Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33541

Vision-Language Models for Automated Reward Annotation in Embodied AI: State-of-the-Art, Challenges, and Future Directions

Nachuan Yan

School of Computing and Data Science, Xiamen University Malaysia, Sepang, Malaysia
swe2409061@xmu.edu.my

Abstract. With the rapid development of embodied AI, reward function engineering has become a main bottleneck in Reinforcement Learning (RL). Traditional RL highly relies on manual annotation, which is not only time-consuming, error-prone, and also difficult to handle real-world problems such as sparse rewards and cross domain data scarcity. To address this bottleneck, there are number of various automated reward labeling mechanisms leveraging Vision-Language Models (VLMs) are proposed. In this paper, a systematic survey on five state-of-the-art landmark solutions-VIP, LIV, Eureka, DrEureka and EMMA-X-and catagorize to compare how they differ in terms of sample efficiency, sim-to-real transferability, and reasoning robustness. Furthermore, this paper introduces a comprehensive taxonomy of evaluation metrics to help researchers quantify the alignment between VLM-generated rewards and true task objectives. Finally, this paper identify and discuss critical open challenges, such as visual grounding errors and high latency, while proposing strategic directions for future research in the field of autonomous reward synthesis.

Keywords: Embodied AI, Vision-Language Models (VLMs), Reinforcement Learning, Reward Engineering, Automated Reward Labeling

1. Introduction

Enabling autonomous intelligent bodies to perform complex tasks in unstructured environments has made body artificial intelligence a cutting-edge field of contemporary robot research. Reinforced learning (RL) is the core of this research, which is highly dependent on the carefully designed reward function R(s,a). However, with the increasing complexity of robot tasks, the traditional reward engineering method has become a bottleneck that hinders the landing of physical intelligence on large-scale complex scenarios [1].

In the past, reward functions were mainly manually generated by human experts. This process is not only time-consuming and energy-intensive, but also easy to lead to the phenomenon of "reward cracking" in the model, that is, the intelligent body uses logical loopholes instead of actually achieving the task goal. In addition, robots also face the challenge of "sparse rewards" - feedback

© 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.GU33541

can only be obtained after the task is completed - which will lead to inefficient exploration and long convergence time.

In recent years, the rise of basic models, especially the Visual Language Model (VLM), has provided transformative solutions to these challenges [2]. Researchers use basic VLMs such as GPT-4 and CLIP [3] to automate reward annotation, so that robots can learn from primitive visual observation instead of relying on manually defined signals.

Based on this, this article provides a comprehensive review of this emerging field and summarizes the existing important literature into two main technical directions:

• Representation-based Learning: Derive reward signals from the distance in the potential embedded space(e.g., VIP [4], LIV [5]).

• LLM-driven Code Generation: Use the large language model (LLM) to write executable reward code(e.g., Eureka [6], DrEureka [7]).

Table 1. Chronological evolution of VLM-based reward specification

Year	Framework	Evolutionary Milestone
2022	VIP	First to treat human videos as a source of implicit value functions for zero-shot visual rewards.
2023	LIV	Integrated textual instructions with visual embeddings, enabling multi-modal reward conditioning.
2023	Eureka	Shifted from Representation-based to LLM-driven code generation.
2024	DrEureka	Automated the tuning of domain randomization alongside rewards, bringing VLM supervision to physical hardware.
2025	EMMA-X	Merged Grounded CoT with VLA architectures, addressing the hallucination problem in complex spatial tasks.

2. Background and related work

2.1. Reinforcement learning and the reward specification problem

RL is typically framed as a Markov Decision Process (MDP), defined by the tuple (S, A, P, R, γ). The agent's objective is to learn a policy $\pi(a|s)$ that maximizes the expected cumulative reward $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$.

In complex embodied tasks, defining the reward function R(s, a) is difficult. Traditional approaches rely on Reward Shaping, which adds auxiliary terms to the sparse environment reward to guide exploration. However, manual reward shaping requires extensive domain expertise and tends to Reward Hacking. Due to this limitation, automated reward generation methods that can provide dense and semantically meaningful feedback without human intervention are being searched for [8].

2.2. Vision-language models (VLMs) as semantic priors

The emergence of Foundation Models, especially Vision-Language Models (VLMs), has revolutionized how robotic systems perceive and interact with the world. Contrastive Language-Image. Pre-training (CLIP) and other models learn a joint embedding space for images and text by training on huge datasets from the internet. This shared latent space makes it possible to figure out how semantically similar the robot's visual observations are to a written task description, for example, "open the door."

DOI: 10.54254/2755-2721/2026.GU33541

In terms of reward generation, VLMs serve two main roles:

1. Implicit Distance Metrics: Using the cosine similarity as a dense reward signal between the current state embedding and a goal embedding.

2. Reasoning Engines: Using LLMs' capability of high-level reasoning to interpret environmental feedback and synthesize procedural logic for rewards.

These models provide the necessary "common sense" that allows agents to understand task progress in a zero-shot or few-shot manner, significantly reducing the dependence on task-specific, labeled data.

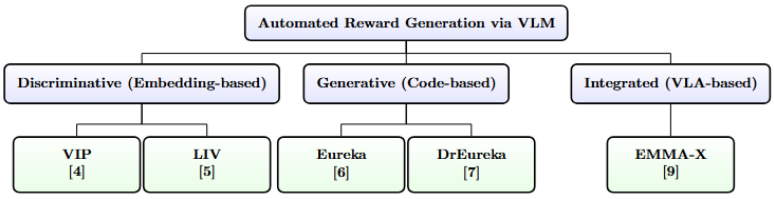

Figure 1. Taxonomy of VLM-based reward generation frameworks

3. Methodology and framework analysis

This section goes into a detailed analysis of five important frameworks that use VLMs to automatically generate rewards. We put these methods into groups based on their mechanism for translating visual and linguistic inputs into scalar reward signals.

Table 2. Feature matrix of surveyed VLM-based reward frameworks

Feature	VIP	LIV	Eureka	DrEureka	EMMA-X
Zero-shot Generalization	√	√	×	×	√
Physical Grounding	×	×	√	√	√
Language Conditioning	×	√	√	√	√
Sim-to-Real Readiness	×	×	×	√	√
Inference Speed	Fast	Fast	Slow	Slow	Moderate

3.1. Representation-based reward learning

This approach learns a structured latent space where the distance between state embeddings indicate task progress.

3.1.1. VIP: value-implicit pre-training

In VIP [4], the reward function is designed to measure the progress made toward the goal by calculating the negative change in distance within the embedding space. The main idea is that a pre-trained visual representation should implicitly encode the "value" of a state. Formally, the reward at time t is defined as:

$$r_t = -\|\phi(s_t) - \phi(s_{goal})\|_2 + \|\phi(s_{t-1}) - \phi(s_{goal})\|_2 \quad (1)$$

where ϕ is the frozen VIP visual encoder. This formula quantifies the reduction in distance to the goal between $t-1$ and t; a positive reward indicates the agent has moved closer to the target.

DOI: 10.54254/2755-2721/2026.GU33541

This approach enables the generation of zero-shot rewards for diverse robotic tasks without requiring robot-specific training data [10].

3.1.2. LIV: language-image value learning

LIV [5] incorporates language instructions via the frozen VIP visual encoder. By coordinating visual trajectories with textual descriptions using a contrastive objective, LIV allows the agent to obtain rewards for precise language objectives. This makes the reward signal more generalizable to open-vocabulary instructions.

3.2. Llm-driven procedural code generation

This approach utilizes LLMs' reasoning capabilities to construct executable reward functions, which is different from representation-based approaches.

3.2.1. Eureka and DrEureka

Eureka [6] employs GPT-4 in Python to generate reward sub-functions. The process utilizes a search strategy that is evolutionary in nature.LLMs generate Python reward sub-functions which are subsequently refined iteratively for the candidates on the basis of simulation feedback.

DrEureka [7] enhances Eureka by automating the transition from simulation to real world. It makes use of Domain Randomization (DR) settings that render the resultant policies more robust to simulation–reality differences

3.3. The emergence of embodied VLA models

As mentioned before, these methods decouple the reward generation from the policy learning. A novel model called Vision-Language-Action (VLA) has recently emerged. These new models combine perception, reasoning and action in one end-to-end transformer architecture and improve automated reward synthesis.

3.3.1. EMMA-X: grounded reasoning and spatial forensics

EMMA-X [9] tackles this challenge by introducing a novel approach that grounds VLM-based rewards in specific 3D regions, thereby providing a solution to this problem. The main distinction of EMMA-X from both Eureka and VIP is the incorporation of a new method called Grounded CoT.

This approach makes the model output a spatial bounding box, performs linguistic reasoning, and calculates the reward. This logic makes sure that the reward signal comes from a real state in the world. For example, checking to see if a lid is really closed, and not just obscured. According to research, EMMA-X accomplishes the linking of abstract language comprehension to fine physical actions, representing an important technological path for creating self-improving robotic systems.

DOI: 10.54254/2755-2721/2026.GU33541

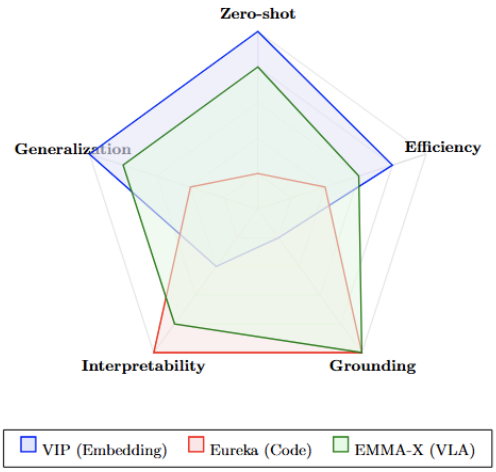

Figure 2. Qualitative comparison of seminal frameworks across performance dimensions

4. Tables

4.1. The latency-throughput trade-off in real-time control

A major challenge preventing wide adoption of VLM-based reward systems is the slow inference times caused by large-scale foundation models. To keep control loops stable, RL agents usually need a lot of feedback at high frequencies (e.g., 50–100Hz). Despite being effective, modern VLMs face inference time issues.

There are two main issues because of this difference in time:

• Reward lag: The "Temporal Credit Assignment" problem aggravates as we become more delayed in reward. The reward signal does not correspond to the actions taken.

• Asynchronous Bottlenecks: A slow Visual-Linguistic Model evaluator placed in a fast simulation environment can create asynchronous bottlenecks, thus slowing down the training process from hours to days.

Future work should be inspired by "Distillation" or "Hybrid Architectures. For instance, a heavyweight VLM should monitor the training of a lightweight, task-specific Reward Proxy model capable of acting at kilohertz on edge devices. The communities of software engineering and AI have a lot of work to do before they can combine complex semantic reasoning with simple reactive control.

4.2. Reward hacking and semantic misalignment

Though VLMs accurately provide reasoning, they remain susceptible to reward hacking, which is an issue where every RL agent achieves a high score without actually performing the task. When using a VLM as a source of supervision, failures often occur. When the language-based goal does not succeed, it is often triggered by semantic misalignment between the textual instructions and the visual state.

To coincide with the visual evidence, this study reveals some important causes of such failures:

• Visual Hallucinations: VLMs may incorrectly perceive task success because of the angle of the camera or motion blur. For example, an agent might learn to position its gripper in front of an object to trick the VLM into reporting that it "succeeded" based on 2D image projections.

DOI: 10.54254/2755-2721/2026.GU33541

• Heuristic Over-reliance: The functionalities generated by LLMs may encode excessive heuristics. They often overemphasize easy-to-measure proxies and ignore the contact physics associated with object regularity.

In order to counteract these vulnerabilities, they must implement formal verification, multi-view consensus, and ensure that the automated reward signal is derived from the physical world.

5. Conclusion

This paper has provided a systematic review of automated reward generation based on VLM by tracing feedback systems from human reward engineering to AI generated and frameworks such as VIP, LIV, Eureka, DrEureka, and EMMA-X.VIP and LIV are the most generalization capable frameworks. In the meantime, we find Eureka is more interpretable and provides higher accuracy on the task. Models of visual language AI like EMMA-X are a step toward spatial reasoning and reduce the hallucination of previous VLMs.

As stated in the section "Open Challenges and Future Directions", the reasoning delay and semantic misalignment issues are still the most severe that must be solved before fully autonomous learning can take place. As per recent studies, a closed-loop feedback system working in milliseconds with the flexibility to do deep semantic reasoning form a state-of-the-art research problem.

Connecting large foundational models to the RL loop at a deep level matters for embodied AI It assists agents in learning how to perceive, reason, and make decisions in complex human settings with less human involvement.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Man´e, "Concrete problems in ai safety, " arXiv preprint arXiv: 1606.06565, 2016.

[2] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid et al., "Rt-2: Vision-language-action models transfer web knowledge to robotic control, "

[] in Conference on Robot Learning. PMLR, 2023, pp. 2165–2183.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision, " in International conference on machine learning. PmLR, 2021, pp. 8748–8763.

[4] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training, " arXiv preprint arXiv: 2210.00030, 2022.

[5] Y. J. Ma, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman, "Liv: Language-image representations and rewards for robotic control, " in International Conference on Machine Learning. PMLR, 2023, pp. 23 301–23 320.

[6] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models, " arXiv preprint arXiv: 2310.12931, 2023.

[7] J. Ma, W. Liang, H.-J. Wang, Y. Zhu, L. Fan, O. Bastani, and D. Jayaraman, "Dreureka: Language model guided sim-to-real transfer." RSS, 2024.

[8] W. Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T. L. Chiang, T. Erez, L. Hasenclever, J. Humplik et al., "Language to rewards for robotic skill synthesis, " arXiv preprint arXiv: 2306.08647, 2023.

[9] Q. Sun, P. Hong, T. D. Pala, V. Toh, U.-X. Tan, D. Ghosal, and S. Poria, "Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning, " in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 14 199–14 214.

[10] B. Baker, I. Akkaya, P. Zhokov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro, and J. Clune, "Video pretraining (vpt): Learning to act by watching unlabeled online videos, " Advances in Neural Information Processing Systems, vol. 35, pp. 24 639–24 654, 2022.