

IPR-1: Interactive Physical Reasoner

Mingyu Zhang^{1,2,*} Lifeng Zhuo^{1,*} Tianxi Tan^{1,2} Guocan Xie¹ Xian Nie¹ Yan Li¹
 Renjie Zhao^{1,**} Zizhu He¹ Ziyu Wang¹ Jiting Cai^{1,3,**} Yong-Lu Li^{1,2,†}

¹Shanghai Jiao Tong University ²Shanghai Innovation Institute ³Carnegie Mellon University

sjtuzmy2003@sjtu.edu.cn yonglu_li@sjtu.edu.cn

Abstract

Humans learn by observing, interacting with environments, and internalizing physics and causality. Here, we aim to ask whether an agent can similarly acquire human-like reasoning from interaction and keep improving with more experience. We study this in a Game-to-Unseen (G2U) setting, curating 1,000+ heterogeneous games with diverse physical and causal mechanisms, and evaluate at three human-like levels: Survival, Curiosity, Utility, from primitive intuition to goal-driven reasoning. Our analysis reveals complementary failures: VLM/VLA agents reason but lack look-ahead in interactive settings, while world models imagine but imitate visual patterns rather than analyze physics and causality. We therefore propose **IPR (Interactive Physical Reasoner)**, using world-model rollouts to score and reinforce a VLM’s policy, and introduce **PhysCode**, a physics-centric action code aligning semantic intent with dynamics to provide a shared action space for prediction and reasoning. Pretrained on 1,000+ games, our IPR performs robustly on three levels, matches GPT-5 overall, and surpasses it on Curiosity. We find that performance improves with more training games and interaction steps, and that the model also zero-shot transfers to unseen games. These results support physics-centric interaction as a path to steadily improving physical reasoning. **Our code will be publicly available.**

1. Introduction

Humans do not learn physics and causality from labels; we earn them through *interaction*. As experience accumulates with age, our prediction sharpens, our reasoning stabilizes, and our abilities scale. This motivates a central question for embodied AI: *what learning paradigm enables human-like reasoning to learn through interactive experience, and to improve steadily with more interaction?*

We assume that, if an agent is exposed to *diverse, inter-*



Figure 1. **Game-to-Unseen (G2U) problem.** Humans accumulate interactive experience and rapidly adapt to new games. Despite different visuals and interfaces, many games share underlying physical/causal mechanisms. We pretrain on 1,000+ visually and physically diverse games to test whether an agent can internalize these shared mechanisms and generalize to *unseen* games.

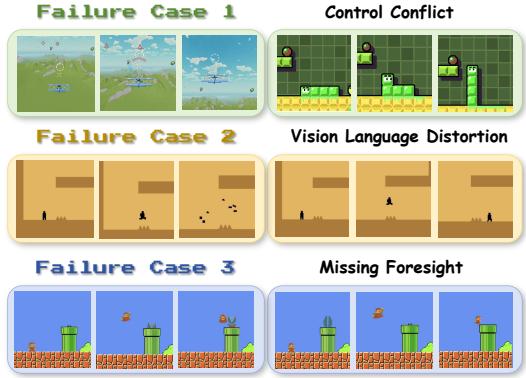
active worlds and trained to capture *shared physical and causal mechanisms*, rather than domain-specific appearance or action interfaces, it would scale its physical reasoning ability reliably and *transfer* to new scenarios. This view aligns with prior reasoning works [6, 22, 57]. Large VLM/VLA models [11, 52, 53], pretrained on vast corpora, exhibit basic reasoning abilities and a certain degree of transfer; yet in interactive scenarios they lack forward prediction *visually* and often fail when *precise anticipation of action consequences* is required (e.g., timing, contact, momentum) (Fig. 2b). World models and other prediction-based methods [3, 10, 21] can *imagine* futures via differentiable latent dynamics and *interactively* optimize trajectories toward goal-aligned representations, but they tend to collapse into target-chasing imitation rather than genuine causal reasoning, failing in long-horizon and sparse-reward tasks. This naturally raises a question: can we *blend* the open-ended reasoning of VLMs with the predictive grounding of world models to support *interactive reasoning* in novel environments and yield competence that improves steadily with experience?

*Equal contribution. †Corresponding author.

**Work conducted at Shanghai Jiao Tong University.



(a) **Three-level evaluation inspired by Maslow’s hierarchy of needs.** We organize tasks into a pyramid of Survival, Curiosity, and Utility. **Survival** measures how long the agent can stay alive by avoiding risks; **Curiosity** measures how broadly it visits novel states; and **Utility** measures how well it achieves downstream goals. The three levels progress from physical intuition to goal-directed reasoning. Our IPR performs robustly across the entire pyramid.



(b) **Motivating failure cases in control semantics, language grounding, and prediction.** (1) *Control conflict*: the same key (e.g., UP) triggers different semantics across games (camera tilt up v.s. character move up), causing console aliasing. (2) *Vision-language distortion*: text-only actions cannot specify precise visual magnitudes (e.g., jump height/speed), leading to systematic amplitude errors. (3) *Missing foresight*: without imagination, the agent cannot anticipate upcoming hazards during interaction (e.g., spikes, moving enemies).

Figure 2. Overview: three-level evaluation pyramid (left) and failure cases of previous VLM-based model (right), motivating our IPR.

In this way, we propose **IPR** (Interactive Physical Reasoner): a paradigm where world model *prediction* reinforces a VLM policy to adapt its physical reasoning in interactive environments (Fig. 3). A natural obstacle arises when naively piping VLM outputs into a predictor: keyboard induces *interface mismatches* across games and language distorts visual details (Fig. 2b). We therefore introduce **PhysCode**, a *physics-centric action code* that fuses action semantics with visual dynamics into a compact discrete representation. Concretely, we follow a Genie [10] style discretization over fused features extracted from video frames, optical flow, and action semantics. Each code is encouraged to align with (i) *domain-agnostic* dynamical primitives (e.g., momentum change) and (ii) *domain-specific* visual affordances. Instead of issuing raw keys or free-form language, the reasoning policy outputs multiple **PhysCode** sequences, which are scored by the world model in the same latent space so that the best candidate would be executed and its imagined rewards are used to reinforce the policy.

To evaluate the paradigm at scale, we curate over 1,000 heterogeneous games spanning visual styles, control interfaces, and physical and causal mechanisms. Games form a low-cost, controllable testbed for physical reasoning: they afford rich interaction, physics closely resembling the real world, and effectively *unlimited* rollouts. We further organize evaluation into three levels inspired by Maslow’s hierarchy of needs [24]: *Survival*, *Curiosity*, *Utility*, covering a spectrum from physical intuition to goal-directed reasoning (Fig. 2a). The result on three levels verifies two failure modes: reasoning-based VLM/VLA lack forward consequence prediction to explore (Curiosity), while prediction-

based world models explore broadly yet fail at goal-driven tasks (Utility), which motivates our design.

Across this suite, our **IPR** remains robust on all three levels, while RL-based and prediction-based baselines often collapse on one or more of them. Trained under the IPR paradigm, an 8B backbone matches GPT-5 overall and even *surpasses* it on curiosity. Moreover, competence scales with the number of training games and interaction steps (Fig. 5) and *zero-shot transfers* to novel environments, highlighting the potential of interactive learning for physical reasoning at scale.

In general, our contributions are: (1) We formulate the **G2U** problem and curate 1,000+ heterogeneous games with a hierarchical evaluation (*Survival/Curiosity/Utility*), diagnosing the strengths and weaknesses of prevalent prediction-based, RL-based, and VLM-based methods. (2) We propose **IPR**: world-model rollouts *score* and *reinforce* VLM in the same action space, enabling interactive experience to steadily build up physical reasoning ability. (3) We introduce **PhysCode**, a physics-centric action code fusing action semantics with visual dynamics, bridging WM prediction and VLM reasoning.

2. Related Works

Action Space Discovery. Research on action spaces spans hand-designed controls, language-based interfaces, and learned latent representations. Early embodied agents operated over environment-specific key bindings, torques, or joystick signals [8, 16, 30, 42], which offer precise control but entangle behavior with platform-specific layouts

and hinder cross-domain transfer. A second line adopts *language*-based action spaces, issuing natural-language commands or tool calls [1, 13, 45, 51, 52]; while language affords semantic generality, it abstracts away timing, force, and perception-action couplings, often leading to imprecise or under-grounded control [39, 41]. A complementary direction learns *latent* action spaces directly from interaction data. Discrete or continuous latent codes—via VQ-VAE [47] or sequence models—have been explored for planning, control, and world models [10, 12, 28, 31, 44]. Recent VLM/VLA systems integrate such latent tokens into large multimodal models [23, 40], but these codes often remain entangled across domains and lack mechanisms to capture shared physical principles versus environment-specific affordances. Our work addresses this gap by learning a *physics-centric* latent action space that captures reusable dynamical patterns across games, instead of binding actions to domain-specific visuals and control layouts.

Agents in Interactive Environments. Research on game-playing agents has largely followed three threads. *RL-based* agents, from DQN and PPO/SAC to large-scale systems like AlphaStar and OpenAI Five [16, 29, 32, 42, 48, 50], learn policies directly from pixels and rewards and achieve strong title-specific performance, but remain sample-inefficient, brittle to interface changes, and struggle with long-horizon credit assignment and cross-game transfer. *Prediction-based* (world-model) agents such as World Models, PlaNet, the Dreamer family, and Genie [10, 15, 17–19] first learn latent dynamics and then plan or optimize in imagination, improving exploration and sparse-reward learning, yet degrade when learned dynamics or action semantics drift from the test environment and typically optimize task or pixel losses rather than reasoning quality. *VLM/VLA-based* agents like Gato, RT-2, Voyager, MineDojo, and recent VLA frameworks [9, 13, 39, 51] cast acting as sequence modeling over images, text, and actions and excel at zero-shot instruction following, but rely heavily on static corpora, heuristic wrappers, and weakly grounded forward prediction. Our IPR paradigm aims to inherit the strengths of these lines by using a physics-centric latent action space where a world model provides imagination-based value estimates and a reasoning VLM policy is reinforced through interactive experience in the *same* latent space.

Benchmarks and Evaluation. Interactive environments have long served as testbeds for learning control, exploration, and generalization: Atari/ALE provided dense step-wise rewards for RL training and evaluation [5, 29], while later platforms such as *Minecraft*, *VizDoom*, and *StarCraft* introduced long-horizon goals, partial observability, and sparse rewards [13, 26, 49, 51]. With the rise of VLM/VLA agents, web-based benchmarks and browser environments

have been proposed to test generalization to novel tasks and interfaces [36, 56]. Following this line, we evaluate agents on a diverse suite of games and adopt simple game-agnostic metrics grouped into three levels—*survival*, *curiosity*, and *utility*—to provide their performance from physical intuition to reasoning and their scaling with experience.

3. Preliminaries

3.1. Problem Setting

We consider a family of interactive environments $\{\mathcal{E}_m\}_{m=1}^M$, each formalized as a POMDP:

$$\mathcal{M}_m = (\mathcal{S}, \mathcal{A}, T_m, R_m, \mathcal{O}, \gamma; \varphi_m), \quad (1)$$

where φ_m are latent *physics parameters* (e.g., gravity g , friction μ , mass M). At time t , the environment emits an image $x_t \sim \mathcal{O}(\cdot | s_t)$, which we encode as $z_t = \phi_{\text{enc}}(x_t)$; the agent executes $a_t \in \mathcal{A}$ and transitions according to

$$s_{t+1} \sim T_m(s_{t+1} | s_t, a_t; \varphi_m), \quad r_t = R_m(s_t, a_t), \quad (2)$$

where physics resides in T_m , and causality in R_m .

Control may use one of several interfaces $A \in \{\text{KEYBOARD}, \text{LANGUAGE}, \text{LATENT}\}$; a goal-conditioned VLM selects actions in the chosen space via

$$a_t^{(A)} \sim \pi_\omega^{(A)}(\cdot | z_t, \text{prompt}_t), \quad a_t \equiv a_t^{(A)} \in \mathcal{A}. \quad (3)$$

A feature-level world model f_θ then rolls out imagined futures under selected action sequences in the same action space A . Given a horizon $H \in \mathbb{N}$, initialize $\hat{z}_t := z_t$ and choose an action sequence $\{a_{t+k}^{(A)}\}_{k=0}^{H-1}$. The rollout is defined by

$$\hat{z}_{t+k+1} = f_\theta(\hat{z}_{t+k}, a_{t+k}^{(A)}), \quad k = 0, 1, \dots, H-1, \quad (4)$$

where k indexes the step inside the imagined trajectory from time t to $t + H$.

3.2. PhysCode: Physics-centric Action Code

Motivated by the issues of raw-key semantic aliasing and the distortion of fine-grained visual dynamics when expressed in language, we propose *PhysCode*, a discrete latent action representation built on a VQ codebook $\mathcal{C} = \{v_k\}_{k=1}^K$. At step t , an action is a short code sequence $a_t^{\text{LAT}} = \langle c_{t,1:L} \rangle$ with embedding obtained by looking up and pooling $\{v_{c_{t,\ell}}\}$.

Each code is conditioned on three cues: (i) *domain-specific* visual appearance via DINOv3 [46] features $\phi_{\text{img}}(x_t)$, (ii) *domain-agnostic* motion via optical flow [14] $\phi_{\text{flow}}(\text{Flow}(x_t, x_{t+1}))$, and (iii) lightweight semantic hints extracted by a T5 encoder [38], with $\phi_{\text{sem}}(y_t) = \text{Enc}_{\text{T5}}(y_t)$. Since natural language alone cannot express fine-grained dynamics (e.g., impulse magnitude, frictional slip), we rely on flow and visual features to carry these details while keeping semantics as guidance. By design, the resulting codes

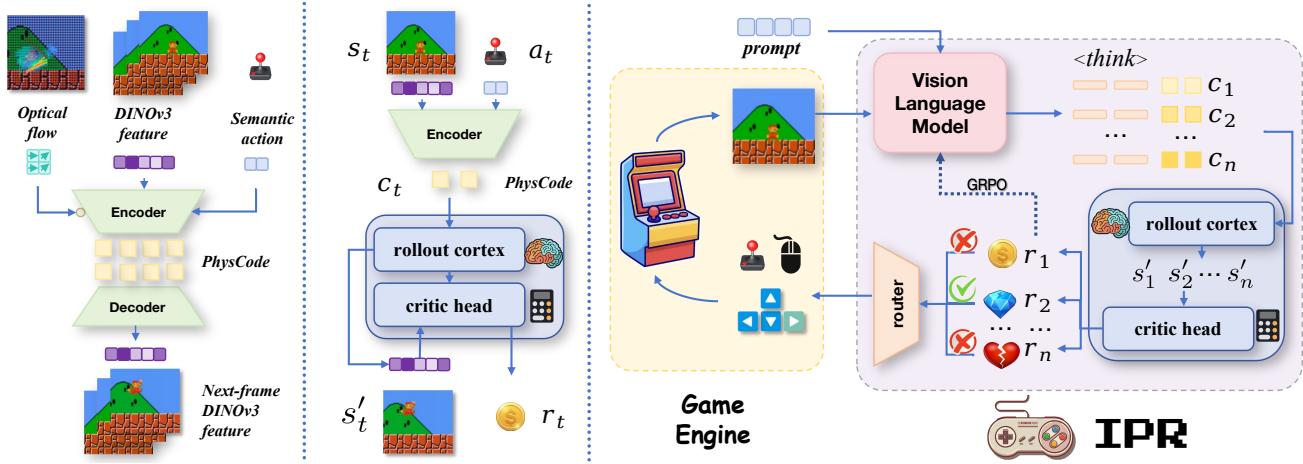


Figure 3. **IPR training pipeline.** **Stage 1: PhysCode pre-training.** Video clips with optical flow and action semantics are fed to a VQ-based latent action model to learn discrete codes (*PhysCode*) that represent dynamics. **Stage 2: Latent-conditioned world model.** Given current features and *PhysCode* sequences, a world model is trained to predict future features and rewards under latent actions. **Stage 3: Prediction-reinforced reasoning.** A VLM reasons over the scene and generates candidate *PhysCode* sequences. The world model rolls them out in imagination, and the predicted rewards/values are used to select the best actions and to optimize the VLM policy.

capture *physics-relevant* intervention primitives that *share* across environments with similar underlying physics and *separate* when physics differ, enabling consistent reuse under matched physics and discrimination under shifted dynamics.

4. Method

In this section, we introduce three components of **IPR** (Fig. 3): (1) learning a *physics-centric action code vocabulary* across diverse physical principles and causal mechanisms; (2) training a *latent-conditioned world model* that predicts future features and rewards under sequences of latent actions; and (3) *reinforcing VLM with world model rollout prediction* in the interactive environment, using aligned latent action code. In inference, the VLM proposes candidate latent actions, queries the world model for short-horizon imagination and value estimates to score them, and executes the highest-scoring action.

Inducing the Latent Action Vocabulary. Using the cues in Sec. 3.2 (DINOv3 appearance $f_t, f_{t+\Delta}$, optical flow u_t , and lightweight semantics e_t), a small gated fusion module forms a fused representation h_t . A spatio-temporal encoder E_ψ maps h_t to a continuous code z_t , which is vector-quantized to an index $a_t \in \{1, \dots, K\}$ with codebook $\mathcal{C} = \{c_k\}_{k=1}^K$, and a decoder D_ψ predicts the future feature $\hat{f}_{t+\Delta}$ from (f_t, c_{a_t}) . We train with a standard VQ-VAE objective

$$\begin{aligned} \mathcal{L}_{\text{LA}} = & \|\hat{f}_{t+\Delta} - f_{t+\Delta}\|_2^2 \\ & + \beta \|\text{sg}[z_t] - c_{a_t}\|_2^2 + \gamma \|z_t - \text{sg}[c_{a_t}]\|_2^2, \end{aligned} \quad (5)$$

augmented with modality dropout on flow and a mild gate-sparsity regularizer to avoid over-reliance on optional cues. Since optical flow is only available during pretraining, it acts as privileged information that helps shape a physics-centric codebook, while dropout and gate sparsity distill this structure into an encoder that, at test time, relies only on appearance and semantic cues. At inference, we disable the flow gate and reuse the same encoder to obtain z_t and its quantized index a_t from appearance+semantics only. The resulting discrete vocabulary yields temporally predictive tokens that cluster under matched physics and separate under different dynamics, providing a shared interface for VLM reasoning and world-model prediction.

Training the Latent-Level World Model with a Critic. With the latent action vocabulary fixed, we train a feature-level world model to predict future features conditioned on latent actions, replacing raw controls with their *PhysCode* indices. For triples $(f_t, a_t, f_{t+\Delta})$, we embed a_t to e_{a_t} and compute

$$(\hat{f}_{t+\Delta}, V_\theta(f_t, a_t)) = P_\theta(f_t, e_{a_t}). \quad (6)$$

We predict in the *latent space*, since features compress appearance variance and rendering noise, making dynamics more shareable across games. Concretely, we first train the world model with a feature-prediction loss $\mathcal{L}_{\text{pred}} = \|\hat{f}_{t+\Delta} - f_{t+\Delta}\|_1$, and then learn a critic head with a Q-learning-style objective $\mathcal{L}_{\text{value}} = \ell_Q(V_\theta(f_t, a_t), y_t)$, where y_t is a target value computed from rollout returns via standard TD backups.

Prediction-Reinforced Interactive Reasoning. We strengthen interactive reasoning with prediction: a world model imagines rollouts, and a VLM plans in the same latent action space. We adopt Qwen3-VL-8B [54] as the backbone and extend its tokenizer with *PhysCode* tokens so the VLM can directly emit discrete latent actions while preserving its language ability.

We first align perception and action by supervised training on (f_t, c_t) pairs, where f_t is the DINOv3 feature of the current frame and c_t the latent action learned in Stage 1. Given the current context and goal g , the VLM samples B candidate *PhysCode* sequences $\{\mathbf{a}^{(b)}\}_{b=1}^B$, and the world model runs short-horizon imagined rollouts to assign each a predicted return, from which we compute advantages $A^{(b)}$. We then update the policy with GRPO [43]:

$$\mathcal{L}_{\text{GRPO}} = \frac{1}{B} \sum_{b=1}^B A^{(b)} \log \pi_\phi(\mathbf{a}^{(b)} | f_t, g) - \beta \text{KL}(\pi_\phi \| \pi_0), \quad (7)$$

In inference, the VLM proposes latent action candidates, the world model scores and prunes them via short-horizon rollouts, and a router T_{env} maps the selected *PhysCode* to environment controls. Through repeated interaction under this prediction-in-the-loop scheme, the experience collected from imagined and executed trajectories reinforces the VLM, improving its physical reasoning in interactive environments.

5. Experiments

In this section, we aim to answer three questions: (1) Why is *PhysCode* necessary compared with raw keyboard inputs or language instructions? (2) How would world model prediction reinforce VLM reasoning? (3) Would IPR show scaling potential to transfer to unseen games?

5.1. Setup: Datasets, Tasks, and Metrics

Sources. We curate a multi-source benchmark covering **863** open-source retro titles (via `stable-retro` [35]), **134** lightweight HTML/Canvas games, and **3** commercial games. This breadth exposes agents to heterogeneous visuals, action interfaces, and underlying physics/causal mechanisms, encouraging models to capture shared physical-causal regularities rather than overfit to domain-specific biases.

Diversity axes. We characterize each environment along seven axes to enable structured generalization analysis: (1) *Game category*, with emphasis on physical interaction (e.g., platformer, shooter, sports); (2) *Control interface*, such as GameBoy-style discrete keys, keyboard-mouse combinations, and high-dimensional hybrids; (3) *Visual complexity*, ranging from low-resolution pixel art to high-fidelity

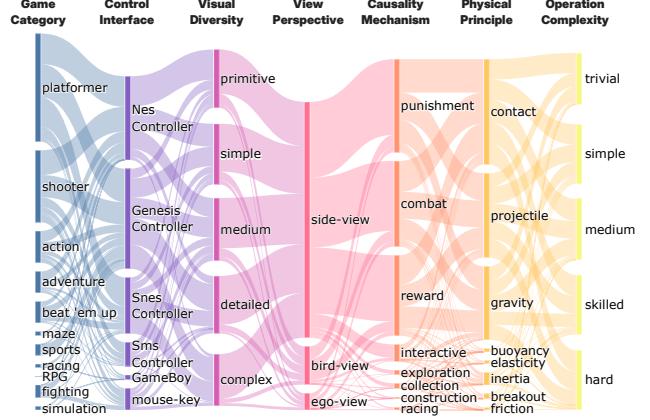


Figure 4. **Game data distribution.** Our dataset spans over 1,000 games categorized by *game category*, *control interface*, *operation* and *visual complexity*, *physical* and *causal mechanisms*. This wide coverage enables agents to experience diverse domains and learn transferable physical and causal understanding.

3D; (4) *View perspective*, e.g. ego-centric, top-down, and side views; (5) *Causal mechanism*, e.g. damage/health dynamics, collection, punishment; (6) *Physical principle*, e.g. gravity, contact, and inertia; (7) *Operational difficulty*, approximated by the entropy and frequency of human control actions, reflecting how precisely and how often players must operate to succeed; Fig. 4 summarizes the distributions over sources, game types, and these axes; detailed per-environment statistics are provided in the *supplementary*.

Data collection and preprocessing. Across the **1,000**-game corpus, we record human play at 60 FPS for 4 minutes per title and obtain per-game annotations covering *physical principles*, *causal mechanisms*, *action semantics*, and *game instructions*. We perform a series of preprocessing, including normalizing time intervals, removing non-interactive segments, rebalancing extended idle/no-op periods, etc. More details are in the *supplementary*.

Hierarchical level design. Inspired by Maslow’s hierarchy of needs [24], we treat gameplay as a three-level progression: *Survival* \rightarrow *Curiosity* \rightarrow *Utility* (Fig. 2a), from intuition to reasoning.

📍 **Survival.** The objective is to remain alive as long as possible, ignoring the original goal and avoiding risks. We report *survival time* normalized per game, $H = \mathbb{E}[T]/T_{\text{typ}}$, where T is episode length (steps) and T_{typ} is a per-game reference horizon (e.g., median survival under a random policy).

🧭 **Curiosity.** The goal is to visit *novel states* like a baby to uncover regularities in the environment’s dynamics and causal mechanisms. Following Magnipy [27], we embed frames with a pretrained CLIP visual encoder [37], compute

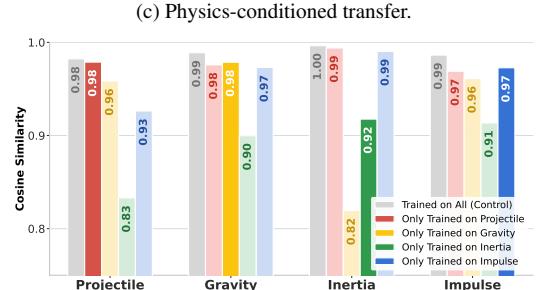
Table 1. **PhysCode validation.** **Left:** Joint training across heterogeneous-physics games reveals cross-game conflicts for keyboard/mouse; language partially alleviates this via semantics, while *PhysCode* separates actions by dynamics, reducing interface aliasing and showing minimal degradation under physics shifts. **Middle:** Leave- n -out transfer: training on all but 10 titles and evaluating zero-shot on the held-out set, *PhysCode* transfers more reliably than keyboard or language interfaces. **Right:** Physics-conditioned transfer: zero-shot performance is relatively higher when target environments *match* the training set’s physical mechanisms, indicating that *PhysCode* captures reusable physical principles rather than game-specific bindings.

(a) Confusion test for joint training.			
Latent-Predict	Cosine \uparrow	MSE \downarrow	L1 \downarrow
Ad-hoc	0.9939	0.0121	0.0495
Keyboard	0.9894	0.0211	0.0772
Language	0.9892	0.0216	0.0758
<i>PhysCode</i>	0.9919	0.0204	0.0737

(b) Leave- n -out transfer.			
Latent-Predict	Cosine \uparrow	MSE \downarrow	L1 \downarrow
Pre-trained	0.9856	0.0230	0.0846
Keyboard	0.9784	0.0430	0.1153
Language	0.9790	0.0418	0.1132
<i>PhysCode</i>	0.9798	0.0403	0.1212

(c) Physics-conditioned transfer.			
	FID \downarrow	SSIM \uparrow	PSNR \uparrow
Ad-hoc	87.83	0.7062	23.86
Keyboard	110.0	0.6110	20.82
Language	82.51	0.6960	23.52
<i>PhysCode</i>	80.35	0.7240	23.82

Latent-Predict	Cosine \uparrow	MSE \downarrow	L1 \downarrow
Pre-trained	127.3	0.7438	22.11
Keyboard	315.0	0.3340	12.46
Language	320.2	0.1670	9.389
<i>PhysCode</i>	297.0	0.3533	13.04



the trajectory’s multi-scale *metric-space magnitude* curve $M(\tau)$, and define the exploration score as the area under this curve: $E = \text{AUC}(M(\tau))$, where larger E indicates broader state-space coverage.

Utility. Utility measures how well an agent *realizes Bentham’s utility of life* [7]: devoting itself to goal completion with higher reward and shorter time. We evaluate downstream goals according to the game types (completion, score, checkpoint time) and report the *human-normalized score* (*HNS*) [4] per game:

$$\text{HNS} = \frac{m - m_{\text{rnd}}}{m_{\text{hum}} - m_{\text{rnd}}} , \quad (8)$$

where m is the agent metric, m_{rnd} the random baseline, and m_{hum} human performance.

5.2. Why is PhysCode Necessary

We first investigate whether **PhysCode** is necessary compared with raw keyboard/mouse inputs and natural-language instructions. First, we assess robustness under mixed-game joint training with heterogeneous physics (Tab. 1a), examining which action space best performs in diverse physical mechanisms and different console/game interfaces. Second, we test transfer (Tab. 1b, Tab. 1c): a *shared* PhysCode learned on source games improves zero-shot performance in unseen environments with *matched* physics, demonstrating genuine physics grounding rather than interface memorization.

First, we examine how different action spaces behave when trained jointly across a mixture of games with heterogeneous physics (Tab. 1a). In this regime, raw keyboard/mouse inputs exhibit cross-game conflicts (the same key triggers different behaviors across environments). Language interfaces partially alleviate this via explicit semantics. *PhysCode* separates actions by dynamics, reducing

interface aliasing and showing minimal degradation under physics shifts.

Next, we ask whether sharing the latent space supports transfer. In a leave- n -out protocol (Tab. 1b), we train on all but 10 games and evaluate zero-shot on the held-out titles. We find that *PhysCode* transfers more reliably than keyboard or language instructions.

Moreover, we condition transfer on the physics of the environment. We group games by their dominant physical mechanism, train under one principle (*e.g.*, gravity), and evaluate zero-shot on held-out games with matching or different mechanisms. When targets *match* the training physics, zero-shot performance is *typically* higher (Tab. 1c), with notable exceptions such as *inertia*, which may already be covered by projectile/impulse. This suggests that *PhysCode* captures reusable physical mechanisms rather than game-specific bindings, even though our coarse physics taxonomy does not perfectly align with the agent’s internal abstractions.

5.3. Playing in Diverse Physical Worlds of Games

We evaluate IPR against prevalent baselines on 200 games, chosen to match the full dataset’s distribution of types, action spaces, and physics/causality. The baselines include:

- **RL.** We utilize Multitask PPO [55] (*policy-based*) and shared-parameter DQN [34] (*value-based*) as standard reinforcement learning approaches.
- **VLM.** We employ a range of vision-language models, including closed-source models such as GPT-4o and GPT-5 [33], as well as open-source models like Qwen3-VL-30B-A3B [54].
- **World Model.** We compare three different world models: DreamerV3 [19] (*latent-based*), V-JEPA2 [3] (*pretrained latent-based prediction*), and Genie [10] (*pixel-based prediction*) (we follow GenieRedux implementation [25]).
- **IL.** We apply imitation learning (IL) models, including

Table 2. **Comprehensive comparison across 🍔, 🗺, and 🎉.** Obj: training objective; Mean: normalized score; Avg. Rank: normalized average rank ($1/(n-1)$, higher is better); Ratio@Top-3(%): percentage of games where the method is in the top-3.

Methods	Survival			Curiosity			Utility		
	Mean	Avg. Rank	Ratio@Top-3(%)	Mean	Avg. Rank	Ratio@Top-3(%)	Mean	Avg. Rank	Ratio@Top-3(%)
Control Group									
Random	0.541	0.450	3.5	5.254	0.363	5.8	0.000	0.559	1.1
Human	0.764	0.760	41.9	5.428	0.480	11.7	1.000	0.874	76.1
Imitation Learning (IL) Group									
ACT-BC	0.541	0.474	5.8	6.896	0.476	13.0	0.092	0.487	5.7
Qwen3-VL-8B-BC	0.578	0.578	7.0	6.945	0.502	11.8	0.129	0.560	3.4
Reinforcement Learning (RL) Group									
PPO@survival	0.571	0.581	16.3	5.453	0.415	2.3	0.259	0.636	14.8
PPO@curiosity	0.562	0.566	11.6	5.344	0.373	3.5	0.285	0.659	19.3
PPO@utility	0.554	0.609	17.4	5.419	0.397	2.3	0.268	0.637	18.2
DQN@survival	0.538	0.493	15.1	7.058	0.588	10.6	0.147	0.518	8.0
DQN@curiosity	0.562	0.531	18.6	6.841	0.578	12.8	0.159	0.521	8.4
DQN@utility	0.556	0.515	18.6	6.450	0.563	9.3	0.156	0.536	9.1
World Model Group									
DreamerV3@survival	0.521	0.495	8.1	7.737	0.526	14.0	0.114	0.446	6.8
DreamerV3@curiosity	0.538	0.466	8.1	7.843	0.608	20.9	0.138	0.449	11.5
DreamerV3@utility	0.538	0.481	9.3	7.336	0.530	17.6	0.139	0.446	8.7
V-JEPA2@survival	0.531	0.394	5.8	7.815	0.469	14.0	0.065	0.285	4.0
V-JEPA2@curiosity	0.516	0.343	0.0	8.463	0.610	22.1	0.046	0.246	4.5
V-JEPA2@utility	0.526	0.369	1.2	7.777	0.412	15.1	0.051	0.293	4.5
GenieRedux@survival	0.539	0.435	3.5	7.937	0.573	17.4	0.090	0.334	4.0
GenieRedux@curiosity	0.534	0.421	3.5	8.390	0.675	20.9	0.084	0.293	3.4
GenieRedux@utility	0.540	0.445	4.7	8.067	0.573	17.4	0.085	0.349	4.5
Multimodal Large Language Model (MLLM) Group									
GPT-4o@survival	0.575	0.605	14.0	4.954	0.390	2.2	0.156	0.543	7.0
GPT-4o@curiosity	0.521	0.429	7.0	5.800	0.518	4.7	0.124	0.474	4.5
GPT-4o@utility	0.522	0.474	8.1	5.676	0.465	2.3	0.162	0.549	8.0
GPT-5@survival	0.619	0.630	25.6	5.181	0.358	2.3	0.199	0.593	18.2
GPT-5@curiosity	0.529	0.488	5.8	5.621	0.448	5.8	0.148	0.515	5.7
GPT-5@utility	0.536	0.460	5.8	5.329	0.427	2.3	0.206	0.606	13.6
Qwen3-VL-A30B@survival	0.544	0.514	5.8	5.690	0.522	5.8	0.149	0.539	6.8
Qwen3-VL-A30B@curiosity	0.518	0.450	7.0	7.113	0.600	11.7	0.133	0.500	4.5
Qwen3-VL-A30B@utility	0.545	0.504	9.3	6.214	0.443	4.7	0.146	0.552	6.8
Interactive Physical Reasoner									
IPR (8B) (ranking)	0.589 (2/26)	0.517 (8/26)	10.4 (9/26)	7.874 (5/26)	0.584 (6/26)	15.6 (7/26)	0.178 (7/26)	0.604 (5/26)	8.9 (9/26)

Key Takeaways across 🍔 Survival, 🗺 Curiosity, and 🎉 Utility

- **Prediction-based Methods (WM).** Strong at 🗺, but weaker at 🍔 and 🎉. Trained on broad exploratory trajectories, latent rollouts broaden coverage and reveal dynamics, but tend to imitate visually-alike futures rather than reliably pursue goals. So prediction is useful as a look-ahead prior for risk and candidate actions.
- **RL-based Methods (PPO, DQN).** Strong at 🍔 and 🎉 when rewards are well-shaped, but weaker on 🗺 and tasks without explicit goals. Reward gradients enable effective credit assignment under the right signal, yet sparsity and partial observability induce instability and interface overfitting—so RL works best as an optimization method.
- **Experience-based Methods (Behavior Cloning).** Strong at human-like 🍔, but weaker on 🗺 and 🎉. Deliberately imitate human trajectories and thus excel at low-risk survival, but struggle once tasks require precise control or exploration, and their performance depends strongly on the coverage and quality of the demonstrations.
- **Reasoning-based Pretrained VLMs.** Strong at goal-conditioned 🍔 and 🎉; weaker on 🗺. They excel at instruction-driven reasoning but cannot predict consequences in the visual state space, so they work best as high-level reasoners that need auxiliary prediction modules for outcome-aware decisions.
- **Interactive Physical Reasoning (Ours).** Robust across 🍔, 🗺, and 🎉. We combine the strengths of all three paradigms: VLMs provide goal-driven causal reasoning, the world model supplies rollout prediction, and RL optimizes decisions using imagined rewards, yielding consistently strong performance across all three levels.

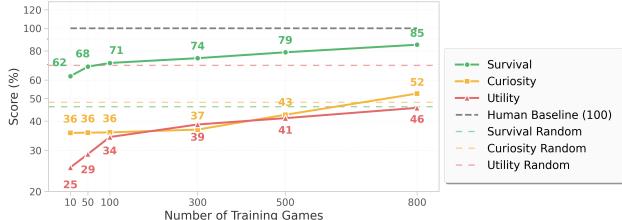


Figure 5. **G2U zero-shot scaling on 50 held-out games.** As the number of training games N increases, zero-shot performance on , , and improves steadily on the unseen set \mathcal{T}_U .

ACT [58] (*end-to-end model*) and Qwen3-VL-8B [54] (*VLM-based model*).

We assess every model on the three hierarchical objectives, instantiating level-specific training or prompting. Further implementation details are provided in the *supplementary*. The key results are reported in Tab. 2. Takeaways are below the table.

5.4. Zero-shot Transferring to Unseen Games

To validate our *Games-to-Unseen (G2U)* setting, we construct a held-out target set \mathcal{T}_U of 50 games that are *never* used for training. From the remaining pool, we form stratified training subsets $\{\mathcal{S}_N\}$ of increasing size N , balanced by physics and causal mechanisms to control for domain bias. For each N , we train our *IPR* paradigm end-to-end on \mathcal{S}_N and *directly* evaluate zero-shot on \mathcal{T}_U without any adaptation or reward re-scaling.

Across all three objectives, performance increases steadily with N , with the steepest early gains on , followed by sustained improvements on and as more diverse interactions are observed. This suggests that training in *physically and causally related* environments helps *IPR* move beyond domain-specific quirks (visual style, control interface) and focus on *shared physical and causal patterns* (e.g., gravity, contact, momentum). In other words, as interactive experience accumulates, *IPR* behaves more *human-like*: it carries over physical priors and causal expectations rather than memorizing domain appearance or controls, demonstrating potential to further scale in richer interactive domains.

5.5. Ablations and Analysis

Does prediction help VLM reasoning? Table 3 compares variants on the same Qwen3-VL-8B backbone. Starting from the pretrained VLM, naive BC barely changes survival ($0.62 \rightarrow 0.63$) but *hurts* curiosity and utility, suggesting that low-quality demonstrations can overwrite useful priors instead of improving control. PPO on top of the VLM achieves the best survival (1.00) and higher utility (1.23), but further suppresses curiosity, and combining PPO with BC degrades all three metrics, indicating

Table 3. Ablation study results for IPR components of World Model prediction and GRPO.

Method	Survival	Curiosity	Utility
VLM (pretrained)	0.62	2.14	0.89
VLM + BC	0.63	1.88	0.87
VLM + PPO	1.00	1.79	1.23
VLM + BC + PPO	0.57	1.86	0.77
IPR	0.76	2.77	1.34

that RL alone tends to overfit short-term rewards under biased data. In contrast, our IPR, which augments the VLM with world-model prediction and GRPO updates, attains the highest curiosity (2.77) while keeping strong survival and utility, showing that prediction-based reinforcement is key to strengthening long-horizon physical reasoning rather than simply pushing for higher immediate scores.

6. Discussion

We study an interactive physical reasoning paradigm in which a general-purpose VLM reasons in language, acts through a physics-centric latent interface (*PhysCode*), and is reinforced by imagined rewards from a world model, asking whether such agents can internalize physical and causal regularities from heterogeneous games and show clear scaling as experience grows. From this perspective, latent-action world models (e.g. Genie, UniVLA [10, 11]) learn discrete action abstractions and latent dynamics for controllable rollouts; imagination-based control methods (e.g. Dreamer, V-JEPA2-AC [2, 20]) optimize policies inside learned world models over device-level actions; and large-scale VLM-based game agents (e.g. Game-TARS [53]) scale vision–language–action models with massive human demonstrations and auxiliary multimodal tasks. Yet, from a physics-centric perspective, these approaches do not explicitly organize actions by shared physical mechanisms across hundreds of games or align VLM’s reasoning ability with prediction competence in a common latent space. IPR combines their advantages to study how physical knowledge and transfer emerge under the unified Survival-Curiosity-Utility evaluation, though it is still limited to game environments and short-horizon imagination, leaving real-world transfer and longer-horizon reasoning to future work.

7. Conclusion

In this work, we introduced *IPR*, a paradigm that *reinforces physical reasoning with prediction* by coupling a physics-centric latent action space (*PhysCode*) with prediction-guided VLM optimization, so that physical and causal regularities are distilled directly from interactive consequences rather than static corpora. On a curated suite of 1,000+ heterogeneous games with *Survival/Curiosity/Utility* evaluation, IPR yields robust gains over VLM-based, prediction-

based, and RL-based baselines, and shows strong zero-shot transfer to unseen games (*survive the 1001st night*). These results suggest that a general-purpose VLM, when grounded in a physics-organized latent interface and trained with imagined rewards, can indeed *learn* and *scale* its physical reasoning ability purely through interaction, providing a step toward interactive agents that acquire reusable physical and causal knowledge.

References

- [1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances, 2022. 3
- [2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. 8
- [3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 1, 6
- [4] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013. 6
- [5] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of artificial intelligence research*, 47:253–279, 2013. 3
- [6] Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007. 1
- [7] Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation*. T. Payne and Son, 1789. 6
- [8] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016. 2
- [9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. 3
- [10] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 3, 6, 8
- [11] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions, 2025. 1, 8
- [12] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion, 2024. 3
- [13] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge, 2022. 3
- [14] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks, 2015. 3
- [15] David Ha and Jürgen Schmidhuber. World models. 2018. 3
- [16] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. 2, 3
- [17] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels, 2019. 3
- [18] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination, 2020.
- [19] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 3, 6
- [20] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. 8
- [21] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models, 2025. 1
- [22] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey, 2023. 1

- [23] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puha Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world, 2024. 3
- [24] William Huitt. Maslow's hierarchy of needs. *Educational psychology interactive*, 23, 2007. 2, 5
- [25] Naser Kazemi, Nedko Savov, Danda Paudel, and Luc Van Gool. Learning generative interactive environments by trained agent exploration, 2024. 6
- [26] Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *2016 IEEE conference on computational intelligence and games (CIG)*, pages 1–8. IEEE, 2016. 3
- [27] Katharina Limbeck, Rayna Andreeva, Rik Sarkar, and Bastian Rieck. Metric space magnitude for evaluating the diversity of latent representations. *Advances in Neural Information Processing Systems*, 37:123911–123953, 2024. 5
- [28] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data, 2021. 3
- [29] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 3
- [30] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015. 2
- [31] Masashi Okada and Tadahiro Taniguchi. Dreaming: Model-based reinforcement learning by latent imagination without reconstruction, 2021. 3
- [32] OpenAI. Dota 2 with large scale deep reinforcement learning, 2019. 3
- [33] OpenAI. Gpt-4o system card, 2024. 6
- [34] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016. 6
- [35] Mathieu Polquin. Stable retro: A maintained fork of openai's gym-retro. <https://github.com/Farama-Foundation/stable-retro>, 2025. 5
- [36] Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Wenyi Zhao, Yu Yang, Xinyue Yang, Jiadai Sun, Shuntian Yao, et al. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. *arXiv preprint arXiv:2411.02337*, 2024. 3
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5
- [38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. 3
- [39] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 3
- [40] Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025. 3
- [41] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 3
- [42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 3
- [43] 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. 5
- [44] Archit Sharma, Michael Ahn, Sergey Levine, Vikash Kumar, Karol Hausman, and Shixiang Gu. Emergent real-world robotic skills via unsupervised off-policy reinforcement learning, 2020. 3
- [45] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks, 2020. 3
- [46] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darct, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprise, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 3
- [47] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. 3
- [48] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning, 2015. 3
- [49] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017. 3
- [50] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Joseph Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun

- Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575:350 – 354, 2019. 3
- [51] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. 3
 - [52] Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jining Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 3
 - [53] Zihao Wang, Xujing Li, Yining Ye, Junjie Fang, Haoming Wang, Longxiang Liu, Shihao Liang, Junting Lu, Zhiyong Wu, Jiazhan Feng, Wanjun Zhong, Zili Li, Yu Wang, Yu Miao, Bo Zhou, Yuanfan Li, Hao Wang, Zhongkai Zhao, Faming Wu, Zhengxuan Jiang, Weihao Tan, Heyuan Yao, Shi Yan, Xiangyang Li, Yitao Liang, Yujia Qin, and Guang Shi. Game-tars: Pretrained foundation models for scalable generalist multimodal game agents, 2025. 1, 8
 - [54] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 5, 6, 8
 - [55] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020. 6
 - [56] Zhecheng Yuan, Sizhe Yang, Pu Hua, Can Chang, Kaizhe Hu, and Huazhe Xu. Rl-vigen: A reinforcement learning benchmark for visual generalization. *Advances in Neural Information Processing Systems*, 36:6720–6747, 2023. 3
 - [57] Mingyu Zhang, Jiting Cai, Mingyu Liu, Yue Xu, Cewu Lu, and Yong-Lu Li. Take a step back: Rethinking the two stages in visual reasoning, 2024. 1
 - [58] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.