MERL: Trust-Calibrated VLA Policy Post-Training with Evolving Imagination

CoRL 2026

Wenzhan Li[1,2] Yiran Qin[3,4,5] Heng Zhou[4,6] Huirui Wang[1] Yulan Guo[1,2] Ruimao Zhang[1*]

[1] Sun Yat-sen University [2] Shenzhen Loop Area Institute [3] The Chinese University of Hong Kong, Shenzhen
[4] Shanghai AI Laboratory [5] University of Oxford [6] University of Science and Technology of China

Abstract

Reinforcement-learning post-training adapts vision–language–action (VLA) policies beyond demonstration coverage. World models offer imagined experience, but simulator updates alone do not ensure reliable rollouts. **Model-Evolving Reinforcement Learning (MERL)** couples an online-updated simulator with **two-level trust**: *stage-level reliability* controls imagined-loss mixing and rollout horizon; *chunk-level reliability* controls replay sampling and optimization weight. A residual predictor scores recursive imagination *without querying future ground truth*. On perturbed LIBERO-PRO, MERL reaches **79.7%** average success and requires **10.9%** fewer high-threshold refinement stages than matched model-free RL. Fixed-data adaptation from 50 demonstrations per task reaches **80.0%** average success on two physical-robot tasks.

1. Background & Motivation

Deployment changes object placement and appearance while *task semantics remain fixed*. Grounded RL adapts VLA policies but **requires environment interaction**; world models offer additional imagined experience.

Two limitations remain:

- **Frozen simulators** *become stale* as policies change.
- **Updated simulators** still yield *unequal-quality chunks*.

Core insight: update the simulator, then calibrate *how imagination affects policy learning*.

2. Paradigm Difference

Imitation learning
Expert demos → VLA policy
Demo-limited; no own-failure feedback.

Model-free RL
Grounded data → RL update
Grounded feedback; interaction-intensive.

Static model-based RL
Frozen simulator → Imagined updates
Fixed imagination use; uniform replay.

Online-updated MBRL
Updated simulator → Imagined updates
Fixed imagination use; uniform replay.

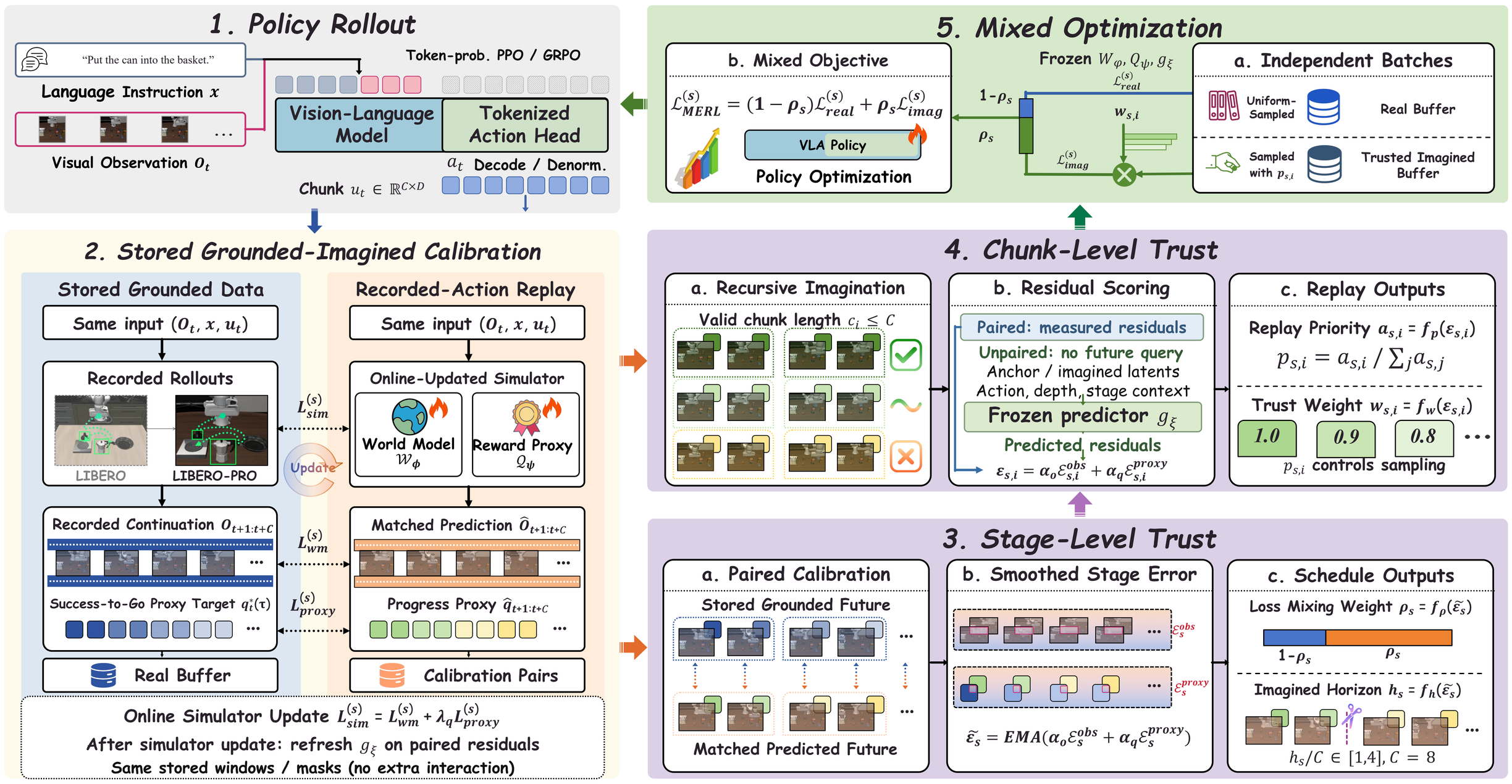
MERL (ours) Reliability guides four learning decisions
Updated simulator → Estimate reliability → Adaptive use → Mixed update

Stage: whole training round		Chunk: each imagined segment	
How much to learn? Loss mixing (ρ)	How far to imagine? Rollout length (h)	Which chunks? Sampling probability (p)	How much influence? Update weight (w)

Matched refinement controls; model-based updates also use grounded data.

3. Method: Evolving Imagination with Two-Level Trust

An evolving simulator generates experience; **two-level trust** *controls its use in policy learning*.



Stored paired replay refreshes the residual predictor after simulator updates; *calibration adds no interaction*.
Recursive imagination is scored *without future-ground-truth queries*. The simulator and residual predictor freeze during policy updates.

4. Simulation Results

LIBERO-PRO · four suites · three seeds (mean ± SD)

Method	Avg. SR (%) ↑	AUC ↑	S2T-H ↓
SimpleVLA-RL	77.1 ± 1.1	67.8 ± 1.0	104.1 ± 7.0
WMPO†	75.4 ± 1.2	63.2 ± 1.3	149.6 ± 10.2
WoVR†	77.8 ± 0.9	68.4 ± 0.9	101.3 ± 6.2
MERL	**79.7 ± 0.7**	**70.6 ± 0.8**	**92.1 ± 4.8**

† Baselines adapted to the same training/evaluation protocol.
AUC: normalized learning-curve area; S2T-H: stages to a high SR threshold.

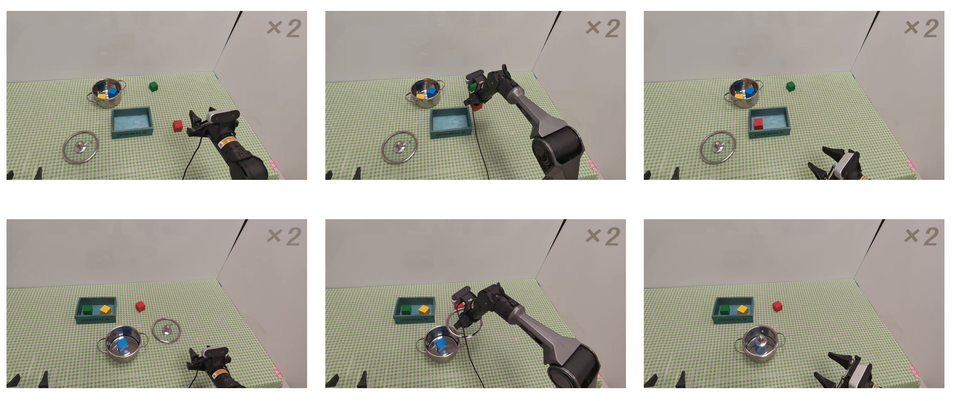
10.9% fewer high-threshold stages vs matched MFRL

103.4 → 92.1; 6 grounded trajectories / stage.
Shared ≤3072-step budget; *calibration reuses collected data*.

5. Physical Robot Evaluation

Fixed-data adaptation: 50 demonstrations / task.
20 closed-loop tests / task; *no extra physical training rollouts*.

MERL-Full: 80.0% (32/40); 95% CI [65.2, 89.5]



MERL-Full closed-loop execution examples; video playback at 2× speed.

* Correspondence: zhangrm27@mail.sysu.edu.cn

Code: github.com/li-wenzhan/MERL