less-visited alternatives. Furthermore, ICPRL trains the policy to internalize this improvement operator via Turn-Aware GRPO, removing the reliance on task-specific fine-tuning. We evaluate our approach on DeepPHY (Xu et al., 2025), covering six diverse dynamic physical environments, to demonstrate ICPRL's superior interactive physical reasoning capabilities.

3 ICPRL

We consider a family of physical simulators $\{\mathcal{E}_m\}_{m=1}^{M}$. At each decision point we observe $s_t \in \mathcal{S}$, choose an action $a_t \in \mathcal{A}$, and the simulator returns an updated observation and a task reward or success indicator. Our ICPRL framework comprises distinct training and inference stages, as shown in Figure 1: **Training (two stages):** We foster **in-context adaptive capabilities** within a VLM **policy model** π_θ through online GRPO (Shao et al., 2024) on diverse multi-episode interaction histories. Concurrently, a separate **world model** $\mathcal{M}_\phi$ is trained offline to acquire robust physical intuition by predicting environment dynamics and outcomes. **Inference:** These two models work in concert. π_θ proposes contextually informed actions, and $\mathcal{M}_\phi$, acting as an **in-context physical simulator**, provides crucial outcome predictions to guide a lightweight **root-node PUCT** (Silver et al., 2017) search. This separation keeps policy π_θ learning stable and simulator-grounded, while enabling rapid context-sensitive adaptation and refined planning at test time without further weight updates.

We formalize the interactive physical reasoning task as a multi-step Partially Observable Markov Decision Process (POMDP) (Kaelbling et al., 1998) augmented with interaction history. At each step, the agent receives a visual observation $o_t \in \mathcal{O}$ and selects a discretized action $a_t \in \mathcal{A}$, receiving a sparse binary reward $r_{GT} \in \{0, 1\}$ at episode end. The policy $\pi_\theta(a_t \mid o_{\leq t}, H)$ conditions on both the current observation sequence and the interaction history $H_k = (\tau_1, \ldots, \tau_{k-1})$ of past attempts. Full formal definitions of all components are provided in Appendix A.

3.1 Training Stage: Policy Model π_θ

Our policy model π_θ is a VLM that generates textual outputs in a structured format (detailed in Section 3.3). These outputs are parsed into discrete actions $a_t \in \mathcal{A}$ by an environment-specific converter, for interaction with simulators across M environments. The reinforcement signal is derived directly from the task's native rewards (*e.g.,* success/failure).

While standard RL algorithms can be applied to this setup by treating the agent's trajectory as a monolithic sequence, this approach often proves suboptimal. It fails to distinguish between agent-generated tokens (reasoning and actions) and environment-provided context, and it applies a uniform temporal discounting that conflates intra-turn and inter-turn credit assignment. To address these limitations, we enhance our Group Relative Policy Optimization (GRPO) framework by incorporating principles from selective token masking and multi-turn credit assignment, inspired by VAGEN framework (Wang et al., 2025a).

First, we employ selective token masking to focus the learning signal strictly on the agent's decision-making process. We introduce a mask $M_t^{\text{loss}} \in \{0, 1\}$, where $M_t^{\text{loss}} = 1$ for all tokens generated by the policy π_θ (reasoning and action tokens) and $M_t^{\text{loss}} = 0$ for all observation and prompt tokens. Second, for multi-turn interaction histories, standard GRPO assigns a single sequence-level advantage, which leads to sparse and unstable gradients over long horizons. To resolve this, we adopt a Turn-Aware Group Advantage Estimation scheme. Instead of a single sequence-level advantage, we calculate a turn-specific advantage $A_{i,k}$ for the k-th turn of the i-th sampled trajectory within a group. This effectively implements an inter-turn discount γ_{turn} for future state transitions, while treating all generated tokens within a single turn (intra-turn, $\gamma_{\text{token}} = 1.0$) as equally responsible for that turn's outcome.

Formally, at a given decision step, we sample a group of G output sequences from the policy model π_θ, denoted as $\mathcal{G} = \{o_1, o_2, \ldots, o_G\}$. The importance sampling ratio for a given token t in the i-th sample is defined as $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$.

Let $A_{i,k}$ denote the relative advantage for the k-th turn of the i-th trajectory, computed by standardizing the turn-specific returns within the group:

$$A_{i,k} = \frac{R_{i,k} - \text{mean}(R_{*,k})}{\text{std}(R_{*,k})} \tag{1}$$

where $R_{i,k}$ is the discounted return from turn k onwards ($R_{i,k} = \sum_{l=k}^{K} \gamma_{\text{turn}}^{l-k} r_{i,l}$). For a trajectory with K turns, the token-level GRPO objective function (to be maximized) is formulated as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\sum_{k=1}^{K}\frac{1}{\sum_t M_{i,k,t}^{\text{loss}}}\sum_{t\in\text{turn}_k} M_{i,k,t}^{\text{loss}}\Bigg[\min\left(r_{i,t}(\theta)\cdot A_{i,k},\ \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\cdot A_{i,k}\right) - \beta D_{\text{KL}}\left(\pi_\theta(\cdot \mid q, o_{i,<t}) \,\|\, \pi_{\text{ref}}(\cdot \mid q, o_{i,<t})\right)\Bigg] \tag{2}$$

Here, $M_{i,k,t}^{\text{loss}}$ indicates whether the t-th token in turn k is generated by the policy, ε is the clipping hyperparameter, and β is the coefficient for the Kullback-Leibler (KL) divergence penalty against the frozen reference model π_{ref}. By coupling GRPO with multi-turn masking, we achieve precise token-level credit assignment without the computational overhead of training a separate Critic model.

3.2 Training Stage: World Model $\mathcal{M}_\phi$

Given an observation-action pair (o, a), the world model $\mathcal{M}_\phi$ is trained to return a strictly structured JSON containing a predicted success probability and a natural language prediction.

$$\mathcal{M}_\phi(o, a) \Rightarrow \begin{cases} \hat{p}_{\text{succ}} \in [0, 1], \\ \hat{p}_{\text{pred}} : \text{NL prediction} \end{cases} \tag{3}$$

The predicted success probability $\hat{p}_{\text{succ}}$ is used directly for planning, while $\hat{p}_{\text{pred}}$ provides qualitative insights for analysis.

To enhance planning robustness, a stability score, $\hat{p}_{\text{stab}}$, is derived from the model's own predictions. This score is calculated by evaluating the model's success predictions over a neighborhood of perturbed actions. For an environment-specific action metric $d_{\mathcal{A}}$ and a radius δ, we define the perturbation set as: $\mathcal{B}_\delta(a) = \{a' : d_{\mathcal{A}}(a, a') \leq \delta\}$. The stability $\hat{p}_{\text{stab}}$ is the expected model-predicted success rate over actions sampled from this perturbation set. Let $\hat{p}_{\text{succ}}(o, a')$ denote the success probability predicted by $\mathcal{M}_\phi(o, a')$. The stability is then defined as: $\hat{p}_{\text{stab}}(o, a) = \mathbb{E}_{a' \sim D(\mathcal{B}_\delta(a))}[\, \hat{p}_{\text{succ}}(o, a')\,]$.

In practice, this expectation is estimated for each sample by querying the model $\mathcal{M}_\phi$ with a small, finite set of "jittered" actions and averaging the resulting $\hat{p}_{\text{succ}}$ predictions.

To train world model $\mathcal{M}_\phi$, we first curate a high-quality dataset with a balanced representation of outcomes (formalized in Algorithm 2 in Appendix). The data generation process for each initial state s is as follows:

We first enumerate all possible successful actions from state s. If k successful actions are found, we then sample and filter to obtain a corresponding set of k diverse, failing actions. This ensures a balanced 1:1 ratio of positive to negative examples for each state, preventing model bias. In practice, this loop terminates rapidly. In physical puzzle environments, the vast majority of actions result in failure; successful actions occupy only a small fraction of the action space. Consequently, each random sample drawn from $\mathcal{A}$ is overwhelmingly likely to be a failing action, ensuring that k diverse failure samples are collected within a small, bounded number of iterations across all evaluated environments.

Each successful and failing action is executed in the simulator. We record the entire chain reaction following the action, from its completion until the environment reaches a stable state. From this resulting video sequence, we uniformly sample 5 frames to represent the dynamic evolution of the environment post-action. The collected data—comprising the initial state (raw and annotated images), the action, the five dynamic frames, and the ground-truth success/fail label—is used to generate rich training signals. We employ a large VLM with a few-shot prompting strategy to produce the NLP ground truth for $\hat{p}_{\text{pred}}$. The VLM is prompted to generate text describing the objects and their relations in the initial state (**Grounding**), predicting the chain reaction that will occur as a consequence of the action (**World Modeling**), and providing a **Success or Fail Label**. All automatically generated annotations are then manually inspected and corrected by human reviewers to ensure their accuracy and quality.

With this curated dataset, we train world model $\mathcal{M}_\phi$. The training objective combines a regression loss for success prediction and a language modeling loss for the textual output. The total loss $\mathcal{L}_{\text{WM}}$ is a weighted sum: $\mathcal{L}_{\text{WM}} = \mathcal{L}_{\text{succ}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}}$, where $\mathcal{L}_{\text{succ}} = \text{BCE}(\hat{p}_{\text{succ}}, y)$ is the Binary Cross-Entropy loss between the predicted success probability and the ground-truth label y, and $\mathcal{L}_{\text{text}}$ is a standard cross-entropy loss for training the language prediction component.

3.3 Inference Stage

At inference time, for each decision point, we employ a root-node search procedure that integrates a learned policy prior with scores from our world model, $\mathcal{M}_\phi$. This process (detailed in Algorithm 1 in Appendix) unfolds in four stages: candidate generation, scoring, search, and execution.

Candidate Generation. First, to form a discrete set of candidate actions $A = \{a_i\}_{i=1}^{S}$, we draw S samples from our policy network $\pi_\theta(\cdot \mid o)$. Based on these samples, we establish a **frequency prior** $P(a)$ over the candidate set, defined as $P(a) = c(a)/\sum_{a' \in A} c(a')$, where $c(a)$ is the count of a.

Scoring Strategy. Next, each candidate action $a \in A$ is evaluated by $\mathcal{M}_\phi$. To handle the distinct characteristics of varying physical environments—ranging from stable to chaotic dynamics—we employ two tailored scoring strategies (detailed per-environment in Table 7 in Appendix C).

Strategy 1: Action Space Perturbation. For environments with stable physics (*PHYRE*, *I-PHYRE*, *Angry Birds*), we estimate stability by sampling a neighborhood of J perturbed actions $\{a'_j\}_{j=1}^{J}$. The stability score is the average success probability of these neighbors: $\hat{p}_{\text{stab}}(o, a) = \frac{1}{J}\sum_{j=1}^{J} \hat{p}_{\text{succ}}(o, a'_j)$. Specifically:

- *PHYRE*: We perturb grid coordinates (x, y) by ± 1 and radius r by ± 1 ($J = 12$).
- *I-PHYRE*: We apply temporal jittering $\delta \in \{\pm 0.5\text{s}, \pm 1.0\text{s}\}$ to event timestamps, ensuring causal order is preserved.
- *Angry Birds*: We perturb launch angle θ by $\pm 5°$ and power p by ± 0.1.

We combine the mean success prediction μ_p and this stability score into: $\text{score}(a \mid o) = (1 - \lambda_{\text{PUCT}})\,\mu_p(o, a) + \lambda_{\text{PUCT}}\,\hat{p}_{\text{stab}}(o, a)$.

Strategy 2: Model Confidence (LCB). For environments sensitive to initial conditions (*Kinetix*, *Pooltool*, *Cut the Rope*), we instead estimate the model's epistemic uncertainty. We perform $K = 8$ stochastic forward passes with temperatures sampled from $[0.1, 1.0]$ to obtain success probabilities $\{p^{(j)}\}_{j=1}^{K}$. We then compute a Lower Confidence Bound (LCB) score: $\text{score}(a|o) = \mu_p - \lambda_{\text{LCB}} \cdot \sigma_p$, where μ_p and σ_p are the mean and standard deviation of predictions, and $\lambda_{\text{LCB}} = 0.2$. This penalizes uncertain actions in more chaotic domains.

PUCT Search. These scores guide a search at the root node. For a planning budget of B iterations, we select the action a_t that maximizes the PUCT criterion:

$$a_t = \arg\max_{a \in A} \left[Q(a) + c_{\text{PUCT}} \cdot P(a) \cdot \frac{\sqrt{\sum_{b \in A} N(b)}}{1 + N(a)} \right] \tag{4}$$

Upon selecting a_t, its world model score is used to update $Q(a_t)$ and $N(a_t)$. We employ an **early-stopping** mechanism if the budget B is exhausted, the same action is chosen 3 times consecutively, or a highly successful action ($\mu_p > 0.8$) is found. Note that unlike standard MCTS where $Q(a)$ is updated via actual environment rollouts, here $Q(a)$ is maintained as a running average of world model scores v, serving as a proxy value estimate within the root-node planning budget.

Execution. Finally, the action $a^* = \arg\max_a Q(a)$ is executed for a single step.

4 Experiments

4.1 Experimental Setup

This subsection details the experimental environments, comparison baselines, evaluation metrics, and the implementation details of ICPRL.

4.1.1 Environments & Task Specifications

We adopt DeepPHY (Xu et al., 2025) as our evaluation platform. This benchmark comprises six diverse and interactive physical simulation environments: ***PHYRE*** (Bakhtin et al., 2019), ***I-PHYRE*** (Li et al., 2024), ***Kinetix*** (Matthews et al., 2025), ***Pooltool*** (Kiefl, 2024), ***Angry Birds***[1], and ***Cut the Rope***[2]. Encompassing a wide spectrum of physical properties (*e.g.,* gravity, elasticity, collisions), these environments present challenges ranging from single-step planning (*PHYRE*) to complex multi-step sequential control (*Kinetix*, *Cut the Rope*).

To ensure a rigorous interactive evaluation, we formally define the task structure and success metrics below:

[1] https://apps.apple.com/us/app/rovio-classics-angry-birds/id1596736236
[2] https://apps.apple.com/cn/app/cut-the-rope/id1024507512

Formal Definitions.

- **Episode (Task Instance):** An episode refers to the complete problem-solving process for a single specific level or puzzle configuration (*e.g.,* PHYRE template 00002:001). An episode concludes when the agent successfully solves the task or exhausts the maximum allowed attempts.
- **Attempt (Trial):** An episode consists of a sequence of up to K attempts. At the start of attempt k, the agent receives the visual observation and the text history of previous failed trajectories $H_{k-1} = \{\tau_1, ..., \tau_{k-1}\}$ to perform in-context learning. For PHYRE and I-PHYRE (training environments), the limit is $K = 10$. For evaluation environments like Pooltool and Angry Birds, limits are set according to Table 2.
- **Action Horizon (Per Attempt):** This defines the temporal complexity of a single plan within one attempt.
  - *Single-step / In-advance:* The agent outputs a complete static plan (*e.g.,* placing one ball in PHYRE) or a sequence of timed actions (I-PHYRE) at the start. The simulator then executes the physics until stability.
  - *Sequential / On-the-fly:* The agent interacts with the environment in a turn-based manner (*e.g.,* Kinetix, Angry Birds), observing intermediate states between actions.

Environment Specifications. We provide a detailed summary of the input modalities, action spaces, and success criteria for all environments in Table 1 and Table 2.

Table 1: Detailed Specifications of DeepPHY Environments (Part I): Input Modalities and Action Spaces.

Environment	Input Modality	Action Space Format
PHYRE	Image + 8×8 Grid Overlay	Discretized Selection: `Cell: [1-64], Radius: [1-8]`
I-PHYRE	Image + Index IDs	JSON Sequence: `[{"time": t, "index": i}, ...]`
Kinetix	Image + Motor/Thruster IDs	JSON Vector: `[m_1, m_2, ..., t_1, t_2]`
Pooltool	2D Top-down View	Discretized Selection: `Speed: [Low/Med/High], Strikespot: [Spin]`
Angry Birds	Screenshot (Annotated)	Code: `[shoot(angle=int, power=float)]`
Cut the Rope	Screenshot (Annotated)	Code: `[cut_pin(id)], [pop_bubble(id)]`, etc.

Table 2: Detailed Specifications of DeepPHY Environments (Part II): Horizons and Success Criteria.

Environment	Horizon	Success Criteria ($r_{GT} = 1$)
PHYRE	Single-step	Green ball touches Target.
I-PHYRE	Sequence	All Red balls fall into abyss.
Kinetix	Multi-step (Max 16)	Green shape touches Blue, avoids Red.
Pooltool	Multi-step (Max 15)	Pot the 9-ball legally.
Angry Birds	Multi-step (Birds avail.)	All pigs are eliminated.
Cut the Rope	Multi-step	Candy reaches *Om Nom*'s mouth.

4.1.2 Evaluation Protocol

To rigorously assess the acquisition of transferable physical intuition, we strictly define our evaluation protocol by partitioning the DeepPHY suite into source and target domains.

We train the Policy Model (π_θ) and World Model ($\mathcal{M}_\phi$) exclusively on *PHYRE* and *I-PHYRE*. The models are evaluated on *Kinetix*, *Pooltool*, *Angry Birds*, and *Cut the Rope*. We enforce a strict zero-shot inference setting on the Target Domains. No weight updates, fine-tuning, or parameter calibration are permitted on these environments. The agent must rely solely on its learned physical intuition and in-context adaptation capabilities to solve these unseen tasks.

4.1.3 Baselines

We compare ICPRL with leading closed-source and open-source VLMs, as detailed in Table 6. These closed-source models represent the current SOTA in zero-shot physical reasoning. We use the random action agent (*MOCK*) (provided by DeepPHY) as a lower bound on performance.

We evaluate two distinct variants of our own method: i) **ICPRL (Policy Only):** This variant performs inference using only our policy model, π_θ, trained with GRPO, without leveraging the world model or PUCT search. This baseline is designed to measure the performance of the adaptive policy itself. And ii) **ICPRL (Full):** Our complete model, which combines the adaptive policy π_θ with the PUCT search guided by World Model $\mathcal{M}_\phi$, showcasing how both work in tandem.

4.1.4 Evaluation Metrics

We employ two core metrics for our evaluation: **Success Rate (S.R.)**, which is the percentage of tasks solved successfully; and **Average Attempts** (Avg. Att.), the mean number of attempts taken, calculated only on successfully solved tasks. We report Success Rate and Average Attempts over three runs under the different settings.

4.1.5 Implementation Details

Model Architecture: Both Policy Model π_θ and World Model $\mathcal{M}_\phi$ are built and trained based on Qwen2.5-VL-3B/7B-Instruct (Qwen Team, 2025).

Training: We conducted training for Policy Model π_θ and World Model $\mathcal{M}_\phi$ on the first two environments (*PHYRE* & *I-PHYRE*[3]), and then evaluated their generalization performance on the remaining four. Policy Model π_θ is trained online using GRPO (Shao et al., 2024). For training, we set the policy learning rate to 1×10^{-6}. The policy updates are regularized with a KL penalty coefficient (β) of 0.001. We employ a Turn-Aware Group Advantage Estimation scheme with a turn-level discount factor $\gamma_{\text{turn}} = 0.95$ and a token-level discount $\gamma_{\text{token}} = 1.0$. During the online data collection (rollout) phase, actions are sampled using a temperature of 0.7 and top-p nucleus sampling of 0.95. World Model $\mathcal{M}_\phi$ is trained offline on the curated dataset, with the loss weight for the text component, λ_{text}, set to 0.2.

Inference: The PUCT search is configured with a candidate sample size of $S = 32$, a planning budget of $B = 32$, a score mixing weight of $\lambda_{\text{PUCT}} = 0.25$, and an exploration coefficient of $c_{\text{PUCT}} = 1.5$. For environments using Strategy 1 (Stability), we use $J = 12$ perturbed neighbors. For environments using Strategy 2 (LCB), we use $K = 8$ stochastic forward passes with $\lambda_{\text{LCB}} = 0.2$.

4.2 Main Results

The main results, shown in Table 3, provide a comprehensive evaluation of the **ICPRL** framework and its components. Unlike the original DeepPHY benchmark, which evaluates on the complete task pool, we report **S.R.** on a held-out test split for *PHYRE* and *I-PHYRE*, as these environments are used for training π_θ and $\mathcal{M}_\phi$. **Crucially, all models in this table were re-evaluated under this identical protocol**, ensuring a fair and consistent comparison. Their analysis validates our core hypotheses regarding in-context physical reinforcement learning, the role of world models, and the importance of training paradigms for generalization in complex physical reasoning tasks.

Our finetuned methods, including both SFT and GRPO, significantly improve the performance of the utilized base open-source VLM, Qwen2.5-VL. The base models struggle to effectively solve these tasks, which demonstrates that our training pipeline successfully triggers and guides the VLMs to acquire physical intuition. Furthermore, after undergoing the identical fine-tuning process, the 7B models consistently outperform their 3B counterparts, highlighting the benefits of scaling model capacity. The premier configuration, **Qwen-7B$^{\text{P\&I GRPO}}$** paired with **Qwen-7B$^{\text{P\&I}}$** (Row #36), consistently achieves state-of-the-art or comparable performance across nearly all tasks. This result validates our central hypothesis: the synergy between an

[3]Although similarly named, the environments pose distinct challenges: *PHYRE* centers on single-step actions that trigger complex chain reactions, while *I-PHYRE* requires multi-step planning with precise temporal and sequential control.

Table 3: Overall Performance on DeepPHY. Model naming conventions are detailed in Appendix.

Model Configuration			In-Domain (Training)				Out-of-Distribution (Zero-shot)				
Row Ann.	Policy Model π_θ	World Model $\mathcal{M}_\phi$	PHYRE		I-PHYRE		Kinetix	Pooltool		Angry Birds	Cut the Rope
			Att. 1	Att. 10	Att. 1	Att. 10		Att. 1	Att. 15		
# 01.	MOCK	-	1.50%	10.80%	23.33%	53.33%	21.40%	2.33%	48.00%	17.65%	11.36%
Open-Source Models											
# 02.	Qwen-3B	-	0.17%	5.85%	13.33%	16.67%	16.22%	0.00%	50.00%	17.65%	7.95%
# 03.	Qwen-7B	-	0.67%	10.10%	32.42%	32.42%	13.51%	23.50%	26.50%	20.59%	9.09%
# 04.	Qwen-32B	-	0.03%	8.70%	0.17%	5.60%	15.20%	0.00%	14.29%	26.47%	6.82%
# 05.	Qwen-72B	-	2.43%	14.92%	13.33%	43.33%	14.86%	0.00%	18.00%	29.41%	13.64%
# 06.	Qwen-72B	Qwen-72B	1.78%	10.39%	10.00%	36.67%	12.16%	0.00%	14.00%	26.47%	11.50%
Close-Source Models											
# 07.	Claude 4.0 Opus	-	1.73%	10.63%	36.67%	56.67%	23.20%	0.00%	49.00%	32.35%	26.14%
# 08.	Claude 4.0 Opus	Claude 4.0 Opus	1.11%	7.44%	30.00%	50.00%	20.50%	0.00%	42.00%	29.41%	23.86%
# 09.	Gemini-2.5-Pro	-	2.17%	22.07%	20.00%	63.33%	24.10%	36.50%	68.00%	35.29%	22.73%
# 10.	Gemini-2.5-Pro	Gemini-2.5-Pro	2.17%	12.33%	16.67%	53.33%	21.80%	25.00%	60.00%	29.41%	20.45%
# 11.	GPT-o3	-	3.03%	30.77%	30.00%	86.67%	26.89%	0.00%	25.67%	35.29%	18.18%
# 12.	GPT-o3	GPT-o3	0.17%	25.60%	25.00%	76.67%	24.50%	0.00%	22.00%	29.41%	16.50%
Fine-tuned Models											
Qwen2.5-VL-3B-Instruct Fine-tuned Series											
# 13.	$\text{Qwen-3B}^{\text{P}}$ *SFT*	-	1.33%	25.58%	20.00%	30.00%	13.70%	0.00%	51.00%	23.53%	10.23%
# 14.	$\text{Qwen-3B}^{\text{P}}$ *SFT*	$\text{Qwen-3B}^{\text{P}}$	1.52%	28.69%	19.87%	32.45%	13.55%	0.00%	53.67%	26.47%	11.85%
# 15.	$\text{Qwen-3B}^{\text{I}}$ *SFT*	-	0.33%	9.83%	23.33%	45.56%	9.60%	0.00%	43.00%	20.59%	7.80%
# 16.	$\text{Qwen-3B}^{\text{I}}$ *SFT*	$\text{Qwen-3B}^{\text{I}}$	0.74%	11.88%	23.21%	47.34%	9.82%	0.00%	45.33%	23.53%	9.13%
# 17.	$\text{Qwen-3B}^{\text{P\&I}}$ *SFT*	-	10.42%	29.00%	20.00%	50.00%	15.07%	0.00%	57.00%	26.47%	11.60%
# 18.	$\text{Qwen-3B}^{\text{P\&I}}$ *SFT*	$\text{Qwen-3B}^{\text{P\&I}}$	10.70%	33.56%	21.29%	52.91%	15.11%	0.00%	59.67%	32.35%	14.05%
# 19.	$\text{Qwen-3B}^{\text{P}}$ *GRPO*	-	9.50%	29.03%	13.33%	26.67%	12.33%	0.00%	58.00%	26.47%	11.05%
# 20.	$\text{Qwen-3B}^{\text{P}}$ *GRPO*	$\text{Qwen-3B}^{\text{P}}$	9.85%	32.00%	12.78%	28.00%	12.33%	0.00%	61.00%	29.41%	12.98%
# 21.	$\text{Qwen-3B}^{\text{I}}$ *GRPO*	-	0.17%	9.67%	36.67%	48.89%	8.20%	0.00%	50.00%	20.59%	7.75%
# 22.	$\text{Qwen-3B}^{\text{I}}$ *GRPO*	$\text{Qwen-3B}^{\text{I}}$	0.66%	13.48%	37.60%	51.00%	8.07%	0.00%	52.33%	23.53%	9.88%
# 23.	$\text{Qwen-3B}^{\text{P\&I}}$ *GRPO*	-	14.00%	34.50%	36.67%	60.00%	19.18%	0.00%	55.00%	29.41%	12.10%
# 24.	$\text{Qwen-3B}^{\text{P\&I}}$ *GRPO*	$\text{Qwen-3B}^{\text{P\&I}}$	14.11%	39.49%	35.66%	61.99%	19.45%	0.00%	57.33%	35.29%	15.80%
Qwen2.5-VL-7B-Instruct Fine-tuned Series											
# 25.	$\text{Qwen-7B}^{\text{P}}$ *SFT*	-	3.67%	36.13%	13.33%	19.33%	15.07%	2.00%	66.00%	38.24%	15.35%
# 26.	$\text{Qwen-7B}^{\text{P}}$ *SFT*	$\text{Qwen-7B}^{\text{P}}$	3.40%	38.67%	13.57%	22.15%	15.31%	2.00%	67.33%	41.18%	17.50%
# 27.	$\text{Qwen-7B}^{\text{I}}$ *SFT*	-	0.67%	9.33%	12.22%	50.00%	13.70%	1.00%	64.00%	23.53%	8.21%
# 28.	$\text{Qwen-7B}^{\text{I}}$ *SFT*	$\text{Qwen-7B}^{\text{I}}$	0.82%	13.45%	11.34%	51.50%	13.49%	1.00%	66.67%	26.47%	10.36%
# 29.	$\text{Qwen-7B}^{\text{P\&I}}$ *SFT*	-	7.50%	42.67%	20.00%	63.33%	15.07%	2.00%	67.00%	41.18%	16.55%
# 30.	$\text{Qwen-7B}^{\text{P\&I}}$ *SFT*	$\text{Qwen-7B}^{\text{P\&I}}$	7.95%	44.85%	21.41%	65.38%	14.88%	2.00%	69.00%	44.12%	19.33%
# 31.	$\text{Qwen-7B}^{\text{P}}$ *GRPO*	-	13.00%	40.00%	16.67%	28.89%	13.70%	0.00%	66.00%	41.18%	16.10%
# 32.	$\text{Qwen-7B}^{\text{P}}$ *GRPO*	$\text{Qwen-7B}^{\text{P}}$	13.18%	43.66%	15.68%	30.00%	13.91%	0.00%	69.00%	44.12%	18.88%
# 33.	$\text{Qwen-7B}^{\text{I}}$ *GRPO*	-	0.67%	8.33%	26.67%	86.67%	16.44%	0.00%	60.00%	23.53%	7.95%
# 34.	$\text{Qwen-7B}^{\text{I}}$ *GRPO*	$\text{Qwen-7B}^{\text{I}}$	0.67%	13.04%	27.82%	89.31%	16.44%	0.00%	63.00%	26.47%	10.15%
# 35.	$\text{Qwen-7B}^{\text{P\&I}}$ *GRPO*	-	14.63%	43.17%	13.33%	90.00%	19.18%	0.00%	69.00%	45.61%	17.05%
# 36.	$\text{Qwen-7B}^{\text{P\&I}}$ *GRPO*	$\text{Qwen-7B}^{\text{P\&I}}$	14.92%	45.56%	13.12%	93.33%	18.96%	0.00%	71.00%	47.06%	21.20%

adaptive policy (π_θ), trained via GRPO on multi-episode interaction histories, and an independently-trained world model ($\mathcal{M}_\phi$), that guides planning, is critical for mastering complex physical challenges. Notably, this model not only excels in the trained environments (*PHYRE* and *I-PHYRE*) but also exhibits remarkable generalization to entirely unseen tasks in the other four diverse settings, demonstrating the acquisition of robust and transferable physical intuition.

A direct comparison between models trained with SFT (*e.g.,* Row #29) and those trained with GRPO (*e.g.,* Row #35) reveals the clear superiority of the latter. While SFT enables learning from expert trajectories, GRPO's online policy optimization teaches the agent to dynamically **adapt its strategy in-context** by conditioning on a history of successes and failures. This capability is fundamental to the ICRL paradigm, resulting in a more robust and adaptive policy that is particularly effective in tasks that inherently involve trial-and-error and iterative problem-solving.

The contribution of World Model ($\mathcal{M}_\phi$) is pivotal, transforming the agent from a purely reactive decision-maker into a deliberative planner. Across all finetuned pairs in the table, the full ICPRL configuration (Policy + World Model) consistently outperforms the policy-only variant. This confirms that the world model, acting as an in-context physical simulator, provides crucial foresight. The resulting "propose-verify-select" mechanism, where the policy generates candidate actions and the world model guides a PUCT search to select the most promising one, elevates the agent's reasoning from simple reaction to improved look-ahead planning. Conversely, employing a generic, non-fine-tuned VLM as a world model (comparing pairs from Rows #5 vs. #6 to #11 vs. #12) degrades performance. This finding, consistent with the original DeepPHY, underscores that general-purpose VLMs currently lack fine-grained interactive physical reasoning capabilities.

Moreover, models trained jointly on both *PHYRE* and *I-PHYRE* (marked as **P&I**) demonstrate superior performance on unseen tasks, when compared to models trained on either environment in isolation (*cf.* Rows #31, #33, and #35). Specifically, the **Qwen-7B$^{\text{P\&I GRPO}}$** policy (Row #35) outperforms its single-environment counterparts across three unseen testbeds. This supports our hypothesis that exposure to diverse physical dynamics enables the model to internalize a more generalizable "policy-improvement operator," facilitating the transfer of learned physical intuition to new domains.

In the *Kinetix* environment, however, our models do not exhibit the same magnitude of performance gain as seen in the other environments. We attribute this to its unique nature, which demands precise fine-grained control of sub-object components and their tight interactions. In contrast, the other environments primarily involve reasoning about the holistic behavior and trajectory of whole objects.

Cut the Rope presents a distinct challenge of its own: it combines GUI-based interaction (cutting ropes, popping bubbles, and timing actions) with complex multi-body physics simulation, making it arguably the most demanding environment in the benchmark. Despite this, ICPRL achieves 21.20% success rate—more than doubling the performance of the untrained Qwen-7B base model (9.09%)—demonstrating that the physical intuition acquired during training does transfer meaningfully to this compound interactive domain, even under strict zero-shot conditions.

4.3 Ablation Studies

Our primary SFT models are trained on a dataset of successful solution trajectories from expert models (Gemini-2.5-Pro, GPT-4o, etc., detailed in Appendix), which often include multiple attempts (typically 5–10) before reaching a solution. This strategy is founded on the hypothesis that exposure to this trial-and-error process, even within a supervised framework, enables the model to internalize an iterative problem-solving strategy, aligning with the principles of ICRL.

To validate this hypothesis, we conduct a controlled ablation study detailed in Table 4. We compare our standard SFT models against ablated variants (denoted with a 'single' subscript). These ablated models are trained exclusively on a dataset of first-attempt successful trajectories generated via a systematic enumeration process. To ensure the comparison is fair, the number of training samples was kept identical for both SFT methods within the same environment.

Table 4: **Ablation Study:** Performance of Multi-Attempt vs. Single-Attempt Training Trajectories.

(a) *PHYRE* Benchmark.

Policy Model Only	Avg. Att.	Att. 1	Att. 4	Att. 7	Att. 10
MOCK	5.00	1.50%	5.87%	8.60%	10.80%
Qwen-3B	3.98	0.17%	4.69%	5.67%	5.85%
Qwen-3B$^{\text{P}_{\text{single}}}$ *SFT*	2.60	7.63%	17.20%	19.83%	20.97%
Qwen-3B$^{\text{P}}$ *SFT*	2.89	1.33%	16.58%	20.33%	25.58%
Qwen-3B$^{\text{P}}$ *GRPO*	3.23	9.50%	25.13%	28.80%	29.03%
Qwen-7B	5.10	0.67%	5.33%	8.17%	10.10%
Qwen-7B$^{\text{P}_{\text{single}}}$ *SFT*	2.61	13.17%	21.58%	26.97%	29.14%
Qwen-7B$^{\text{P}}$ *SFT*	3.74	3.67%	27.33%	34.20%	36.13%
Qwen-7B$^{\text{P}}$ *GRPO*	5.70	13.00%	33.17%	41.77%	40.00%
Qwen-32B	3.97	0.03%	3.10%	6.93%	8.70%
Qwen-72B	4.48	2.43%	9.53%	12.40%	14.92%

(b) *I-PHYRE* Benchmark.

Policy Model Only	Avg. Att.	Att. 1	Att. 4	Att. 7	Att. 10
MOCK	3.81	23.33%	43.33%	46.67%	53.33%
Qwen-3B	2.87	13.33%	16.67%	16.67%	16.67%
Qwen-3B$^{\text{I}_{\text{single}}}$ *SFT*	2.97	33.43%	33.43%	34.54%	34.54%
Qwen-3B$^{\text{I}}$ *SFT*	3.72	23.33%	30.00%	43.33%	45.56%
Qwen-3B$^{\text{I}}$ *GRPO*	3.81	36.67%	43.33%	45.56%	48.89%
Qwen-7B	2.20	32.42%	32.42%	32.42%	32.42%
Qwen-7B$^{\text{I}_{\text{single}}}$ *SFT*	3.44	35.56%	35.56%	35.56%	35.56%
Qwen-7B$^{\text{I}}$ *SFT*	3.53	12.22%	46.67%	50.00%	50.00%
Qwen-7B$^{\text{I}}$ *GRPO*	4.37	26.67%	43.33%	53.33%	86.67%
Qwen-32B	1.40	0.17%	5.60%	5.60%	5.60%
Qwen-72B	2.00	13.33%	30.00%	43.33%	43.33%

While the 'single' models demonstrate strong initial performance (Att. 1), they often plateau, a behavior particularly pronounced on the *I-PHYRE* benchmark. In contrast, models trained on multi-attempt data exhibit a significantly steeper improvement curve, ultimately achieving superior performance by the final attempts (Att. 10). This confirms that learning from a history of failures and recoveries is crucial for developing a more robust policy. Furthermore, the success rate for our models trained on multi-attempt trajectories consistently and monotonically increases as the interaction history lengthens. A compelling trend is found where the average number of attempts for successful solutions (Avg. Att.) tends to increase in tandem with the model's capability (*e.g.,* from **SFT** to **GRPO** variants). This suggests that more advanced models are not simply more efficient, but are capable of tackling more complex problems that inherently require a longer iterative process. This provides direct empirical evidence that the models have learned

Table 5: **Ablation Study:** Impact of Dynamic Visual Feedback on World Model Training.

(a) *PHYRE* Benchmark.

Policy Model	World Model	Att. 1	Att. 10
Qwen-3B^{P} *SFT*	-	1.33%	25.58%
Qwen-3B^{P} *SFT*	**Qwen-3B$^{P}_{w/o\ 5\ Frames}$**	1.40%	27.19%
Qwen-3B^{P} *SFT*	**Qwen-3B^{P}**	1.52%	28.69%
Qwen-3B^{P} *GRPO*	-	9.50%	29.03%
Qwen-3B^{P} *GRPO*	**Qwen-3B$^{P}_{w/o\ 5\ Frames}$**	9.95%	30.20%
Qwen-3B^{P} *GRPO*	**Qwen-3B^{P}**	9.85%	32.00%
Qwen-7B^{P} *SFT*	-	3.67%	36.13%
Qwen-7B^{P} *SFT*	**Qwen-7B$^{P}_{w/o\ 5\ Frames}$**	3.40%	37.47%
Qwen-7B^{P} *SFT*	**Qwen-7B^{P}**	3.40%	38.67%
Qwen-7B^{P} *GRPO*	-	13.00%	40.00%
Qwen-7B^{P} *GRPO*	**Qwen-7B$^{P}_{w/o\ 5\ Frames}$**	12.68%	41.76%
Qwen-7B^{P} *GRPO*	**Qwen-7B^{P}**	13.18%	43.66%

(b) *I-PHYRE* Benchmark.

Policy Model	World Model	Att. 1	Att. 10
Qwen-3B^{I} *SFT*	-	23.33%	45.56%
Qwen-3B^{I} *SFT*	**Qwen-3B$^{I}_{w/o\ 5\ Frames}$**	22.81%	45.04%
Qwen-3B^{I} *SFT*	**Qwen-3B^{I}**	23.21%	47.34%
Qwen-3B^{I} *GRPO*	-	36.67%	48.89%
Qwen-3B^{I} *GRPO*	**Qwen-3B$^{I}_{w/o\ 5\ Frames}$**	36.97%	47.29%
Qwen-3B^{I} *GRPO*	**Qwen-3B^{I}**	37.60%	51.00%
Qwen-7B^{I} *SFT*	-	12.22%	50.00%
Qwen-7B^{I} *SFT*	**Qwen-7B$^{I}_{w/o\ 5\ Frames}$**	11.34%	50.30%
Qwen-7B^{I} *SFT*	**Qwen-7B^{I}**	11.34%	51.50%
Qwen-7B^{I} *GRPO*	-	26.67%	86.67%
Qwen-7B^{I} *GRPO*	**Qwen-7B$^{I}_{w/o\ 5\ Frames}$**	27.12%	87.51%
Qwen-7B^{I} *GRPO*	**Qwen-7B^{I}**	27.82%	89.31%

to leverage ICL to interact with the environment and parse physical intuition. Finally, the outstanding performance of the **GRPO** models further underscores the efficacy of our approach.

Table 5 evaluates the importance of providing the World Model with visual information about the dynamic consequences of an action during its training phase. We compare the full model against a variant where the World Model was trained without the five uniformly sampled post-action frames (detailed in Algorithm 2). Across both *PHYRE* and *I-PHYRE*, and for all policy model configurations, the full ICPRL framework—which leverages a world model trained with post-action visual frames—consistently outperforms the variant employing an ablated world model (data curated by `w/o 5 Frames`). This performance delta provides strong evidence that incorporating explicit visual feedback of an action's dynamic consequences is crucial for training an effective world model, directly translating to more effective guidance for the PUCT search procedure at inference time. Furthermore, it is noteworthy that even the ablated world model generally provides some performance lift over the policy-only baseline, underscoring the fundamental utility of our decoupled two-component architecture for deliberative planning.

5 Conclusion

In this work, we introduced **ICPRL** (In-Context Physical Reinforcement Learning), a framework designed to address the limitations of existing VLMs in interactive reasoning within dynamic physical environments. Our framework integrates an adaptive policy model with a world model that provides explicit physical intuition. Our extensive evaluations on the diverse DeepPHY benchmark demonstrate that ICPRL not only achieves significant performance gains over strong baselines but, crucially, maintains robust generalization and enables the genuine in-context acquisition of an environment's physical dynamics directly from interactive experience.

While ICPRL represents a significant step forward, this work also highlights promising avenues for future research. First, the policy and world models in the current framework are trained independently. Future work could explore synergistic training paradigms where the models are co-trained, allowing data generated by one to inform and enhance the learning process of the other, potentially creating a virtuous cycle of improvement. Second, our analysis revealed that while ICPRL excels at learning overarching physical principles, its performance gains were less pronounced on tasks like *Kinetix*, which demands high-dexterity, component-level manipulation. This distinction suggests that such fine-grained control problems may constitute a distinct class of challenges, representing another promising direction for future investigation.

References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Anthropic. System Card: Claude Opus 4 & Claude Sonnet 4, 05 2025. URL https://www.anthropic.com/model-card. Accessed: 2025-07-01.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems, NeurIPS*, 32, 2019.

Jakob Bauer, Kate Baumli, Feryal M. P. Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor, Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Hannah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez Nieves, Nemanja Rakicevic, Tim Rocktäschel, Yannick Schroecker, Satinder Singh, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl, and Lei M. Zhang. Human-timescale adaptation in an open-ended task space. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML*, 2023. URL https://proceedings.mlr.press/v202/bauer23a.html.

Peng Chen, Pi Bu, Yingyao Wang, Xinyi Wang, Ziming Wang, Jie Guo, Yingxiu Zhao, Qi Zhu, Jun Song, Siran Yang, Jiamang Wang, and Bo Zheng. CombatVLA: An efficient vision-language-action model for combat tasks in 3D action role-playing games. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. LLMPhy: Complex physical reasoning using large language models and world models, 2024. URL https://arxiv.org/abs/2411.08027.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding, 2025. URL https://arxiv.org/abs/2501.16411.

Daniel J. H. Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. Theoretical physics benchmark (TPBench) – a dataset and study of ai reasoning capabilities in theoretical physics, 2025. URL https://arxiv.org/abs/2502.15815.

Gheorghe Comanici and et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre M. Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/985e9a46e10005356bbaf194249f6856-Abstract.html.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Sam Earle and Julian Togelius. Autoverse: An evolvable game language for learning robust embodied agents, 2024. URL https://arxiv.org/abs/2407.04221.

Kevin Frans and Phillip Isola. Powderworld: A platform for understanding generalization via rich task distributions. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023. URL https://openreview.net/forum?id=AWZgXGmsbA.

Samuel Garcin, James Doran, Shangmin Guo, Christopher G. Lucas, and Stefano V. Albrecht. DRED: zero-shot transfer in reinforcement learning via data-regularised environment design. In *Forty-first International Conference on Machine Learning, ICML*, 2024. URL https://openreview.net/forum?id=uku9r6RROl.

Jake Grigsby, Linxi Fan, and Yuke Zhu. Amago: Scalable in-context reinforcement learning for adaptive agents. *arXiv preprint arXiv:2310.09971*, 2023.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy P. Lillicrap. Mastering diverse control tasks through world models. *Nat.*, 640(8059):647–653, 2025.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob N. Foerster, Edward Grefenstette, and Tim Rocktäschel. Replay-guided adversarial environment design. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/0e915db6326b6fb6a3c56546980a8c93-Abstract.html.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021b. URL http://proceedings.mlr.press/v139/jiang21b.html.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://doi.org/10.1016/S0004-3702(98)00023-X.

Nathaniel S Keplinger, Baiting Luo, Iliyas Bektas, Yunuo Zhang, Kyle Hollins Wray, Aron Laszka, Abhishek Dubey, and Ayan Mukhopadhyay. Ns-gym: Open-source simulation environments and benchmarks for non-stationary markov decision processes. *arXiv preprint arXiv:2501.09646*, 2025.

Evan Kiefl. Pooltool: A python package for realistic billiards simulation. *Journal of Open Source Software*, 9 (101):7301, 2024.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M. Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal M. P. Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. In *The Thirteenth International Conference on Learning Representations, ICLR*. OpenReview.net, 2025. URL https://openreview.net/forum?id=CjwERcAU7w.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Stenberg Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *The Eleventh International Conference on Learning Representations, ICLR*. OpenReview.net, 2023. URL https://openreview.net/forum?id=hy0a5MMPUv.

Shiqian Li, Kewen Wu, Chi Zhang, and Yixin Zhu. I-PHYRE: interactive physical reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024.

Jieru Lin, Zhiwei Yu, and Börje F. Karlsson. SWITCH: Benchmarking modeling and handling of tangible interfaces in long-horizon embodied scenarios. *arXiv preprint arXiv:2511.17649*, 2025. URL https://arxiv.org/abs/2511.17649.

Hao Luo, Jiechuan Jiang, and Zongqing Lu. Model-based decentralized policy optimization. *arXiv preprint arXiv:2302.08139*, 2023.

Michael T. Matthews, Michael Beukman, Chris Lu, and Jakob Nicolaus Foerster. Kinetix: Investigating the training of general agents through open-ended physics-based control tasks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, 2025.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML*, 2016. URL http://proceedings.mlr.press/v48/mniha16.html.

OpenAI. OpenAI o3 and o4-mini System Card, April 2025. URL https://openai.com/index/o3-o4-mini-system-card/.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob N. Foerster, Edward Grefenstette, and Tim Rocktäschel. Evolving curricula with regret-based environment design. In *International Conference on Machine Learning, ICML*, 2022. URL https://proceedings.mlr.press/v162/parker-holder22a.html.

Han Qi, Haocheng Yin, Aris Zhu, Yilun Du, and Heng Yang. Strengthening generative robot policies through predictive world modeling. *arXiv preprint arXiv:2502.00622*, 2025.

Qwen Team. Qwen2.5-VL, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Alexander Rutherford, Michael Beukman, Timon Willi, Bruno Lacerda, Nick Hawes, and Jakob N. Foerster. No regrets: Investigating and improving regret approximations for curriculum discovery. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2024.

Gabriel Sarch, Lawrence Jang, Michael J. Tarr, William W. Cohen, Kenneth Marino, and Katerina Fragkiadaki. VLM agents generate their own memories: Distilling experience into embodied programs of thought. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, Nat McAleese, Nathalie Bradley-Schmieg, Nathaniel Wong, Nicolas Porcel, Roberta Raileanu, Steph Hughes-Fitt, Valentin Dalibard, and Wojciech Marian Czarnecki. Open-ended learning leads to generally capable agents, 2021. URL https://arxiv.org/abs/2107.12808.

Fahim Tajwar, Yiding Jiang, Abitha Thankaraj, Sumaita Sadia Rahman, J Zico Kolter, Jeff Schneider, and Ruslan Salakhutdinov. Training a generally curious agent. *arXiv preprint arXiv:2502.17543*, 2025.

Weihao Tan, Wentao Zhang, Xinrun Xu, Haochong Xia, Ziluo Ding, Boyu Li, Bohan Zhou, Junpeng Yue, Jiechuan Jiang, Yewen Li, Ruyi An, Molei Qin, Chuqiao Zong, Longtao Zheng, Yujie Wu, Xiaoqiang Chai, Yifei Bi, Tianbao Xie, Pengjie Gu, Xiyun Li, Ceyao Zhang, Long Tian, Chaojie Wang, Xinrun Wang, Börje F. Karlsson, Bo An, Shuicheng Yan, and Zongqing Lu. Cradle: Empowering foundation agents towards general computer control. In *Forty-second International Conference on Machine Learning (ICML)*, 2025. URL https://proceedings.mlr.press/v267/tan25h.html.

Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao, Linjie Li, Qineng Wang, Hanyang Chen, Chi Wan, Yiping Lu, Zhengyuan Yang, Lijuan Wang, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, Yejin Choi, and Manling Li. VAGEN: Reinforcing world model reasoning for multi-turn VLM agents. In *Advances in Neural Information Processing Systems 38, NeurIPS 2025*, 2025a. URL https://openreview.net/forum?id=xpjWEgf8zi.

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, Mingyu Ding, Lei Bai, Wanli Ouyang, Shixiang Tang, Aoran Wang, and Xinzhu Ma. Physunibench: An undergraduate-level physics reasoning benchmark for multimodal models, 2025b. URL https://arxiv.org/abs/2506.17667.

Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. Newton: Are large language models capable of physical reasoning?, 2023. URL https://arxiv.org/abs/2310.07018.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are scalable world models. *Advances in Neural Information Processing Systems*, 37: 68082–68119, 2024.

Xinrun Xu, Pi Bu, Ye Wang, Börje F. Karlsson, Ziming Wang, Tengtao Song, Qi Zhu, Jun Song, Zhiming Ding, and Bo Zheng. DeepPHY: Benchmarking agentic VLMs on physical reasoning. *arXiv preprint arXiv:2508.05405*, 2025. URL https://arxiv.org/abs/2508.05405.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Honglin Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Haodong Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Haoqi Yuan, Yu Bai, Yuhui Fu, Bohan Zhou, Yicheng Feng, Xinrun Xu, Yi Zhan, Börje F. Karlsson, and Zongqing Lu. Being-0: A Humanoid Robotic Agent with Vision-Language Models and Modular Skills. *arXiv preprint arXiv:2503.12533*, 2025. URL https://arxiv.org/abs/2503.12533.

Yiming Zhang, Yingfan Ma, Yanmei Gu, Zhengkai Yang, Yihong Zhuang, Feng Wang, Zenan Huang, Yuanyuan Wang, Chao Huang, Bowen Song, Cheng Lin, and Junbo Zhao. ABench-Physics: Benchmarking physical reasoning in LLMs via high-difficulty and dynamic physics problems, 2025. URL https://arxiv.org/abs/2507.04766.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62178–62209. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhou24t.html.

Ilya Zisman, Vladislav Kurenkov, Alexander Nikulin, Viacheslav Sinii, and Sergey Kolesnikov. Emergence of in-context reinforcement learning from noise distillation. In *Forty-first International Conference on Machine Learning, ICML*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Y8KsHT1kTV.

Appendix: Model Abbreviations

The abbreviations used throughout the paper and the full details of the corresponding open-source and closed-source models are listed in Table 6.

To clearly distinguish the fine-tuned models developed in our work, we adopt a systematic naming convention, which we illustrate using the Qwen-3B model as an example. This convention applies equally to Qwen-7B-based models. We use GREEN to denote Policy Models and ORANGE for World Models.

Policy Models. The nomenclature for our Policy Models follows the format: $\text{BaseModel}^{\text{Env}_{\text{Data-Type}}\ \textit{Method}}$. Each component in the naming structure is defined as follows:

- **Env:** Indicates the environment(s) used in fine-tuning:
  - **P**: Denotes fine-tuning exclusively on *PHYRE*.
  - **I**: Denotes fine-tuning exclusively on *I-PHYRE*.
  - **P&I**: Denotes a model trained using tasks from both *PHYRE* and *I-PHYRE*.
- **Method:** Specifies the training algorithm used:
  - **SFT**: Supervised Fine-Tuning.

Table 6: List of Model Abbreviations

Model	Reference	Abbreviation
MOCK	-	*MOCK*
Open-Source Models		
Qwen2.5-VL Series		
Qwen2.5-VL-3B-Instruct	-	Qwen-3B
Qwen2.5-VL-7B-Instruct	-	Qwen-7B
Qwen2.5-VL-32B-Instruct	-	Qwen-32B
Qwen2.5-VL-72B-Instruct	-	Qwen-72B
Closed-Source Models		
Claude 4.0 Opus	(Anthropic, 2025)	Claude 4.0 Opus
Gemini-2.5-Pro-06-17	(Comanici & et al., 2025)	Gemini-2.5-Pro
GPT-o3-0416	(OpenAI, 2025)	GPT-o3

  - **GRPO**: Group Relative Policy Optimization (Shao et al., 2024).

- **Data-Type:** An optional subscript specifying the nature of the training data:
  - **(No subscript)**: The model was trained on a dataset of successful solution trajectories collected from expert models (*e.g.,* Gemini-2.5-Pro, GPT-4o). These trajectories were often generated over multiple attempts (typically 5–10) to find a solution. This strategy enables the model to learn from a process of trial and error, aligning with the philosophy of ICRL.
  - $^{\textbf{single}}$: The model was trained exclusively on a dataset of first-attempt successful trajectories, generated via systematic enumeration over the action space.

For example, **Qwen-3B**$^{\textbf{P}_{\textbf{single}}\ SFT}$ denotes a policy model based on Qwen2.5-VL-3B-Instruct, trained via SFT on single-attempt successful trajectories from the *PHYRE* environment.

World Models. The naming for our World Models is more concise. Since all World Models are trained via SFT, the naming directly specifies the base model and the environment. For example, **Qwen-3B**$^{\textbf{P}}$ indicates the model is based on Qwen-3B and trained on *PHYRE*.

World models may include one subscript: "$_{w/o\ 5\ Frames}$". This subscript stands for "without 5 frames" and represents a critical ablation detail. As described in Subsection 4.3, this variant was trained on a curated dataset that *excluded* the five uniformly sampled post-action video frames, serving as a baseline to validate the importance of dynamic visual feedback.

A Mathematical Definitions

We formalize the interactive physical reasoning task as a multi-step Partially Observable Markov Decision Process (POMDP) (Kaelbling et al., 1998) augmented with interaction history.

- **State Space ($\mathcal{S}$):** The underlying physical state of the simulator at time t (denoted as s_t), such as object positions, velocities, and friction coefficients. Note that s_t is not directly accessible to the agent.
- **Observation Space ($\mathcal{O}$):** The visual rendering of the state $o_t \in \mathcal{O}$ at time step t. This may include annotated overlays as detailed in Section 4.1 and DeepPHY (Xu et al., 2025).
- **Action Space ($\mathcal{A}$):** The set of executable commands. While the physics simulator accepts continuous parameters, we discretize $\mathcal{A}$ into text tokens for VLM processing. For example, in Angry Birds, continuous angles $\theta \in [0, 90]$ are discretized into integer bins.

- **Terminal Reward ($\mathcal{R}$):** We utilize a sparse binary reward signal r_{GT} provided by the environment at the end of an episode of length T. Specifically, $r_{GT} = 1$ denotes success, while $r_{GT} = 0$ denotes failure. No intermediate rewards are provided.

- **Trajectory (τ):** A sequence recording the interaction data of a single attempt from start to finish:

$$\tau = (o_0, a_0, o_1, a_1, \ldots, o_T, a_T, r_{GT})$$

- **Interaction History (H):** A collection of trajectories from past failed attempts within the same problem instance: $H_k = (\tau_1, \tau_2, \ldots, \tau_{k-1})$.

- **Policy Mapping:** The policy $\pi_\theta(a_t \mid o_{\leq t}, H)$ maps the current observation sequence (context) and past interaction history to a probability distribution over the next action a_t.

- **Reference Policy (π_{ref}):** A frozen version of the policy model derived from SFT stage. It serves as the anchor for the KL-divergence penalty, preventing the RL-tuned policy π_θ from deviating excessively from the natural language distribution and maintaining output format validity.

B Algorithm Details

Algorithm 1 presents the complete inference procedure with root-node PUCT search, integrating candidate generation, world model scoring, and action selection as described in Section 3.3. Algorithm 2 details the world model dataset curation process, which constructs balanced training data with paired success/failure examples and VLM-generated annotations.

C Scoring Strategy Details

Table 7 summarizes the environment-specific perturbation and scoring methods used within the PUCT search procedure described in Section 3.3.

Table 7: Calculation Methods of Stability and Uncertainty for PUCT Search.

Environment	Methodology	Action Space	Perturbation / Sampling Details
Strategy 1: Action Space Perturbation for Stability Estimation *Stability Score:* $\hat{p}_{\text{stab}}(a) = \frac{1}{J}\sum_{j=1}^{J} \hat{p}_{\text{succ}}(o, a'_j)$			
PHYRE	Action Space Perturbation	(x, y, r) Grid coord. & radius	Perturb (x, y) by ± 1 (4 directions), and r by ± 1 or 0.
I-PHYRE		Sequence of timed events $[(i_k, t_k)]$	Temporal jittering: perturb timestamps t_k with $\pm\Delta t \in \{\pm 0.5\text{s}, \pm 1.0\text{s}\}$ for a subset of events.
Angry Birds		(θ, p) Angle & power	Perturb θ with $\Delta_\theta \in \{-5°, 0°, 5°\}$ and p with $\Delta_p \in \{-0.1, 0, 0.1\}$.
Strategy 2: Confidence-Based Scoring for Highly Sensitive Environments *Uncertainty Score (LCB):* $score(a\|o) = \mu_p - \lambda_{\text{LCB}}\sigma_p$			
Kinetix	Model Confidence (LCB)	Varies (*e.g.,* timings, forces, positions)	$K = 8$ stochastic forward passes of $\mathcal{M}_\phi(o, a)$ using different temperatures sampled from $[0.1, 1.0]$. This yields a set of probabilities $\{p^{(j)}\}_{j=1}^{K}$.
Pooltool			
Cut the Rope			

Algorithm 1 Inference with Root-Node Search.

Require: Observation o; policy π_θ; world model $\mathcal{M}_\phi$; scoring strategy $\texttt{strat} \in \{1, 2\}$; hyperparameters S, B, c_{PUCT}; Strategy 1: J, λ_{PUCT}; Strategy 2: K, λ_{LCB}.
Ensure: Best action a^*.

▷ ***Stage 1: Candidate Generation and Prior***
$A_{\text{samples}} \leftarrow \emptyset$
for $i = 1 \to S$ **do**
  Sample $a \sim \pi_\theta(\cdot \mid o)$ and add to A_{samples}
end for
$A \leftarrow \textsc{Unique}(A_{\text{samples}})$
for each action $a \in A$ **do**
  $P(a) \leftarrow \textsc{Count}(a, A_{\text{samples}})/S$
end for
▷ ***Stage 2 & 3: PUCT Search Guided by WM Scores***
$N(a) \leftarrow 0, Q(a) \leftarrow 0$ for all $a \in A$
$a_{\text{last}} \leftarrow \text{null},\ n_{\text{same}} \leftarrow 0$
for $t = 1 \to B$ **do**
  $N_{\text{tot}} \leftarrow \sum_{b \in A} N(b)$
  $a_t \leftarrow \arg\max_{a \in A} \left[Q(a) + c_{\text{PUCT}} \cdot P(a) \cdot \frac{\sqrt{N_{\text{tot}}}}{1+N(a)} \right]$
  ▷ Early stopping check
  if $a_t = a_{\text{last}}$ **then** $n_{\text{same}} \leftarrow n_{\text{same}} + 1$
  else $n_{\text{same}} \leftarrow 1$
  end if
  if $n_{\text{same}} \geq 3$ **then break**
  end if
  $a_{\text{last}} \leftarrow a_t$
  ▷ Compute Score
  if $\texttt{strat} = 1$ **then** ▷ Strategy 1: Action Space Perturbation (PHYRE, I-PHYRE, Angry Birds)
    $\hat{p}_{\text{succ}} \leftarrow \mathcal{M}_\phi(o, a_t)$
    Sample J perturbed neighbors $\{a'_j\}_{j=1}^{J} \sim \mathcal{B}_\delta(a_t)$
    $\hat{p}_{\text{stab}} \leftarrow \frac{1}{J} \sum_{j=1}^{J} \mathcal{M}_\phi(o, a'_j)$
    $v \leftarrow (1 - \lambda_{\text{PUCT}}) \cdot \hat{p}_{\text{succ}} + \lambda_{\text{PUCT}} \cdot \hat{p}_{\text{stab}}$
  else ▷ Strategy 2: Model Confidence / LCB (Kinetix, Pooltool, Cut the Rope)
    Sample K temperatures $\{\tau_j\}_{j=1}^{K}$ uniformly from $[0.1, 1.0]$
    $p^{(j)} \leftarrow \mathcal{M}_\phi(o, a_t; \tau_j)$ for $j = 1, \ldots, K$
    $\mu_p \leftarrow \frac{1}{K} \sum_{j=1}^{K} p^{(j)}, \quad \sigma_p \leftarrow \sqrt{\frac{1}{K} \sum_{j=1}^{K} (p^{(j)} - \mu_p)^2}$
    $v \leftarrow \mu_p - \lambda_{\text{LCB}} \cdot \sigma_p$
  end if
  ▷ Update search statistics
  $Q(a_t) \leftarrow \frac{Q(a_t) \cdot N(a_t) + v}{N(a_t)+1}$
  $N(a_t) \leftarrow N(a_t) + 1$
end for
▷ ***Stage 4: Final Selection***
$a^* \leftarrow \arg\max_{a \in A} Q(a)$
return a^*

Algorithm 2 World Model Dataset Curation.

Require: Task set $\mathcal{X}$; action space $\mathcal{A}$; simulator $\textsc{Sim}(x, a)$ returns trajectory τ and label $y \in \{0, 1\}$; frames m=5; diversity threshold ε.
Ensure: WM dataset $\mathcal{D}_{\text{WM}}$.

$\mathcal{D}_{\text{WM}} \leftarrow \emptyset$
for each task $x \in \mathcal{X}$ **do**
  Get initial visual states: I_0, I_{ann}
  ▷ **Phase 1: Solution Discovery**
  $S \leftarrow \{a \in \mathcal{A} \mid \textsc{Sim}(x, a).y = 1\}$ ▷ Enumerate or retrieve cached solutions
  $k \leftarrow |S|$
  if $k = 0$ **then continue**
  end if
  ▷ **Phase 2: Balanced Failure Sampling**
  $F \leftarrow \emptyset$
  while $|F| < k$ **do**
    Sample candidate $a \sim \mathcal{A}$
    $d_{\min} \leftarrow \min_{a' \in S \cup F} \text{dist}(a, a')$
    if $d_{\min} \geq \varepsilon$ **and** $\textsc{Sim}(x, a).y = 0$ **then**
      $F \leftarrow F \cup \{a\}$
    end if
  end while
  ▷ **Phase 3: VLM Annotation & Compilation**
  for each action $a \in S \cup F$ **do**
    $(\tau, y) \leftarrow \textsc{Sim}(x, a)$
    Extract m uniformly spaced frames $F_{1:m}$ from τ post-action
    ▷ Generate Grounding, Reasoning, and Label via VLM
    $T_{\text{prompt}} \leftarrow (I_0, I_{\text{ann}}, a, F_{1:m}, y)$
    $T_{\text{output}} \leftarrow \text{VLM}(T_{\text{prompt}})$
    if human verification passes **then**
      $\mathcal{D}_{\text{WM}} \leftarrow \mathcal{D}_{\text{WM}} \cup \{(x, I_0, I_{\text{ann}}, a, F_{1:m}, y, T_{\text{output}})\}$
    end if
  end for
end for
return $\mathcal{D}_{\text{WM}}$

D Implementation Example: Cut the Rope

Figure 3: Examples of static element annotation in the *Cut the Rope* game. Key static props—such as Pins, Active Pins, and Air Cushions—are clearly marked with numerical IDs. This method converts pixel-level visual information into grounded tokens, enabling the Agent to accurately identify and manipulate objects.

To illustrate how continuous gaming environments are adapted for VLM control, we detail the implementation of the *Cut the Rope* task (visualized in Figure 3).

Visual Input: The model receives a screenshot where interactive elements are annotated with numerical IDs. For example, a rope anchored to a pin might be labeled "Pin 1", and a floating bubble containing candy might be labeled "Bubble 2".

Action Space: Instead of continuous gestures, the model outputs structured Python-like function calls within square brackets. The interpreter parses these tokens into game actions. Supported commands include: `[cut_pin(id=3)]` to cut a specific rope, `[pop_bubble(id=2)]` to release candy, or `[sleep(seconds=0.5)]` to handle swing dynamics.