DRILL: Training World Models to Improve Policies, Not Predict Pixels

Zhaolu Kang[1,2], Tailong Luo, Lei Wei[2], Guangyuan Dong, Sihan Liu[2], Yingjie He[2], Siheng Wang, Shuaibo Li, Rongchao Zhang, Zhen Tian, Richeng Xuan[1†], Zhichao Hu[1†]

[1]Hunyuan, Tencent [2]Peking University †Corresponding author

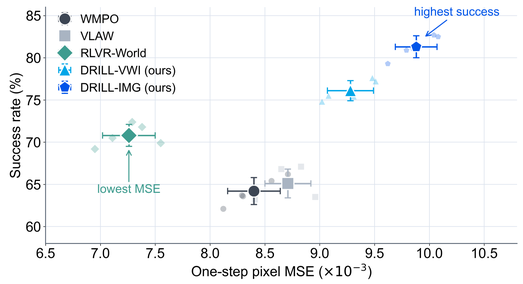
Paper
Abstract & paper

Email
zlkang25@stu.pku.edu.cn

TL;DR **World models for VLA reinforcement learning are trained to predict pixels, not to improve policies. DRILL trains them by the policy's downstream return: +13.1 success points over WMPO despite 17% higher pixel MSE.**

1 The mismatch

- RL post-training of **VLA policies** can need millions of environment steps, so imagined rollouts from **video world models** replace them.
- The standard recipe (WMPO) alternates two steps: fit the world model to demonstration pixels, then run GRPO inside it with a frozen vision-language reward model.
- World models are trained to **predict the next frame**, but their real job is to produce **learning signals** for the policy.
- Pixel MSE weights all pixels equally; policy gradients care about **visited states** and **high-advantage actions**.

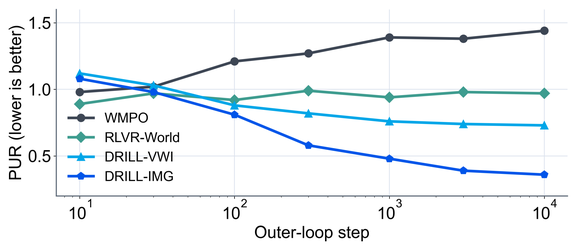


LIBERO-Long: lower pixel MSE does not mean higher success; DRILL sits top right.

- **Evidence**: lower pixel MSE can produce worse policy gradients.

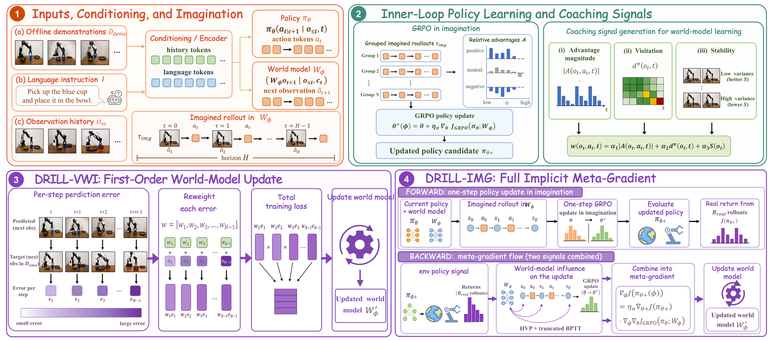
2 A diagnostic: Policy-Utility Residual

- **PUR** measures the angular gap between the pixel-MSE gradient and the policy-improvement direction.
- **Method-agnostic**: computed for every world-model objective we compare, not only DRILL.
- Lower PUR tracks higher success: WMPO 1.42, RLVR-World 0.97, DRILL-VWI 0.73, DRILL-IMG **0.36**.



LIBERO-Long: DRILL drives PUR down during training; WMPO's rises.

3 DRILL: a bilevel imagination loop



Inner loop: GRPO in imagination. Outer loop: update the world model so the updated policy earns more real return.

$$\phi^* = \arg\max_{\phi} J_{(\pi_{\theta^+(\phi)})}, \qquad \theta^+(\phi) = \theta + \eta \nabla_\theta \hat{J}_{\mathrm{GRPO}}(\pi_\theta; W_\phi)$$

J is the return in the target simulator: the world model is never asked to grade itself.

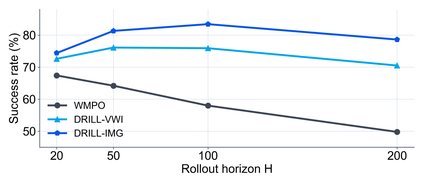
$$w(o_t, a_t) = \hat{d}(o_t) \cdot \mathrm{sgn}(\hat{A}) \, |\hat{A}|^{\alpha} \cdot s(o_t)$$

DRILL-VWI weight on each pixel error.

- Visitation, signed GRPO advantage and confidence all come from the imagined loop; weights are stopped and clipped.

- **DRILL-VWI** swaps pixel MSE for this weighted error: **no extra simulator interaction**.
- **DRILL-IMG** differentiates through the inner GRPO step, with 30 simulator rollouts per outer step and 2-step truncated BPTT.
- **Theory**: VWI is the first-order part of the IMG meta-gradient; IMG converges.

4 Which variant, and how robust?



LIBERO-Long: the IMG–VWI gap widens with horizon H.

- **DRILL-VWI**: drop-in replacement for pixel MSE, no extra simulator cost.
- **DRILL-IMG** pays off on long horizons: 83.4 vs. 75.9 at H = 100.
- On short-horizon SimplerEnv the two tie (64.0 vs. 63.1).

Method	Novel obj.	Distractors	Light/texture	Pose	Avg.
WMPO	37.4	25.3	30.4	27.6	30.2
RLVR-World	31.8	20.4	26.0	24.1	25.6
DRILL-VWI	26.4	17.1	23.5	**18.0**	21.3
DRILL-IMG	**22.5**	**15.3**	**17.6**	**18.9**	**18.6**

Zero-shot out-of-distribution relative success drop on LIBERO (%, lower is better).

5 Results (π_0-3B, 5 seeds)

+13.1 success points vs. WMPO

+7.9 points vs. RLVR-World

+17% pixel MSE, yet better policies

Method	LIBERO	CALVIN	RoboCasa	MT50	Simpler	MSE	PUR
Scratch-GRPO	18.2	22.5	8.1	21.4	15.0	–	–
Sim-PPO	28.5	32.1	12.4	31.6	22.8	–	–
WMPO	64.2	71.8	38.5	67.4	55.1	8.4	1.42
VLAW	65.1	74.0	39.6	68.2	57.9	8.7	1.37
RLVR-World	70.8	75.3	42.7	72.9	61.6	**7.2**	0.97
DRILL-VWI	**76.1**	**80.7**	**51.8**	**75.6**	**64.0**	**9.3**	**0.73**
DRILL-IMG	**81.3**	**83.1**	**55.7**	**79.4**	**63.1**	**9.8**	**0.36**
IMG + RLVR WM	**83.9**	**85.4**	**59.1**	**81.8**	**67.3**	**8.6**	**0.29**

Success rate (%); MSE × 10³. IMG + RLVR WM: DRILL-IMG on an RLVR-World-pretrained world model.

- **Sim-PPO** gets DRILL-IMG's simulator episodes but no imagination and reaches only 28.5 on LIBERO-Long: the world model amplifies scarce interaction.

Method	Steps	SR
WMPO	10k	64.2
WMPO, 1.4×	14k	65.1
WMPO, 2.0×	20k	64.4
DRILL-IMG	**10k**	**81.3**

LIBERO-Long; extra WMPO updates.

- **Not just compute**: 2× more WMPO updates move it only from 64.2 to 64.4.
- **Setup**: LIBERO-Long, CALVIN ABC→D, RoboCasa, MT50, SimplerEnv; Welch's t-tests; about 17k H20 GPU-hours.

DRILL-IMG vs. baseline	DRILL	Tie	Base.	p
WMPO, visual quality	33%	26%	41%	0.29
RLVR-World, visual quality	30%	28%	42%	0.17
WMPO, task completion	**62%**	23%	15%	<0.001

Forced choice, 6 raters, 120 trials per row.

- Raters cannot tell the rollouts apart by look, yet prefer DRILL once task completion is shown.

6 Ablations and transfer

World-model objective	SR	PUR
WMPO (pixel MSE)	58.2	1.42
VA-WM (value-aware)	62.4	1.10
Visitation only	64.2	1.18
Signed advantage only	66.8	1.06
Confidence only	60.1	1.31
DRILL-VWI (all three)	**69.2**	**0.73**
VWI, α = 2	67.9	0.76
DRILL-IMG	**73.4**	**0.36**

Mean SR on LIBERO, CALVIN and RoboCasa.

Method	LIBERO	CALVIN
WMPO	59.7	66.9
RLVR-World	64.9	70.6
DRILL-VWI	72.1	75.8
DRILL-IMG	**77.1**	**78.9**

OpenVLA-7B backbone, SR (%).

- Margin over WMPO grows with world-model size: +7.9 (300M), +13.1 (1.2B), +16.4 (7B).
- Swapping the reward model keeps the gain: +15.6 to +18.4 over WMPO.
- **Limitation**: simulators only, no real-robot results.