IPR-1: Interactive Physical Reasoner

Mingyu Zhang[1,*] Lifeng Zhuo[1,*,‡] Tianxi Tan[1] Guocan Xie[1] Xian Nie[1] Yan Li[1]
Renjie Zhao[1,**] Zizhu He[1] Ziyu Wang[1] Jiting Cai[1,2,**] Yong-Lu Li[1,†]
RHOS
[1]Shanghai Jiao Tong University [2]Carnegie Mellon University
sjtuzmy2003@sjtu.edu.cn yonglu_li@sjtu.edu.cn

Abstract

Humans learn by observing, interacting with environments, and internalizing physics and causality. We explore whether agents can similarly acquire human-like reasoning through interaction and experience. To study this, we introduce a Game-to-Unseen *(G2U) benchmark of 1,000+ heterogeneous games that exhibit significant visual domain gaps. Existing approaches, including VLMs and world models, struggle with underlying physics and causality due to overfitting on visual details. 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, IPR performs robustly on levels from primitive intuition to goal-driven reasoning, and even surpasses GPT-5 overall. 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. Further demos and details can be found at* https://mybearyzhang.github.io/ipr-1.

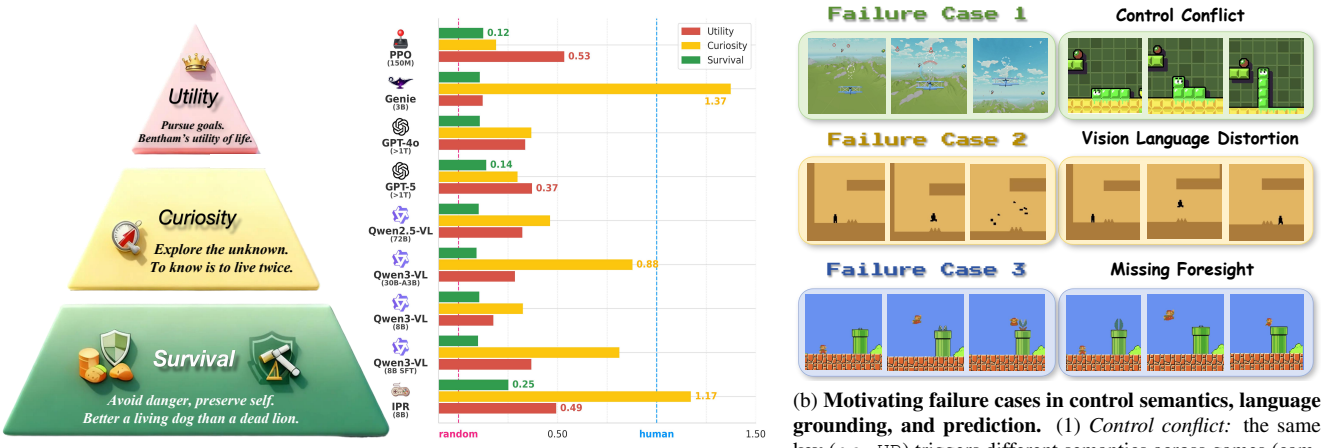
1. Introduction

Humans do not learn physics and causality from labels; we learn 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?*

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.

We assume that, if an agent is exposed to *diverse, interactive worlds* and trained to distill *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 resonates with prior reasoning works [6, 12, 28, 61]. Pretrained VLMs [23, 56] rely on static pattern-matching and lack interactive grounding as open-loop planning; SFT strengthens text-based declarative reasoning but not the predictive grounding required for interactive physical tasks. Behavior-cloned VLAs [11, 55] are limited by demonstration quality and prone to failure under variations. Model-based approaches [33, 57], such as control theory and model-based RL, ensure stability but struggle with complex environment modeling, whereas model-free RL [37, 59] demands massive samples and often overfits to task-specific shortcuts. Recent world models [3, 10, 22] scale latent dynamics but frequently collapse into short-horizon imitation or surface correlations, lacking robust causal reasoning and suffering

*Equal contribution. †Corresponding author.
** Conducted during an internship at Shanghai Jiao Tong University.
‡ Lifeng Zhuo is with the School of Computer Science, SJTU, and the Zhiyuan College, SJTU.



(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-driven 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*.

from compounding errors.

Collectively, these limitations highlight a fundamental gap: while existing paradigms exhibit partial success, they tend to overfit to superficial visual details rather than capturing the underlying physical and causal mechanisms. Approximating these invariant dynamics for robust transfer requires leveraging diverse domains to disentangle core mechanisms from appearance. While RL excels at interactive optimization, it relies on task-entangled signals; similarly, Generative World Models often over-model sensory space, and VLMs lack the predictive grounding for physical consistency. This motivates a "blended" perspective: integrating these strengths via a **Latent World Model** [3] backbone. This scalable reasoner should: **(i)** model only **essential latent dynamics** for consequence anticipation, bypassing pixel reconstruction; **(ii)** utilize **VLM-enriched semantic policies**; and **(iii)** reinforce policies via predictive physical feedback. By predicting abstract representations, the system filters task-irrelevant perceptual noise to capture the "essence" of physical laws over world "appearance."

In this way, we propose *IPR* (**I**nteractive **P**hysical **R**easoner), a paradigm where world model *prediction* reinforces a VLM policy to adapt its physical reasoning in interactive environments (Fig. 3). To evaluate this paradigm at scale, we curate over 1,000 heterogeneous games spanning diverse visual styles, control interfaces, physics configurations, and causal structures. Games provide an ideal testbed for physical reasoning: they offer rich interaction, realistic physics, and effectively *unlimited* rollouts at low cost. Crucially, their heterogeneous visual appearances introduce substantial domain gaps that typically break traditional agents trained environment-by-environment. For IPR, however, these diverse worlds share the same underlying physical and causal principles, allowing it to learn a representation focused enough to transfer across radically different domains.

We organize evaluation into three levels inspired by Maslow's hierarchy of needs [25]: *Survival*, *Curiosity*, and *Utility*, spanning physical intuition to goal-directed reasoning (Fig. 2a). Results reveal two failure modes: reasoning-based VLMs/VLAs lack forward prediction for exploration (*Curiosity*), while prediction-based world models fail at goal-driven tasks (*Utility*). Across the full suite, our *IPR* remains robust across all levels, whereas RL and prediction-based baselines often collapse on one or more. With an 8B backbone, IPR even *surpasses* GPT-5 overall. Furthermore, performance scales with training games and interaction steps (Fig. 5) and *zero-shot transfers* to novel environments, highlighting the potential of large-scale interactive physical reasoning. Future work will extend this paradigm to real-world robotic tasks.

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 interfaces, and learned latent representations. Early agents used environment-specific key bindings or torques [8, 17, 32, 45], offering precision but hindering cross-domain transfer due to platform-specific layouts. Alternatively, *language*-based actions [1, 14, 48, 54, 55] provide semantic generality but often suffer from imprecise or under-grounded control [42, 44] by abstracting away low-level dynamics. A complementary approach learns *latent* action spaces from interaction data. Discrete or continuous codes via VQ-VAE [50] or sequence models have been used for planning and world models [10, 13, 30, 34, 47]. While recent VLM/VLA systems [24, 43] integrate these tokens, they remain entangled across domains and fail to distinguish shared physical principles from environment-specific affordances. Our work addresses this by learning a *physics-centric* latent space that captures reusable dynamical patterns across games, rather than binding actions to domain-specific visuals.

Agents in interactive environments. Research on game-playing agents follows three primary threads. *RL-based* agents, ranging from DQN to large-scale systems like AlphaStar and OpenAI Five [17, 31, 35, 45, 51, 53], learn policies from pixels and rewards to achieve expert performance; however, they remain sample-inefficient, brittle to interface changes, and struggle with long-horizon credit assignment. *Prediction-based* (world-model) agents, such as PlaNet, Dreamer, and Genie [10, 16, 18–20], learn latent dynamics to plan in imagination, improving exploration in sparse-reward settings; yet, they degrade under dynamics drift and often prioritize pixel-level reconstruction over high-level reasoning. *VLM/VLA-based* agents like Gato, RT-2, and Voyager [9, 14, 42, 54] treat acting as sequence modeling over multimodal tokens to excel at zero-shot instruction following, but rely on static corpora and lack grounded physical prediction (Fig. 2b). Our IPR paradigm integrates these strengths 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.

Benchmarks and evaluation. Interactive environments have long served as testbeds for control, exploration, and generalization. While *Atari/ALE* provided dense rewards for early RL training [5, 31], later platforms such as *Minecraft*, *VizDoom*, and *StarCraft* introduced long-horizon goals, partial observability, and sparse rewards [14, 27, 52, 54]. To evaluate VLM/VLA agents, web-based benchmarks have recently been proposed to test generalization across novel tasks and interfaces [39, 60]. Following this line, we evaluate agents on a diverse suite of games using three-level metric levels: *survival*, *curiosity*, and *utility*. These metrics measure performance from physical intuition to high-level 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:

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

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

$$s_{t+1} \sim T_m\big(s_{t+1} \mid s_t, a_t; \varphi_m\big), \qquad r_t = R_m(s_t, a_t), \tag{2}$$

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

$$a_t^{(A)} \sim \pi_\omega^{(A)}\big(\cdot \mid z_t, \text{prompt}_t\big), \qquad a_t \equiv a_t^{(A)} \in \mathcal{A}. \tag{3}$$

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

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

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 [49] features $\phi_{\text{img}}(x_t)$, (ii) *domain-agnostic* motion via optical flow [15] $\phi_{\text{flow}}(\text{Flow}(x_t, x_{t+1}))$, and (iii) lightweight semantic hints extracted by a T5 encoder [41], with $\phi_{\text{sem}}(y_t) = \text{Enc}_{\text{T5}}(y_t)$. As 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 capture *physics-relevant* intervention primitives that *share* across domains with similar underlying physics and *separate* when physics differ, enabling consistent reuse under matched physics and discrimination under shifted dynamics.

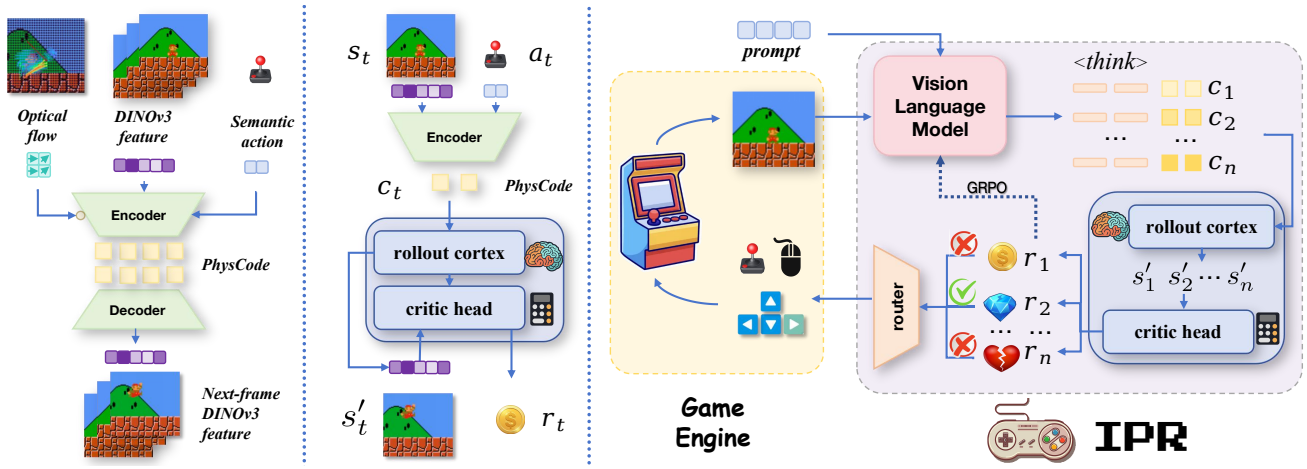


Figure 3. **IPR training pipeline. Stage 1: PhysCode pre-training.** Video clips with optical flow and action semantics are fed to a VQ-based latent action model to learn discrete codes (*PhysCode*) that represent dynamics. **Stage 2: Latent-conditioned world model.** Given current features and PhysCode 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.

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

$$\mathcal{L}_{\text{LA}} = \big\|\hat{f}_{t+\Delta} - f_{t+\Delta}\big\|_2^2 + \beta\big\|\text{sg}[z_t] - c_{a_t}\big\|_2^2 + \gamma\big\|z_t - \text{sg}[c_{a_t}]\big\|_2^2, \tag{5}$$

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

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

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 $\mathcal{L}_{\text{pred}} = \big\|\hat{f}_{t+\Delta} - f_{t+\Delta}\big\|_1$, and then learn a critic head with a Q-learning–style objective $\mathcal{L}_{\text{value}} = \ell_{\text{Q}}\big(V_\theta(f_t, a_t),\, y_t\big)$, 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 [58] 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 [46]:

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

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` [38]), **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 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*.

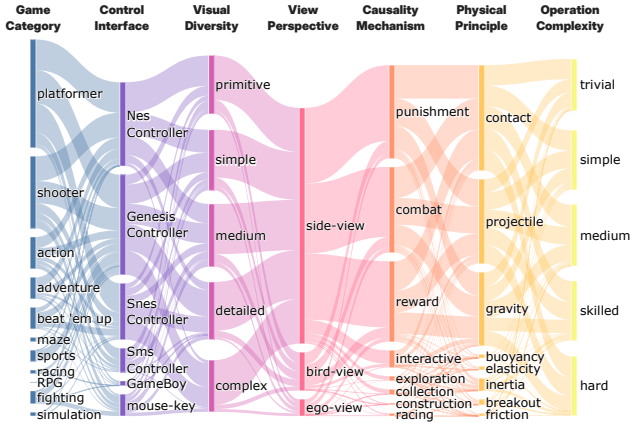


Figure 4. **Game data distribution.** Our dataset spans over 1,000 games categorized by *game category*, *control interface*, *operation and visual complexity*, *physical and causal mechanisms*. This wide coverage enables agents to experience diverse domains and learn transferable physical and causal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

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 [25], we treat gameplay as a three-level progression: *Survival* → *Curiosity* → *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 [29], we embed frames with a pretrained CLIP visual encoder [40], compute the trajectory's multi-scale *metric-space magnitude* curve $M(\tau)$, and define the exploration score as the area under this curve: $E = \mathrm{AUC}(M(\tau))$, 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 com-

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.

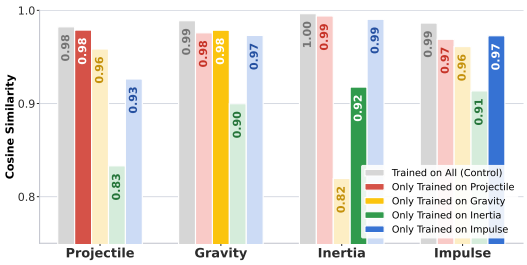
(a) Confusion test for joint training.

Latent-Predict	Cosine ↑	MSE ↓	L1 ↓
Ad-hoc	0.9939	0.0121	0.0495
Keyboard	0.9894	0.0211	0.0772
Language	0.9892	0.0216	0.0758
PhysCode	**0.9919**	**0.0204**	**0.0737**
Pixel-Predict	**FID ↓**	**SSIM ↑**	**PSNR ↑**
Ad-hoc	87.83	0.7062	23.86
Keyboard	110.9	0.6110	20.82
Language	82.51	0.6960	23.52
PhysCode	**80.35**	**0.7240**	**23.82**

(b) Leave-n-out transfer.

Latent-Predict	Cosine ↑	MSE ↓	L1 ↓
Pre-trained	0.9856	0.0230	0.0846
Keyboard	0.9784	0.0430	0.1153
Language	0.9790	0.0418	**0.1132**
PhysCode	**0.9798**	**0.0403**	0.1212
Pixel-Predict	**FID ↓**	**SSIM ↑**	**PSNR ↑**
Pre-trained	127.3	0.7438	22.11
Keyboard	315.0	0.3340	12.46
Language	320.2	0.1670	9.389
PhysCode	**297.0**	**0.3533**	**13.04**

(c) Physics-conditioned transfer.



pletion 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:

$$\text{HNS} = \frac{m - m_{\text{rnd}}}{m_{\text{hum}} - m_{\text{rnd}}}, \tag{8}$$

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

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 [59] *(policy-based)* and shared-parameter DQN [37] *(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 [36], as well as open-source models like Qwen3-VL-30B-A3B [58].
- **World Model.** We compare three different world models: DreamerV3 [20] *(latent-based)*, V-JEPA2 [3] *(pretrained latent-based prediction)*, and Genie [10] *(pixel-based prediction)* (we follow GenieRedux implementation [26]).
- **IL.** We apply imitation learning (IL) models, including ACT [62] *(end-to-end model)* and Qwen3-VL-8B [58] *(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 summarized below the table.

Table 2. **Comprehensive comparison across 🍖, 🧭, and 👑.** "@" denotes the optimization objective. Scores are normalized individually for each game, scaled between random (0) and human (1) benchmarks. **Mean** is the average of these normalized scores, indicating overall competence. **Avg. Rank** is the average relative rank among 30 methods across all games (lower is better). **Ratio@Top-3(%)** is the proportion of games where the method ranks within the top-3. Our IPR demonstrates robust performance across all metrics.

Methods	🍖Survival			🧭Curiosity			👑Utility			Overall
	(Overall) Mean ↑	(Robustness) Avg. Rank ↓	(Competitiveness) Ratio@Top-3(%) ↑	(Overall) Mean ↑	(Robustness) Avg. Rank ↓	(Competitiveness) Ratio@Top-3(%) ↑	(Overall) Mean ↑	(Robustness) Avg. Rank ↓	(Competitiveness) Ratio@Top-3(%) ↑	Avg. Rank ↓
Control Group										
Random	0.000	16.2	6.7	0.000	18.1	3.0	0.000	12.3	12.8	26.9
Human	1.000	5.7	46.3	1.000	7.9	14.0	1.000	2.9	61.6	2.8
Imitation Learning (IL) Group										
ACT-BC	0.088	14.3	17.1	0.793	15.1	12.8	0.255	12.0	13.4	16.6
Qwen3-VL-8B-BC	0.099	12.9	14.0	0.812	12.8	9.1	0.368	9.6	12.8	13.3
Reinforcement Learning (RL) Group										
PPO@survival	0.125	14.0	14.0	0.233	16.5	3.7	0.588	7.3	**30.5**	12.0
PPO@curiosity	0.114	14.9	11.6	0.190	17.3	2.4	**0.609**	**6.9**	29.3	14.8
PPO@utility	0.120	15.0	12.2	0.220	16.8	3.0	*0.534*	8.0	25.6	14.7
DQN@survival	0.121	14.4	15.9	0.856	14.4	8.5	0.497	10.8	15.2	12.2
DQN@curiosity	*0.131*	13.2	*18.3*	0.772	13.4	7.9	0.424	10.9	15.9	10.6
DQN@utility	0.125	13.7	16.5	0.620	14.2	4.9	0.445	10.8	17.1	11.4
World Model Group										
DreamerV3@survival	0.102	15.8	15.2	1.120	*12.5*	16.5	0.298	11.3	16.5	13.1
DreamerV3@curiosity	0.108	14.5	17.7	1.161	13.1	14.0	0.235	10.0	20.1	10.7
DreamerV3@utility	0.097	14.9	17.7	0.964	15.4	11.0	0.139	11.4	18.3	15.4
V-JEPA2@survival	0.102	17.4	4.9	1.150	15.6	17.7	0.191	13.9	16.5	18.3
V-JEPA2@curiosity	0.100	17.8	2.4	**1.402**	15.6	16.5	0.146	14.0	11.6	20.8
V-JEPA2@utility	0.102	17.5	1.8	1.136	14.5	**22.6**	0.152	14.1	11.6	20.2
GenieRedux@survival	0.108	13.7	15.9	1.198	12.5	11.0	0.128	12.7	12.8	14.2
GenieRedux@curiosity	0.104	14.3	14.0	1.374	12.5	9.8	0.100	12.8	12.8	16.1
GenieRedux@utility	0.110	13.7	16.5	*1.248*	**12.4**	14.6	0.122	13.5	14.6	12.4
Multimodal Large Language Model (MLLM) Group										
GPT-4o@survival	0.108	*12.6*	13.4	0.039	17.2	0.6	0.302	9.2	19.5	16.4
GPT-4o@curiosity	0.079	16.8	11.6	0.368	15.3	5.5	0.186	10.6	17.7	19.4
GPT-4o@utility	0.087	15.8	10.4	0.319	14.6	3.7	0.337	10.0	17.1	18.8
GPT-5@survival	0.140	10.5	24.4	0.127	18.3	1.8	0.263	8.0	23.8	13.3
GPT-5@curiosity	0.093	15.3	12.2	0.298	16.4	7.3	0.333	9.8	16.5	17.9
GPT-5@utility	0.108	15.2	11.0	0.185	16.5	0.6	0.371	*7.8*	*26.2*	16.8
Qwen3-VL-30B-A3B@survival	0.091	14.3	11.0	0.325	23.0	0.0	0.289	12.0	14.0	22.7
Qwen3-VL-30B-A3B@curiosity	0.086	15.8	11.6	0.878	20.5	2.4	0.155	11.7	15.2	22.4
Qwen3-VL-30B-A3B@utility	0.108	13.5	12.2	0.528	21.3	4.9	0.285	11.6	14.6	17.6
Interactive Physical Reasoner										
Qwen3-VL-8B w/o IPR	0.105	13.7	14.0	0.325	15.0	4.3	0.176	11.6	12.8	18.2
Qwen3-VL-8B w/ IPR	**0.252**	**2.6**	**72.0**	1.173	13.1	13.4	0.493	8.5	22.0	**4.9**
(IPR ranking w/o control group)	(**1**/28)	(**1**/28)	(**1**/28)	(**5**/28)	(**6**/28)	(**7**/28)	(**5**/28)	(**6**/28)	(**6**/28)	(**4.9**/28)

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 Reasoner (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.
- **Summary.** Prediction-based world models understand dynamics but cannot reliably plan toward long-horizon goals, while reasoning-based VLMs can plan semantically but lack grounded prediction of physical outcomes. IPR combines them by using WM rollouts as physical priors and VLM reasoning to select and pursue feasible futures, surpassing GPT-5 with an 8B backbone.

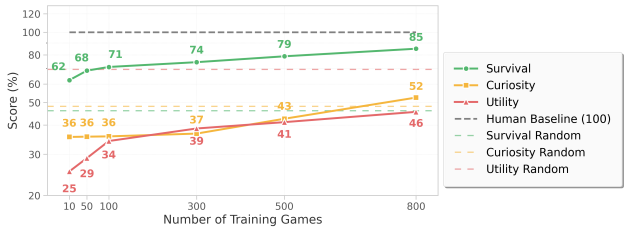


Figure 5. **G2U zero-shot scaling on 50 held-out games.** As the number of training games N increases, zero-shot performance on 🍖, 🧭, and 👑 improves steadily on the unseen set $\mathcal{T}_{\text{U}}$.

5.4. Zero-shot Transfer to Unseen Games

To validate our *Games-to-Unseen (G2U)* setting, we construct a held-out target set $\mathcal{T}_{\text{U}}$ 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 $\mathcal{T}_{\text{U}}$ without any adaptation or reward re-scaling.

Across 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* acts 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→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 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.

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 + GRPO	0.95	1.78	1.22
VLM + BC + PPO	0.57	1.86	0.77
VLM + BC + GRPO	0.55	1.84	0.79
IPR	0.76	**2.77**	**1.34**

6. Discussion

We study an interactive physical reasoner 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, 21]) optimize policies inside learned world models over device-level actions; and large-scale VLM-based game agents (*e.g.* Game-TARS [56]) 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 its 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.

8. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant No. U25A20442, 62306175, Shanghai Municipal Science and Technology Major Project No. 2025SHZDZX025G14.

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, 2, 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. 5

[8] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016. 3

[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, 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] Zixuan Chen, Jiaxin Li, Junxuan Liang, Liming Tan, Yejie Guo, Cewu Lu, and Yong-Lu Li. M3-vos: Multi-phase, multi-transition, and multi-scenery video object segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29193–29202, 2025. 1

[13] 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

[14] 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

[15] 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

[16] David Ha and Jürgen Schmidhuber. World models. 2018. 3

[17] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. 3

[18] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels, 2019. 3

[19] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination, 2020.

[20] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 3, 6

[21] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. 8

[22] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models, 2025. 1

[23] Sihao Hu, Tiansheng Huang, Gaowen Liu, Ramana Rao Kompella, Fatih Ilhan, Selim Furkan Tekin, Yichang Xu, Zachary Yahn, and Ling Liu. A survey on large language model-based game agents. *arXiv preprint arXiv:2404.02039*, 2024. 1

[24] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world, 2024. 3

[25] William Huitt. Maslow's hierarchy of needs. *Educational psychology interactive*, 23, 2007. 2, 5

[26] Naser Kazemi, Nedko Savov, Danda Paudel, and Luc Van Gool. Learning generative interactive environments by trained agent exploration, 2024. 6

[27] 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

[28] Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, Yuan Yao, Siqi Liu, and Cewu Lu. Beyond object recognition: A new benchmark towards object concept learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20029–20040, 2023. 1

[29] 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

[30] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data, 2021. 3

[31] 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

[32] 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. 3

[33] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023. 1

[34] Masashi Okada and Tadahiro Taniguchi. Dreaming: Model-based reinforcement learning by latent imagination without reconstruction, 2021. 3

[35] OpenAI. Dota 2 with large scale deep reinforcement learning, 2019. 3

[36] OpenAI. Gpt-4o system card, 2024. 6

[37] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016. 1, 6

[38] Mathieu Poliquin. Stable retro: A maintained fork of openai's gym-retro. https://github.com/Farama-Foundation/stable-retro, 2025. 5

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[46] 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

[47] 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

[48] 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

[49] 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 Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 3

[50] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. 3

[51] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning, 2015. 3

[52] 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

[53] 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

[54] 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

[55] Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing 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

[56] 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

[57] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019. 1

[58] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 4, 6

[59] 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. 1, 6

[60] 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

[61] 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

[62] 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. 6