北京通用人工智能研究院
Beijing Institute for General Artificial Intelligence

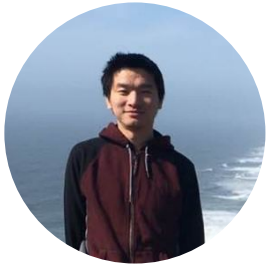
MVR: Multi-view Video Reward Shaping for Reinforcement Learning

Lirui Luo, Guoxi Zhang, Hongming Xu, Yaodong Yang, Cong Fang, Qing Li

Beijing Institute for General Artificial Intelligence (BIGAI)
Peking University

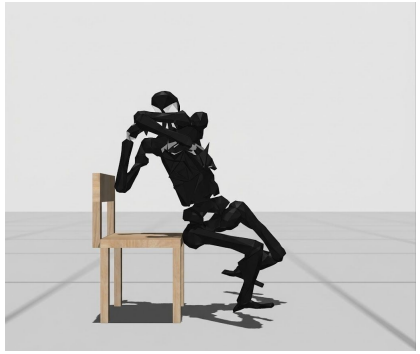
mvr-rl.github.io






The Necessity of VLM Rewards

Task reward guidance leads to incorrect visually fragile motion.



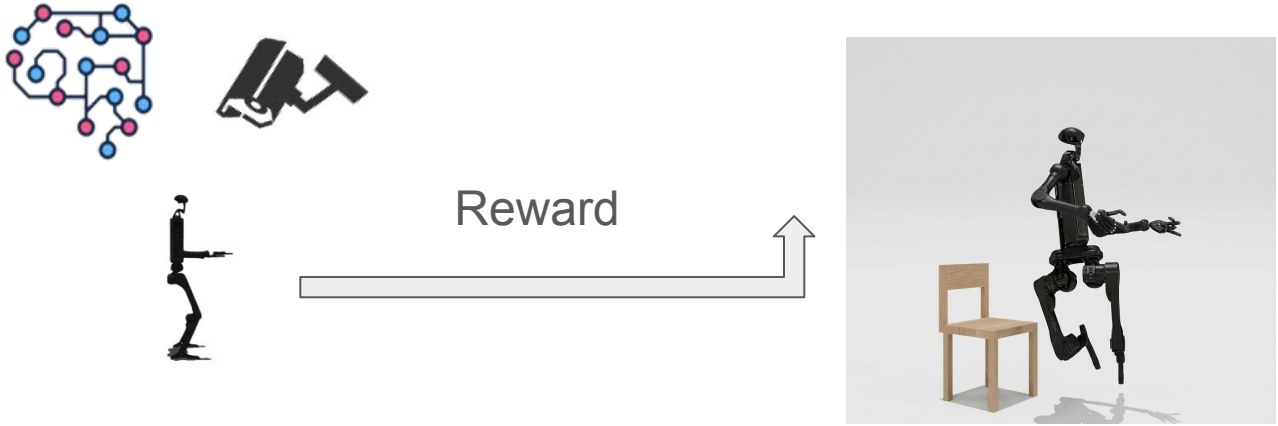
VLM rewards, combined with task rewards, guide towards visually robust motion.

Disadvantages of Traditional VLM Rewards

Single-frame and single-viewpoint approaches can provide **incorrect visual feedback** to the robot.

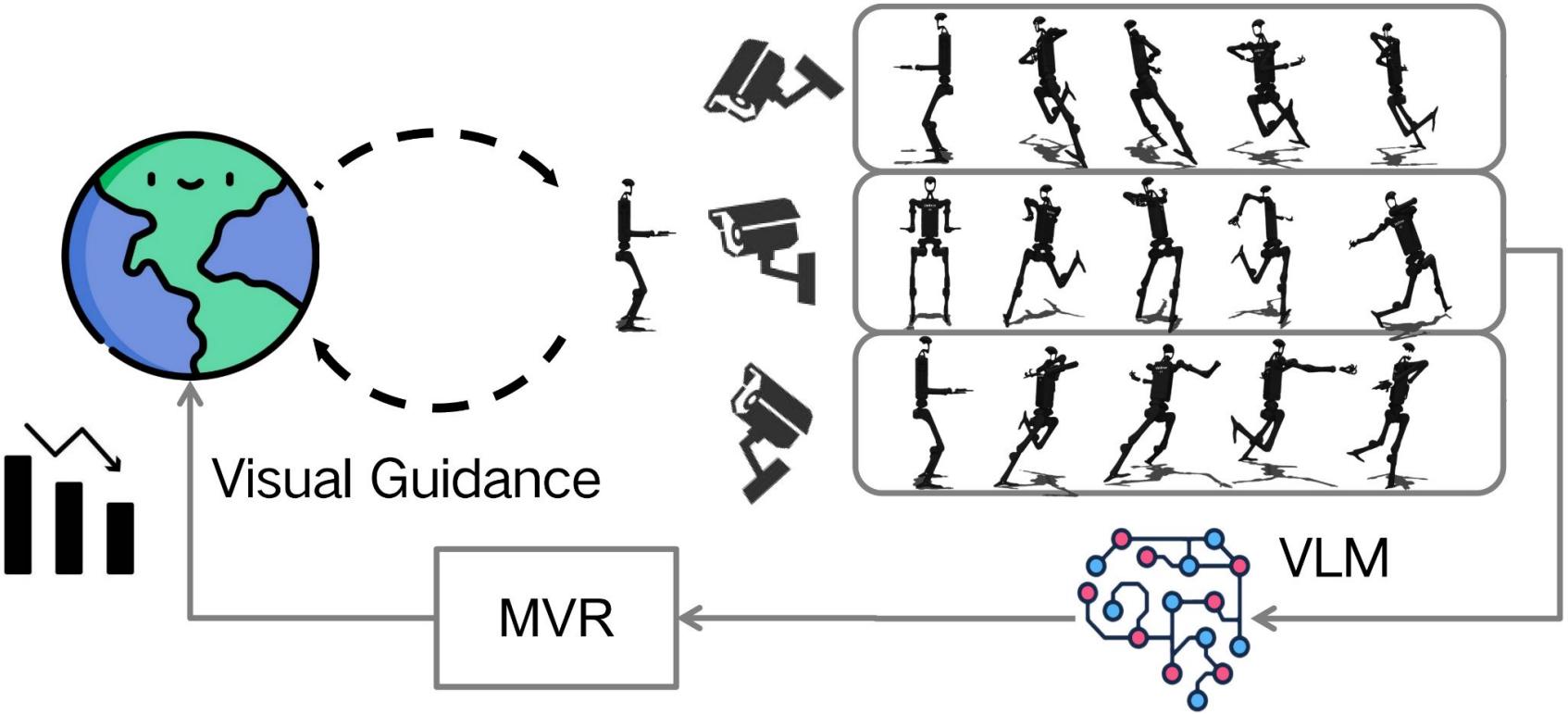
- Single-view leads -> Partial observability.
- Only frame -> No motion information.

Disadvantages of Traditional VLM Rewards



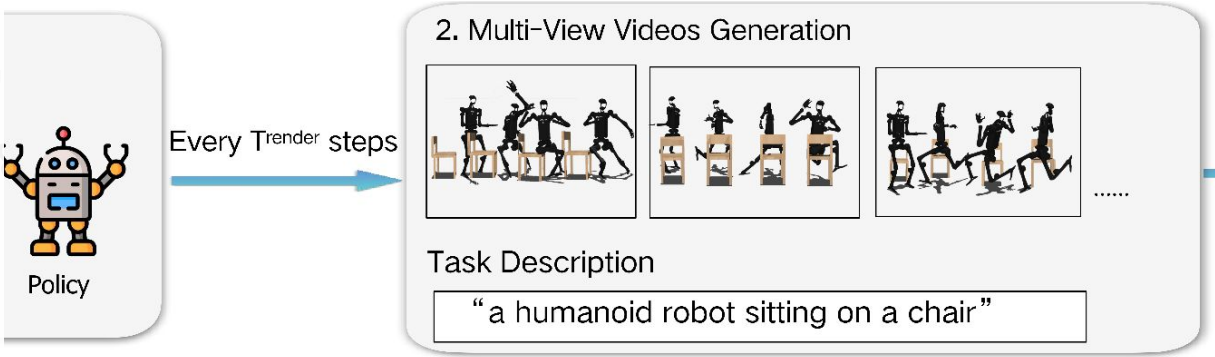
Non-decaying rewards can lead to **suboptimal policy**, such as **avoiding completing the task** in order to maintain a visually appealing pose.

Why MVR



MVR can extract **automatically decaying rewards** from **multi-view video**.

MVR: Multi-View Videos Generation

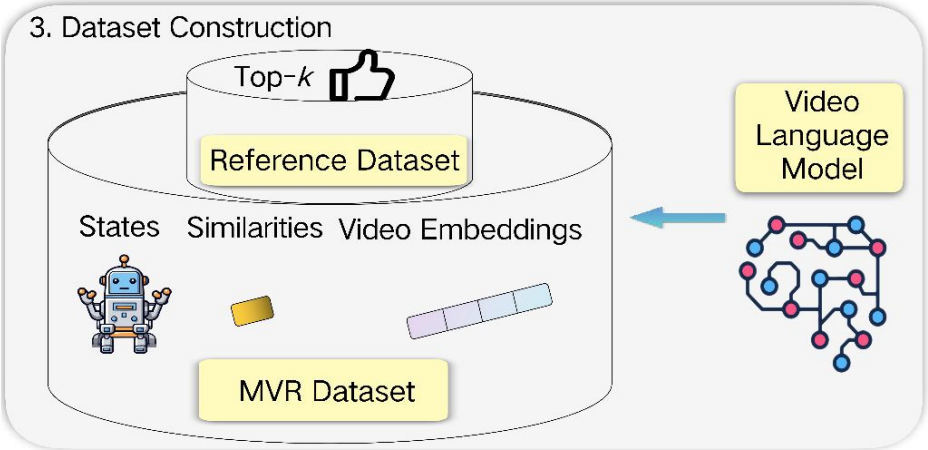


T^{render}-step runs once, so there **is almost no overhead**.

Only one of the four view is rendered at a time, requiring **no additional computing resources**.

MVR: Dataset Construction

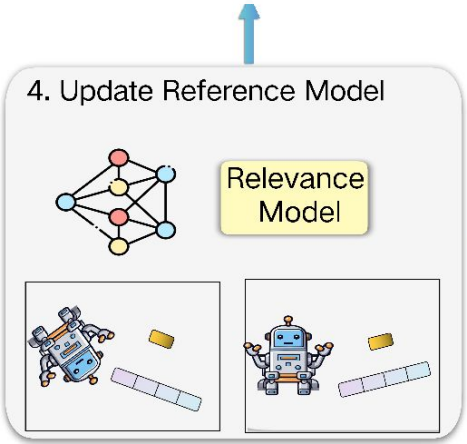
Maintaining a **reference dataset** to learn automatically decaying rewards.



MVR: Update Reference Model

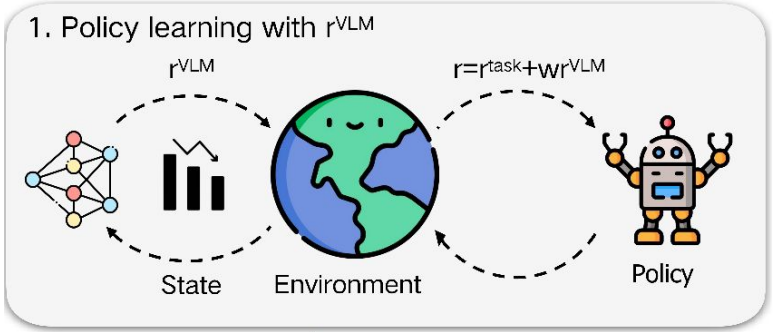
The reference model is **updated periodically**, which has **low overhead**.

The reference model awards **higher rewards to videos** that are **more text-based**.



Every T^{update} steps

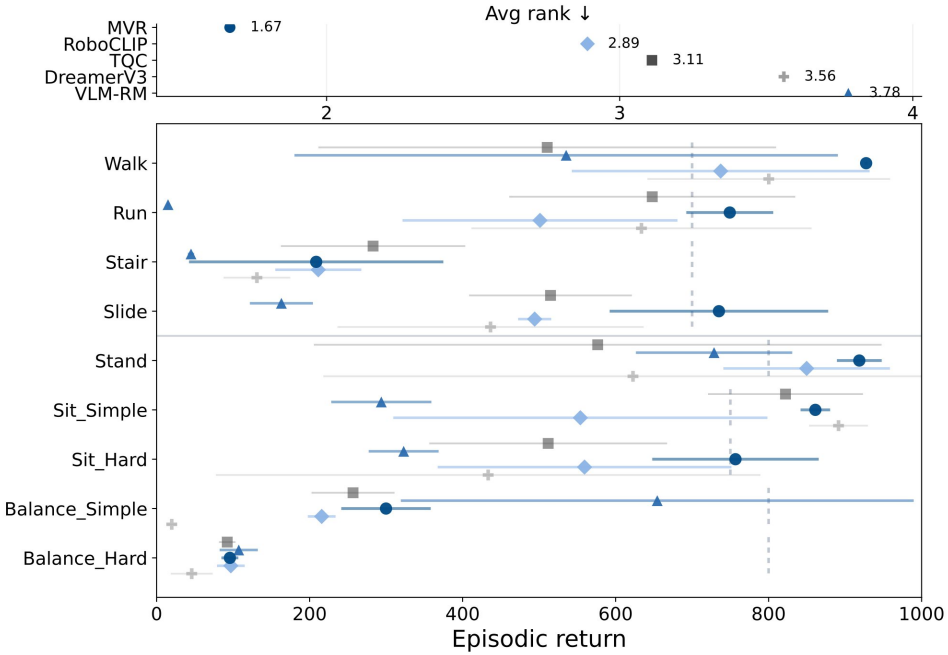
MVR: Policy Learning

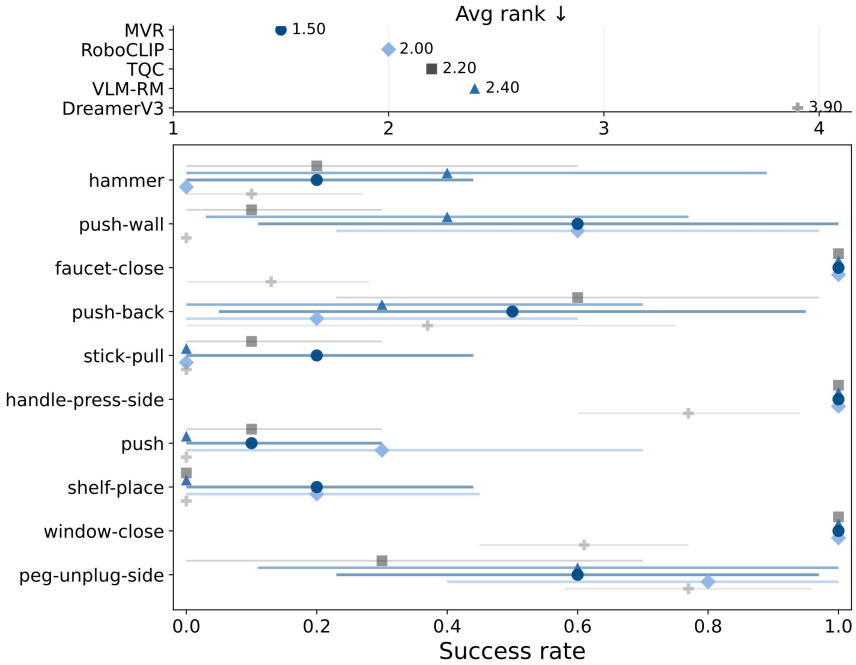


The reward given by MVR will cause the robot to **perform behavior that matches the language**.

Once it approaches the target behavior, the reward will **decay to prevent changing the optimal policy.**

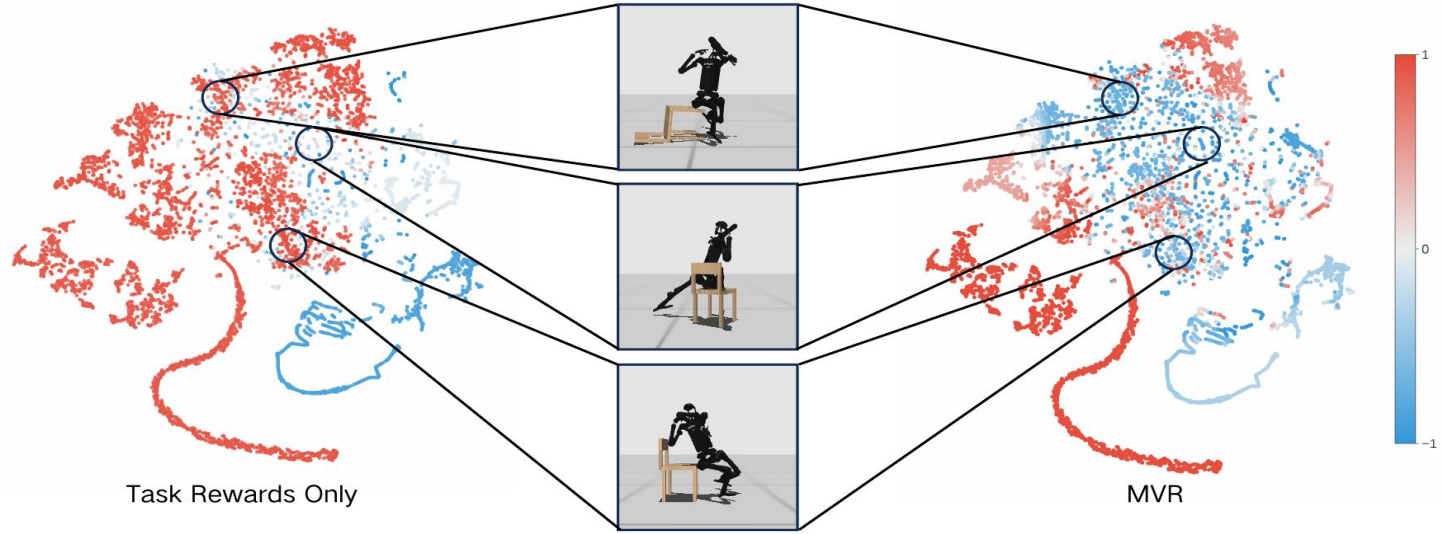
Performances





For Metaworld and Humanoid Bench, MVR gets the best average rank for the reward/success rate.

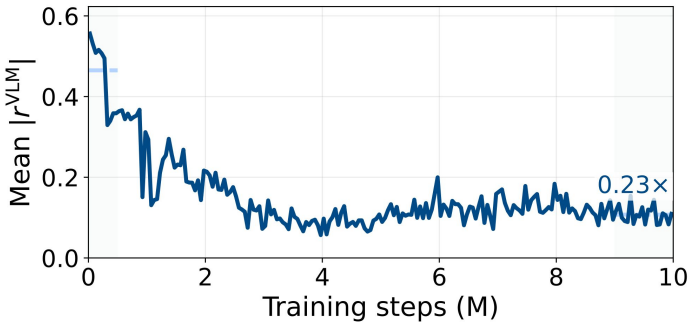
Case Study

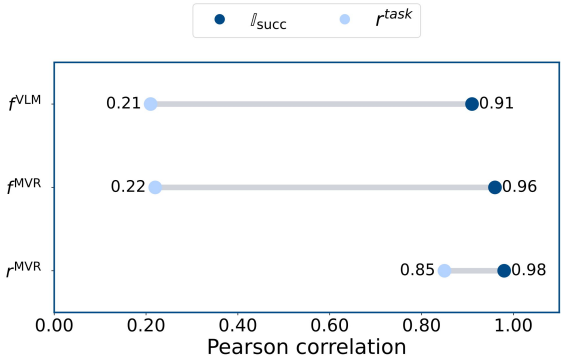


While task rewards are high when the agent is close to the chair, MVR's visual guidance component **assigns low values to improper sitting poses** (sitting on chair's leg, leaning, or sitting at the edge).

Case Study

MVR rewards automatically **decay to 23% of the initial value**, and are **highly correlated with success rate**.





Takeaways

We propose MVR, **a framework for generating automatic decaying feedback from multi-view videos**.

We introduce a **state-dependent reward shaping formulation** that integrates task rewards with VLM-based guidance and decays its influence as behaviour aligns with the target.

We **validate MVR's effectiveness** through simulated experiments on humanoid locomotion tasks in HumanoidBench and manipulation tasks in MetaWorld,.

Code & demo: mvr-rl.github.io

MVR: Multi-view Video Reward Shaping for Reinforcement Learning