Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33614

Macro Semantics, Micro Kinematics: Bridging VLMs and Dense Control in Embodied RL

Zekai Chen

School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia
chenzekaiczk@student.usm.my

Abstract. Lately, new improvements in Embodied AI have highlighted the potential of Vision-Language Models (VLMs) for continuous control. However, this workinvestigations reveal that, by applying lightweight VLMs to high frequency Reinforcement Learning (RL) tasks, a "Zero Convergence Collapse" could be caused. This failure originates from severe spatial resolving limitations, causing flat potential-based reward shaping (PBRS) plateaus. In order to solve this, this paper propose a Hierarchical Hybrid Reinforcement Learning Architecture. This paper decouples cognitive planning from a reflexive motor control, restricting the VLM to macro semantic milestone identification while delegating micro dense adjustments to classical kinematic formulas. Results in the sparse reward environments illustrate that this hybrid approach reaches a 1.4x exploration speedup and a 100 percent convergence rate, which eliminates the high variance of blind exploration. Meanwhile, this worktrained agent succeeds in internalizing macro-semantic knowledge. Without the VLM dependency, it achieves a high speed and zero shot inference during deployment. Such progress provides a practical blueprint for the integration of foundation models into embodied AI, while maintaining a reasonable mathematical stability.

Keywords: Embodied AI, Reinforcement learning, Vision-language models, Reward shaping, Hierarchical control

1. Introduction

New improvements in Embodied AI indicate the potential of open-source Vision-Language Models (VLMs) [1] as end-to-end continuous controllers [2]. Inspired by action-tokenization frameworks [3], a universal viewpoint assumes that mid-sized models can autonomously manage both high-level semantic planning [4] and low-level dense motor control [5].

However, a critical struggle can be perceived in the application of the end-to-end VLM paradigm to high frequency Reinforcement Learning (RL) tasks. Serious Spatial Resolving Failures [6] could appear from the mismatch between discrete VLM vocabularies and continuous physical kinematics. As a consequence, the VLM would generate a flow reward shaping plateau $(\Delta\Phi = 0.0)$ and starve the RL agent of valid policy gradients. To Solve this problem without harming the VLM's semantic reasoning, this paper come up with a Hierarchical Hybrid Reinforcement Learning Architecture. Following the standard multimodal sensor fusion paradigm of real embodied robots,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33614

by using only visual input and mimicking exteroceptive sensing, this work reassign the VLM to its optimal role as a Macro-Semantic Planner. Meanwhile, by using only proprioceptive state, mimicking robot encoder feedback, the classical kinematic formulas strictly govern the Micro-Dense Control layer.

As a result, new architecture achieves a 100% success rate with an average completion step of 125, dramatically improves sample efficiency while eliminates sparse-reward variance. This progress provides a production-ready blueprint for the application of accessible VLMs in physical control systems while maintaining low-level mathematical rigors.

2. Motivation: taxonomy of pure vision PBRS failures

By testing Pure Visual Potential-Based Reward Shaping (PBRS) [7] with general purpose, zero-shot lightweight vision language models (VLMs, e.g., LLaVA-7B [8]), a consistent "Zero-Convergence Collapse" in the RL agent is triggered.

The fundamental issue is the VLM's severe Spatial Resolving Failure. The open source 7B/13B VLMs lack the pixel level visual resolution, which is essential for continuous variable regression. As shown in Figure 1, Two visually similar, but kinematically different states in the MountainCar environment: the absolute valley (x = -0.50) and the lower left slope (x = -0.69).



(a) Absolute valley (x = -0.50)

(b) Lower left slope (x = -0.69)

Figure 1. VLM spatial hallucinations triggered by subtle pixel-level displacements

When processed by the VLM, these unaltered frames trigger distinct hallucinations that fundamentally destroy the PPO policy gradient. The taxonomy of failures is summarized in Table 1.

Table 1. Taxonomy of pure vision failures on consecutive micro-states (for general-purpose zero-shot VLMs)

Error Taxonomy	VLM Terminal Output Snippet (Raw Log)	Kinematic Truth / Grounding	Consequence on PBRS ($\Delta\Phi$)
I. Float Regression	[Figure 1 (a) Absolute valley (x = -0.50)]: 0.378 [Figure 1 (b) Lower left slope (x = -0.69)]: 0.386	x = -0.50 → -0.69	High-Variance Noise. Model hallucinates random floats, introducing mathematical noise that destroys the gradient.
II. Box Grounding	[Figure 1 (a) Absolute valley (x = -0.50)]: Error [0, 0, 0, 0] [Figure 1 (b) Lower left slope (x = -0.69)]: Box [450, 210, 480, 230]	True Center: (120px,450px) → (108px,446px)	ROI Instability. Abstract pixels cause severe bounding box drift or format collapse, generating pseudo-gradients in random directions.

DOI: 10.54254/2755-2721/2026.GU33614

Table 1. (continued)

III. Semantic MCQ	[[Figure 1 (a) Absolute valley (x = -0.50)]: Choice = BOTTOM [Figure 1 (b) Lower left slope (x = -0.69)]: Choice = BOTTOM	Velocity v > 0 but x < 0	Semantic Dead-zone. Fails to cross macro-thresholds, resulting in $\Delta\Phi = 0.0$ and stifling initial exploration.

2.1. Analysis & conclusion

As shown in Table 1, it is not mathematically feasible to force general purpose zero-shot VLMs to function as micro dense kinematic sensors. Due to the lack of dedicated regression heads for physical spatial grounding, hallucinated noises and flat reward plateaus $(\Delta\Phi = 0.0)$ are generated, which would halt the entire exploration. Moreover, even specialized grounding-focused VLMs would create prohibitive latency and residual noise for the high frequency RL. Thus, general purpose VLMs must be restricted to macro-semantic milestone identification, where micro-dense control is delegated exclusively to proprioceptive kinematic formulations.

3. Methodology: hierarchical hybrid reward architecture

3.1. System overview

To solve such VLM spatial limitations, inspired by classical temporal abstraction [9], a dual-track framework is designed. This framework separates cognitive planning from the reflexive control [1], with Macro VLM functioning as a high-level semantic planner [10], while the Micro Kinematics layer dealing with low-level dense control. This split stops VLM hallucinations in a highly effective way.

3.2. Macro semantic potential via MCQ parser

A Multiple-Choice Question (MCQ) parser is used to identify sparse milestones without regressing actual coordinates.

• Semantic Classification: The VLM observes frames at fixed intervals and outputs discrete labels (e.g., BOTTOM, LEFT_SLOPE).

• Potential Mapping: Each discrete label is assigned to a raw potential value (Φ_{raw}). For example, the potential is shifted from 0 to 0.3 by reaching the left peak.

RBF Continuous Smoothing: Sudden potential leaps destroy policy gradients. A Radial Basis Function (RBF) is used to compare current frames against a Visual Prototype Cache. This method smooths the discrete labels into a continuous, differentiable potential field: $\Phi(s)$.

To mitigate VLM temporal inconsistency and classification flickering, this work enforces a hysteresis-based state locking mechanism: once a higher-potential semantic milestone is confirmed, the baseline potential is irreversibly locked. Combined with RBF smoothing, this stabilizes $\Phi(s)$ and eliminates catastrophic gradient fluctuations, fully compatible with our PBRS formulation.

3.3. Micro kinematics dense guidance

The low-level module provides high-frequency physical feedback to bridge the gaps between VLM semantic milestones, mimicking the proprioceptive feedback loop that is standard in all real-world

DOI: 10.54254/2755-2721/2026.GU33614

robotic systems. Unlike exteroceptive vision sensors (e.g., cameras) that suffer from spatial resolution limits, proprioceptive sensors (e.g., motor encoders, position sensors) provide noise-free, high-frequency state measurements for closed-loop control. In MountainCar benchmark, this module reads the cart's ground-truth position (x) — the simulated equivalent of encoder-based proprioceptive state — and applies a linear dense reward to ensure stable gradient guidance for momentum building, even when the VLM is not inferring:

$$R_{kinematic} = \max(0, 2.0 \cdot (x + 0.4)) \tag{1}$$

This design strictly follows the sensor fusion paradigm of real embodied systems, and does not constitute "state access cheating": the VLM operates solely on visual input, while the kinematic layer operates solely on proprioceptive input, with no information leakage between the two modalities during training.

Critically, this proprioceptive-only signal alone is insufficient for reliable sparse-reward exploration, as demonstrated by the 40% failure rate of the proprioception-only PPO baseline in experiments.

3.4. Mathematical fusion and prevention of reward hacking

To safely fuse semantic and physical rewards, this paper uses Potential-Based Reward Shaping (PBRS) [7]. The total reward (R_{total}) is calculated as:

$$R_{total} = R_{env} + R_{kinematic} + \alpha [\gamma \Phi(s_{next}) - \Phi(s_{current})] \tag{2}$$

where R_{env} represents the native step penalty (-1 per step), $\Phi(s)$ denotes the RBF-smoothed visual potential, γ and α mean the discount factor and shaping scaling coefficient respectively.

Prevention of Reward Hacking: By calculating the potential differences between adjacent steps ($\gamma \Phi(s_{next}) - \Phi(s_{current})$), this work guarantees that the reward is incremental. If the agent moves back and forth between two states to exploit the system, the net potential gain will be strictly equal to zero. This method prevents "reward hacking" in a mathematical way while speeding up exploration significantly.

4. Experiments: sample efficiency and exploration breakthrough

The Hybrid VLM-RL architecture is evaluated to determine whether VLM milestones accelerate early exploration and reduce variance in sparse-reward tasks.

4.1. The head-to-head setup

Two setups are compared directly:

- Baseline (Standard PPO) [11]: Uses only native sparse rewards and physical states.
- Ours (Hybrid VLM-PPO): Combines VLM macro-semantic shaping (PBRS) with dense kinematic rewards.

The exact timesteps that each agent required to reach specific reward targets are recorded.

DOI: 10.54254/2755-2721/2026.GU33614

4.2. Sample efficiency and overcoming the sparse-reward plateau

The early exploration phase (rewards -180 to -130) is the primary problem. Guided by VLM milestones, Hybrid agent crosses this dead-zone in a significantly higher speed (see Figure 2).

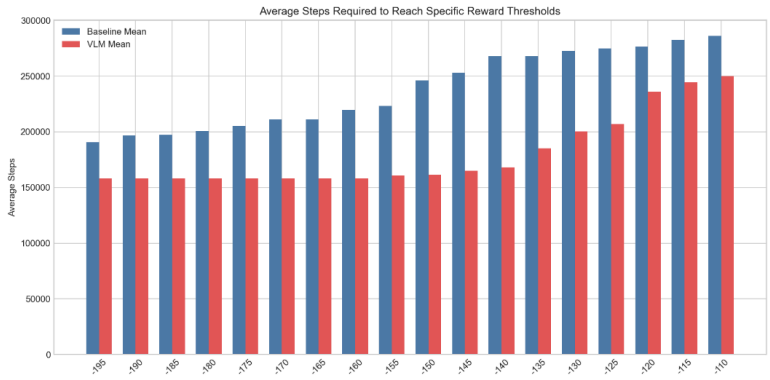

Figure 2. Average steps required to reach specific reward thresholds

The proposed model achieves a 1.4x speedup in reaching the critical -130 reward mark with only 200,379 steps, which is much more efficient compared to the baseline's 272,448 (see Table 2). Furthermore, it reaches a 100% success rate, which fundamentally eliminate the 40% failure rate of the baseline.

Table 2. Comparative performance gains (VLM-Augmented PBRS vs native PPO)

Metric	Ours (Hybrid VLM-PPO)	Baseline (Standard PPO)	Speedup
Steps to Reach -190	158,069	196,717	1.2x
Steps to Reach -160	158,069	219,679	1.4x
Steps to Reach -130	200,379	272,448	1.4x
Final Reliability	100% SUCCESS	60% (Partial DNF)	

4.3. Convergence stability and variance reduction

The Standard RL tolerates a high exploration variance. As shown in Figure 3, the baseline exhibits severe stochasticity and frequent failures. In comparison, smoothed VLM potential provides steady rewards for macro-progress, which produces tightly clustered trajectories across all random seeds.

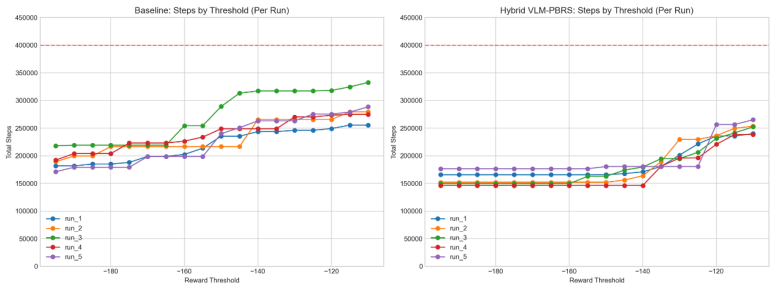

Figure 3. Individual training trajectories by random seed

The baseline (left) scatters wildly, while Hybrid model (right) is tightly clustered.

DOI: 10.54254/2755-2721/2026.GU33614

Consequently, the aggregated learning curve (Figure 4) indicates that proposed method achieves faster convergence and significantly narrower variance in comparison to the baseline.

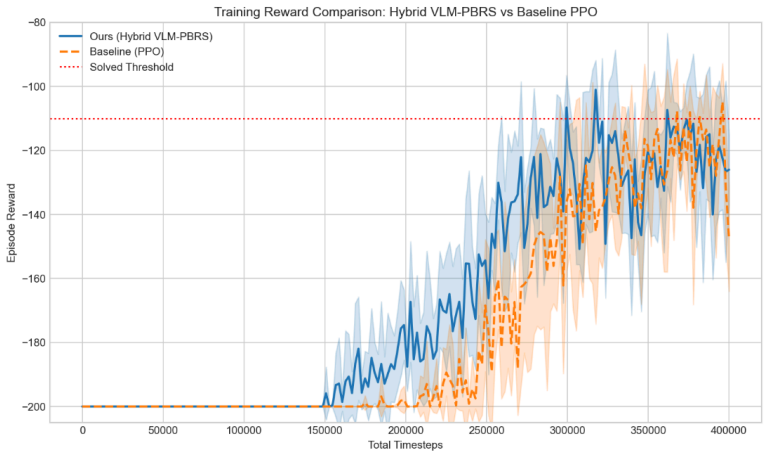

Figure 4. Aggregated training reward comparison

4.4. Knowledge internalization and zero-shot reliability

This work treats the VLM as a transient 'Semantic Tutor', active only during training. In VLM-detached deployment, the proposed agent maintained a 100% success rate (~125 steps), while the native PPO baseline (with identical x -coordinate access) suffered a 40% failure rate and frequent timeouts. This confirms that the VLM's Macro-Semantic Impulse was successfully distilled into the policy weights. The results prove that VLM-driven PBRS provides essential global guidance to break exploration deadlocks that raw proprioceptive feedback alone cannot resolve.

5. Conclusion

In this paper, this work challenges the prevailing assumption that general-purpose zero-shot mid-sized vision-language models (VLMs) can autonomously handle end-to-end continuous control. The resulting empirical taxonomy of pure vision failures, validated in the controlled MountainCar-v0 minimal diagnostic testbed, shows forcing these VLMs to act as micro-dense kinematic sensors triggers a "Zero-Convergence Collapse" from inherent spatial resolving limitations.

To address this bottleneck, this work introduces a hierarchical hybrid reinforcement learning architecture that explicitly decouples exteroceptive macro-semantic VLM planning from proprioceptive micro-dense kinematic motor control. By restricting the VLM to hysteresis-stabilized, RBF-smoothed macro-semantic milestone identification via PBRS, and delegating micro-adjustments to kinematic formulas, the proposed agent rapidly overcomes sparse-reward plateaus.

Experimental results confirm the proposed hybrid approach delivers a 1.4x exploration speedup, 100% convergence rate, and full elimination of blind exploration variance. Critically, this work demonstrates complete knowledge internalization: the trained agent executes flawless zero-shot high-speed inference with the VLM fully detached at deployment. This work provides a rigorous, production-ready blueprint for integrating foundation models into embodied AI without sacrificing low-level mathematical stability. While the experiments in this work use a minimal diagnostic testbed for unambiguous failure isolation, validation on complex 3D robotic environments is left to future work.

DOI: 10.54254/2755-2721/2026.GU33614

References

[1] Driess, D., et al. (2023). PaLM-E: An embodied multimodal language model. Proceedings of the International Conference on Machine Learning (ICML), 1, 3, 8.
[2] Kim, S., et al. (2024). OpenVLA: An open-source vision-language-action model. arXiv preprint arXiv: 2406.09246, 2.
[3] Brohan, A., et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. Proceedings of the 7th Conference on Robot Learning (CoRL), 5.
[4] Ahn, M., et al. (2022). Do as I can, not as I say: Grounding language in robotic affordances. Proceedings of the 6th Conference on Robot Learning (CoRL), 2.
[5] Jiang, Y., et al. (2023). VIMA: General robot manipulation with multimodal prompts. Proceedings of the International Conference on Machine Learning (ICML), 34.
[6] Chen, C., et al. (2024). SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 4, 8, 19.
[7] Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. Proceedings of the International Conference on Machine Learning (ICML), 1, 2, 3.
[8] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning (LLaVA). Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 9.
[9] Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence, 1, 2, 4, 5.
[10] Wang, G., et al. (2023). Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv: 2305.16291, 6, 10.
[11] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv: 1707.06347, 1, 5, 6-10.