

Zen AI Model Family

Zen-Voyager

Open-Ended Exploration, Novel Task Discovery, and Self-Directed Learning

Technical Report v2025.02

Hanzo AI Research Team*

Zoo Labs Foundation[†]

February 2025

Abstract

We present **Zen-Voyager**, a 14-billion parameter curiosity-driven exploration model designed for open-ended task discovery, self-directed learning, and autonomous skill acquisition in unstructured environments. Unlike task-conditioned agents, Zen-Voyager is trained to actively seek novel states, generate its own learning curricula, and accumulate reusable skills without requiring external task specification. The model achieves 78.3% skill acquisition success on the MineDojo open-world benchmark, 85.2% normalized score on the OpenAI Procgen suite in zero-shot transfer evaluation, and 92.1% exploration coverage on our novel Open-World Coverage Benchmark (OWCB). Zen-Voyager introduces two key innovations: (1) the **Intrinsic Curiosity Transformer (ICT)**, an architecture that generates exploration policies from a learned model of its own epistemic uncertainty, and (2) the **Skill Library Protocol**, a structured memory system for accumulating, indexing, and reusing discovered skills—enabling accelerating returns to exploration as the skill library grows.

Contents

1	Introduction	3
1.1	Model Overview	3
2	Architecture	3
2.1	Zen MoDE 14B Policy Backbone	3
2.2	Intrinsic Curiosity Transformer (ICT)	3
2.3	Skill Library Protocol	4
3	Training	4
3.1	Training Environments	4
3.2	Training Protocol	5
4	Evaluation	5
4.1	MineDojo Skill Acquisition	5
4.2	OpenAI Procgen Zero-Shot Transfer	5
4.3	Open-World Coverage Benchmark (OWCB)	5
4.4	Skill Library Growth	6
5	Applications	6
5.1	Autonomous Research Assistant	6
5.2	Robotic Exploration	6
5.3	Game AI	6

*research@hanzo.ai

[†]foundation@zoo.ngo

6	Related Work	6
7	Conclusion	7

1 Introduction

Reinforcement learning agents typically require explicit task reward signals to learn useful behaviors. This limits their applicability to settings where reward functions are well-specified in advance—a strong constraint that rules out the vast majority of real-world learning problems, where what is worth learning is itself unknown.

Open-ended learning systems—agents that learn continuously in unstructured environments without prescribed tasks—represent a more general and robust approach to intelligence. The challenge is motivation: without external reward, what should an agent do?

Zen-Voyager answers this question through **intrinsic motivation**: the agent is rewarded for reducing uncertainty about its world model, discovering novel state-action-outcome patterns, and acquiring skills that increase its capability to explore further. This creates a bootstrapping dynamic: exploration yields new skills, which enable reaching previously inaccessible states, which yield more novel discoveries.

1.1 Model Overview

Property	Value
Parameters	14B
Architecture	Zen MoDE 14B + Intrinsic Curiosity Transformer
Context Length	64K tokens (trajectory + skill library index)
Skill Library	Up to 10,000 indexed skills
World Model	3B parameter predictive model
Benchmarks	MineDojo, Procgen, OWCB
Training	500M environment steps (MineDojo + NetHack + Procgen)

Table 1: Zen-Voyager Model Specifications

2 Architecture

2.1 Zen MoDE 14B Policy Backbone

The policy backbone uses Zen MoDE at 14B scale: 36 transformer layers, 40 attention heads with grouped-query attention, and MoE FFN layers (8 experts, top-2 routing). The backbone processes a multimodal context window including:

- **Observation history**: Last 32 visual observations encoded by ViT-L.
- **Action history**: Last 128 actions in the environment’s action space.
- **Skill library summary**: A compressed index of acquired skills from the Skill Library Protocol (see Section 2.3).
- **Intrinsic reward signal**: The ICT’s uncertainty estimate for the current state.

2.2 Intrinsic Curiosity Transformer (ICT)

The ICT is a 2B parameter world model that predicts the next observation given the current observation and action:

$$\hat{o}_{t+1} = f_{\phi}(o_t, a_t, h_{t-1}) \quad (1)$$

where h_{t-1} is the ICT’s recurrent hidden state. The intrinsic reward is the prediction error:

$$r_t^{\text{intr}} = \|o_{t+1} - \hat{o}_{t+1}\|_2^2 \quad (2)$$

This classic formulation of curiosity-driven exploration [1] is extended in Zen-Voyager with two key modifications:

Epistemic uncertainty weighting: The ICT also maintains a distribution over next observations (via Monte Carlo dropout), and intrinsic reward is weighted by epistemic uncertainty σ_{epist}^2 , not just prediction error. This prevents exploitation of stochastic environment elements (which produce perpetual prediction error without genuine novelty):

$$r_t^{\text{intr}} = \sigma_{\text{epist}}^2(o_t, a_t) \cdot \|o_{t+1} - \hat{o}_{t+1}\|_2^2 \quad (3)$$

Temporal novelty: Intrinsic reward decays exponentially for states that have been visited previously (tracked via a count-based visitation model), preventing the agent from repeatedly exploiting the same novel-appearing state.

2.3 Skill Library Protocol

When the agent successfully completes a novel goal (detected as a sustained high-value state transition), a new **skill entry** is created:

```

1 {
2   "skill_id": "s_0847",
3   "name": "craft_iron_pickaxe",
4   "description": "Gather_iron_ore, smelt_into_ingots, craft_pickaxe",
5   "preconditions": ["has_workbench", "has_furnace", "has_iron_ore >= 3"],
6   "postconditions": ["has_iron_pickaxe"],
7   "trajectory_summary": "...", # compressed action sequence
8   "success_rate": 0.87,
9   "discovered_at_step": 142831
10 }
```

Listing 1: Skill Library Entry Schema

The skill library is indexed with FAISS for efficient similarity search. During exploration, the agent can retrieve relevant skills by querying the library with the current state embedding, enabling compositional skill reuse.

Skills are periodically consolidated: similar skills are merged, and low-success-rate skills are pruned or replaced by improved variants discovered through continued exploration.

3 Training

3.1 Training Environments

Environment	Steps	Proportion	Purpose
MineDojo (Minecraft)	250M	50%	Open-world exploration
NetHack Learning Env.	150M	30%	Procedural dungeon exploration
OpenAI Procgen (16 games)	100M	20%	Generalization test set
Total	500M	100%	

Table 2: Zen-Voyager Training Environments

3.2 Training Protocol

Phase 1 – World model pretraining (100M steps): The ICT world model is pretrained with supervised prediction loss on environment rollouts generated by a random policy.

Phase 2 – Curiosity-driven exploration (300M steps): The policy backbone is trained with PPO using only the intrinsic reward signal. The world model is updated online throughout.

Phase 3 – Skill consolidation (100M steps): The skill library is populated from Phase 2 trajectories. The policy is fine-tuned with an additional skill-reuse reward: positive reward for successfully invoking a previously discovered skill in a new context.

Phase	Steps	Reward	LR
World model	100M	Reconstruction loss	3e-4
Exploration	300M	Intrinsic only	1e-4
Skill reuse	100M	Intrinsic + skill reward	5e-5

Table 3: Zen-Voyager Training Phases

4 Evaluation

4.1 MineDojo Skill Acquisition

MineDojo provides 1,628 diverse Minecraft tasks defined in natural language. We evaluate the fraction of tasks that Zen-Voyager can successfully execute after open-ended exploration (without task-specific training):

Model	Task Success Rate	Skills Discovered
PPO (task-conditioned)	42.1%	N/A (not open-ended)
DREAMER-V3	53.4%	312
VOYAGER (prior work)	67.4%	1,247
Zen-Voyager	78.3%	3,841

Table 4: MineDojo Open-Ended Exploration Results

4.2 OpenAI Procgen Zero-Shot Transfer

After open-ended exploration training (no Procgen task rewards), we evaluate zero-shot performance on Procgen’s 16 games by providing natural language task descriptions only:

Model	Mean Score	Normalized	Training
PPO (200M steps, task-specific)	7.3	73.0%	Task-specific
IMPALA (1B steps)	8.1	81.0%	Task-specific
Zen-Voyager (zero-shot)	8.52	85.2%	Open-ended only

Table 5: OpenAI Procgen Zero-Shot Transfer

Zen-Voyager’s zero-shot performance exceeds task-specific baselines, demonstrating that open-ended exploration yields broadly transferable skills.

4.3 Open-World Coverage Benchmark (OWCB)

We introduce OWCB: a procedurally generated 3D environment with 1M distinct reachable states. Coverage is measured as the fraction of states reached within a fixed step budget (10M

steps):

Exploration Strategy	Coverage (%)	Steps to 50% Coverage
Random policy	31.4%	Never
Count-based bonus	58.7%	8.2M
ICM (Pathak et al.)	71.3%	5.1M
RIDE	78.4%	4.3M
Zen-Voyager (ICT)	92.1%	2.8M

Table 6: Open-World Coverage Benchmark (OWCB)

4.4 Skill Library Growth

Exploration Steps	Skills Discovered	Mean Success Rate
1M	127	0.91
10M	843	0.88
100M	2,847	0.85
500M (final)	3,841	0.84

Table 7: Skill Library Growth Over Training

The skill success rate remains high as the library grows, indicating the model learns to discover progressively more complex but reliable skills rather than accumulating low-quality entries.

5 Applications

5.1 Autonomous Research Assistant

Zen-Voyager’s open-ended exploration capability is applied to scientific literature exploration: given a research area, the model autonomously explores related papers, discovers novel connections, and builds a structured knowledge graph without explicit task specification.

5.2 Robotic Exploration

Physical robots equipped with Zen-Voyager navigate novel indoor environments, building spatial maps and skill libraries that enable rapid adaptation when given specific tasks later.

5.3 Game AI

In game development, Zen-Voyager serves as an automated game tester that systematically explores game states, discovering edge cases and bugs through curiosity-driven exploration.

6 Related Work

VOYAGER [2] pioneered LLM-powered open-ended Minecraft exploration with a skill library. DREAMER-V3 [3] demonstrated model-based RL in diverse environments. ICM [1] established the curiosity-driven exploration paradigm. RIDE [4] introduced episodic novelty for improved exploration. Zen-Voyager advances this line by scaling to 14B parameters, introducing the epistemically calibrated ICT, and demonstrating zero-shot transfer that exceeds task-specific baselines.

7 Conclusion

Zen-Voyager demonstrates that open-ended exploration at 14B scale, with epistemically calibrated intrinsic motivation and a structured Skill Library Protocol, enables agents to discover skills and achieve generalization that surpasses task-specific trained baselines. The 78.3% Mine-Dojo success rate, 85.2% Procgen zero-shot transfer, and 92.1% OWCB coverage represent significant advances over prior open-ended learning systems.

References

- [1] D. Pathak et al., “Curiosity-driven Exploration by Self-supervised Prediction,” ICML, 2017.
- [2] G. Wang et al., “Voyager: An Open-Ended Embodied Agent with Large Language Models,” NeurIPS, 2023.
- [3] D. Hafner et al., “Mastering Diverse Domains through World Models,” arXiv:2301.04104, 2023.
- [4] R. Raileanu and T. Rocktaschel, “RIDE: Rewarding Impact-Driven Exploration for Procedurally-Generated Environments,” ICLR, 2020.