GLANCE: Driving Exploration in VLM Agents via Visual-Linguistic Curiosity

Haoxi Li [1] Qinglin Hou [2] Jianfei Ma [3] Jinxiang Lai [1] Tao Han [1] Sikai Bai [1] Jingcai Guo [4] Jie Zhang [1] Song Guo [1]

[1]The Hong Kong University of Science and Technology [2]University of Southern California [3]National University of Singapore [4]Hong Kong Polytechnic University

Problem Introduction: Passive Reasoning Is Not Exploration

VLM agents increasingly internalize world modeling capabilities into their policies via explicit CoT reasoning, enabling them to mentally simulate futures before acting.

Yet passive reasoning over visited states is insufficient. In sparse-reward tasks, agents lack the epistemic drive to actively seek the "known unknowns" needed to refine their world model.

As a result, an agent may perfectly explain a "dead end"—without ever exploring the alternative path that would reveal its mistake.

Research Question:

> Can VLM agents actively find signals that challenge and refine their internal world model through curiosity-driven exploration?

Prediction Error from Different Environments

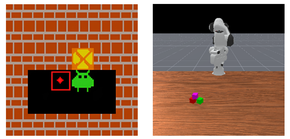

Figure 1. Prediction error examples. Left: Sokoban. Right: PrimitiveSkill.

Key intuition: the discrepancy between a linguistic prediction and visual reality highlights potential weaknesses in the agent's world modeling capability.

Research Motivation: Ground Thinking in Seeing

Standard curiosity methods focus on predicting latent visual futures from visual pasts. For VLM agents, this can decouple exploration from the agent's linguistic reasoning process.

GLANCE instead follows the principle:

> What the VLM agent thinks should predict what it sees.

Effective exploration must drive the VLM agent towards interactions where its **linguistic hypothesis of the future fails to align with visual reality**, forcing the internalized world model to ground itself through active falsification.

Agentic Task Domains

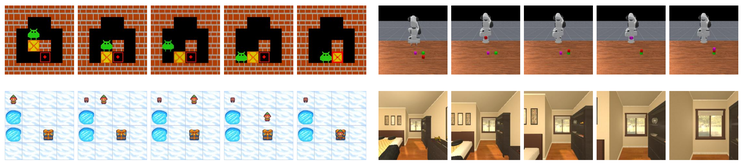

Figure 2. Examples of visual states from four environments used in our study.

We evaluate across a diverse suite of four agentic task domains: **Grid Puzzles**, **Navigation**, **Object Manipulation**, and **Geometric Reconstruction**.

Methodology: GLANCE

GLANCE is a unified framework that bridges reasoning and exploration by grounding the agent's linguistic world model into stable visual representations of an evolving target network.

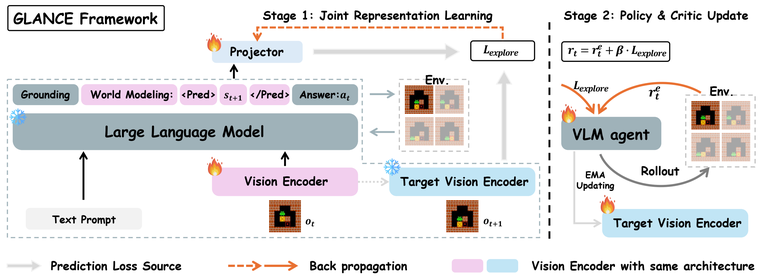

Figure 3. Overview of the GLANCE framework.

The architecture consists of two parallel streams:

- **Online VLM Agent**: generates reasoning and a future-state prediction.
- **Target Network**: provides stable evolving regression targets from the next observation.

Grounding Reasoning via Visual Alignment

Given the extracted linguistic hypothesis state h_{t+1} from the online VLM's last prediction token, we map it from the semantic language space to the visual latent space using a projector g_ψ

$$\hat{y}_{t+1} = g_\psi(h_{t+1}). \tag{1}$$

Simultaneously, the momentum target network processes the next-turn visual observation o_{t+1}

$$y_{t+1} = \mathrm{sg}(f_\phi(o_{t+1})). \tag{2}$$

The prediction loss is defined as the MSE loss between normalized predictions and targets:

$$\mathcal{L}_{\text{explore}}(v, \psi, t) = \left\| \frac{\hat{y}_{t+1}}{\|\hat{y}_{t+1}\|_2} - \mathrm{sg}\left(\frac{y_{t+1}}{\|y_{t+1}\|_2} \right) \right\|_2^2. \tag{3}$$

This selective gradient routing updates the projector g_ψ and the online visual encoder f_v while preserving the LLM's semantic priors.

Curiosity as Active Exploration

To drive active exploration, we interpret the prediction error not merely as a representational loss, but as a proxy for epistemic uncertainty.

$$r_t^i = \beta \cdot \mathcal{L}_{\text{explore}}(v, \psi, t), \qquad r_t = r_t^e + r_t^i. \tag{4}$$

The same discrepancy is used in two ways:

- **Self-supervised grounding objective**: aligns linguistic world modeling with visual reality.
- **Intrinsic curiosity reward**: steers the agent to actively explore areas where its internal model is uncertain.

Experimental Evaluation

Main Results on General Agentic Benchmarks

Table 1. Main results on the general agentic benchmarks.

Model/Method	Sokoban	FrozenLake	Navigation	PrimitiveSkill	Overall
Foundation VLMs					
Qwen2.5-VL-3B	0.13	0.14	0.23	0.00	0.21
Claude 4.5 Sonnet	0.31	**0.80**	0.67	0.53	0.64
Dense Extrinsic Rewards RL with World Model Reasoning for Visual States					
VAGEN-Full	0.79	0.72	0.81	0.97	0.81
GLANCE-Full	**0.85**	0.78	**0.87**	**0.97**	**0.86**
Sparse Extrinsic Reward RL with World Model Reasoning Strategy					
VAGEN-Base	0.61	0.71	0.79	0.91	0.76
GLANCE-Base	0.74	0.73	0.81	0.94	0.80
Turn-level PPO with World Model Reasoning Strategy					
Turn-PPO w/ Mask	0.38	0.68	0.81	0.25	0.58
GLANCE w/ Turn-PPO	0.52	0.70	0.79	0.66	0.69

Finding: GLANCE consistently outperforms exploitation-based RL methods in VLM agents across dense-reward, sparse-reward, and turn-level PPO settings.

Ablation Study on GLANCE Components

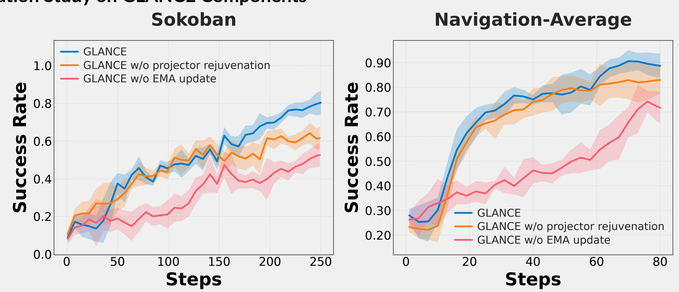

Figure 4. Ablation study on GLANCE components.

- **Projector rejuvenation** prevents intrinsic reward from diminishing too early.
- **Momentum target encoder** provides a critical, consistent anchor for grounding the agent's linguistic hypothesis.

Visualization of Curriculum Dynamics

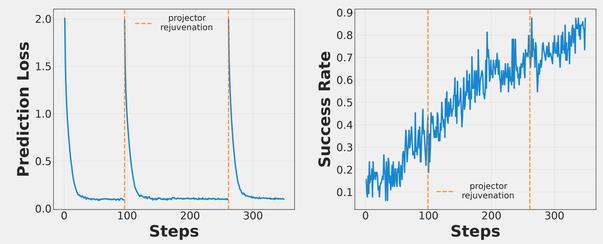

Figure 5. Temporal correlation between prediction loss and task success rate.

The sawtooth-style pattern indicates that each projector rejuvenation intentionally spikes the prediction loss, rejuvenates the intrinsic curiosity signal, and helps the VLM agent break through performance plateaus.