



RLVR-World: Training World Models with Reinforcement Learning

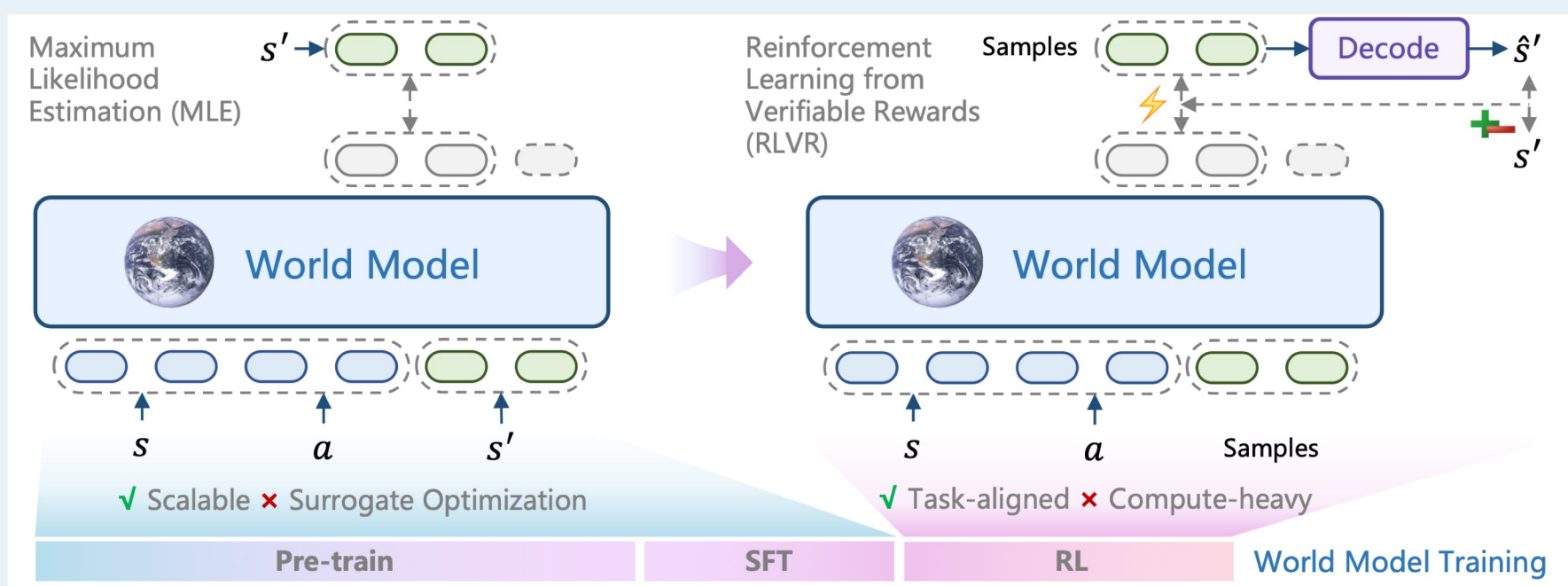


Jialong Wu, Shaofeng Yin, Ningya Feng, Mingsheng Long

TL;DR

Key Insight

As world models are built with more advanced non-end-to-end models (*discrete autoregressive models or diffusion models*), their training methods typically optimizing for **likelihood** are **misaligned** with the actual task objective of world modeling.



Contributions

RLVR as a direct and effective post-training method:

Task-specific prediction metrics serve as verifiable rewards

World models across modalities (language and video):

Unified under a sequence modeling formulation

RL-trained world models enhance decision-making:

Policy evaluation and model predictive control

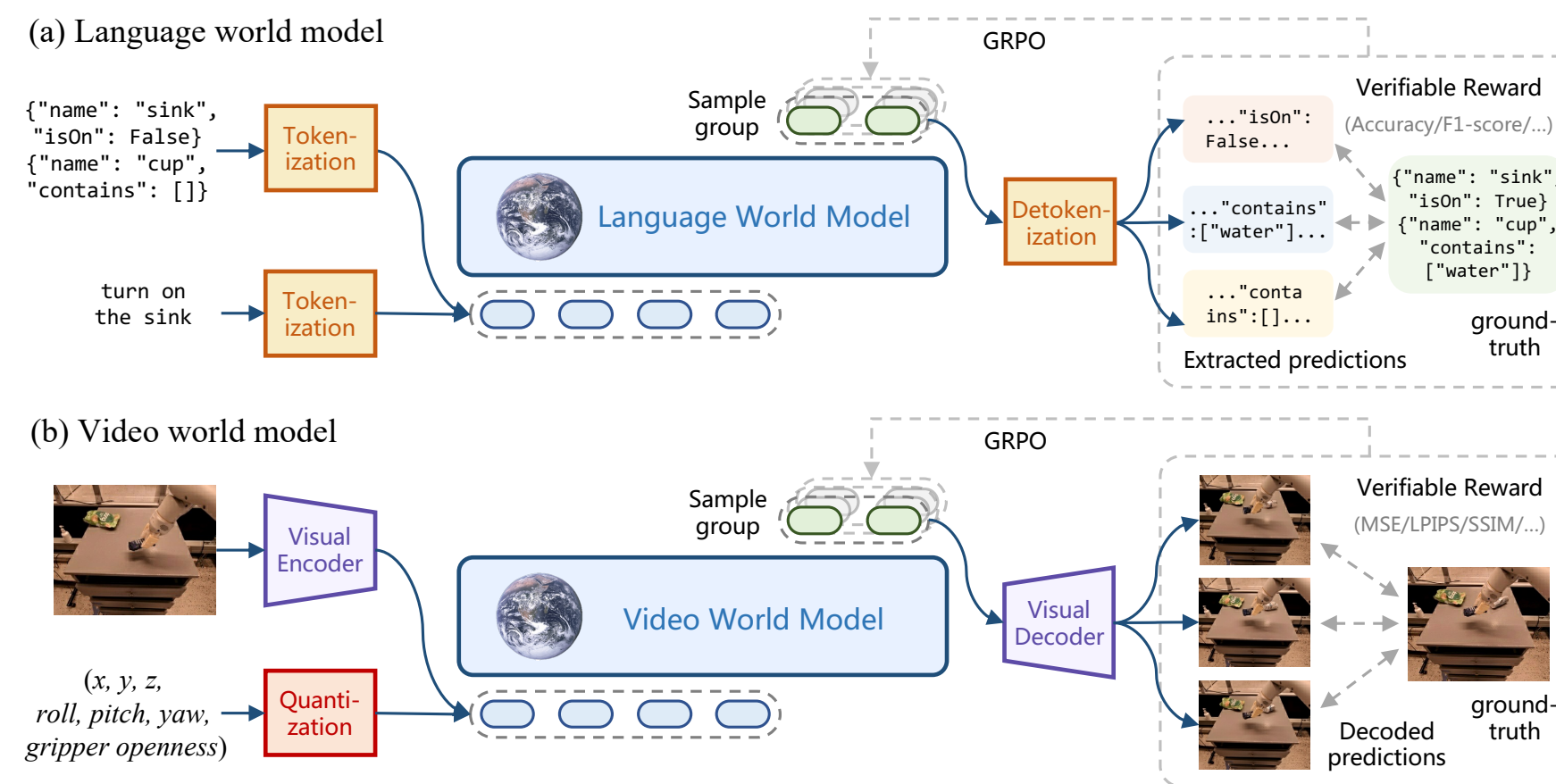
Project Website: thuml.github.io/RLVR-World

Code: github.com/thuml/RLVR-World

Datasets & Models: [hf.co/collections/thuml/rlvr-world](https://huggingface.co/collections/thuml/rlvr-world)



RLVR-Framework



Unified Sequence Modeling Formulation (Diffusion in future work!)

$$\mathcal{J}_{MLE}(\theta) = \log p_{\theta}(o(s') | q(s, a)) = \sum_{t=1}^{|o(s')|} \log p_{\theta}(o_t(s') | q(s, a), o_{<t}(s'))$$

Task-Specific Prediction Metrics as Rewards

GRPO with $R_i = \text{sign}(D) \cdot D(\hat{s}'_i, s')$

$$\text{sign}(D) = -1$$

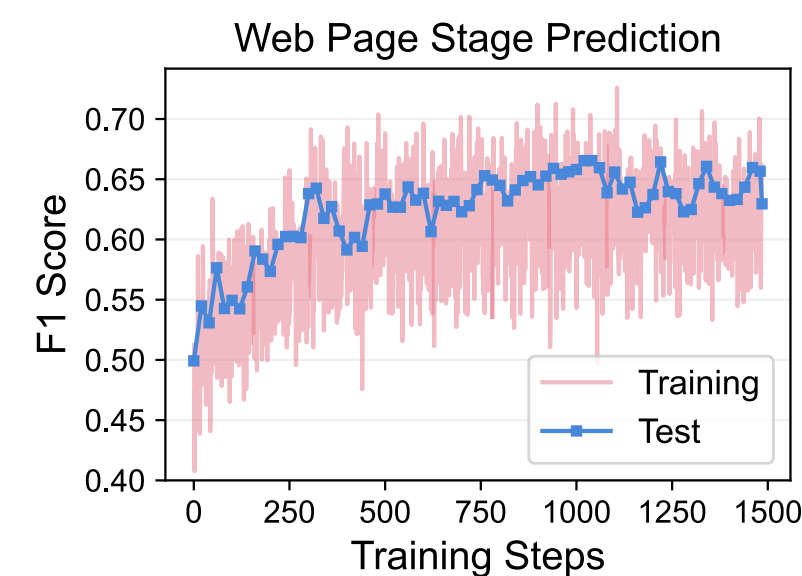
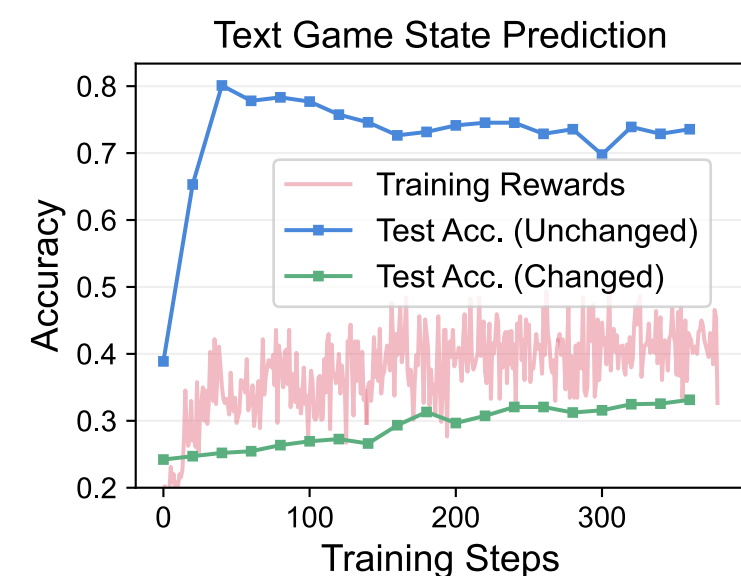
If Lower D is better

$$\text{sign}(D) = 1$$

Otherwise

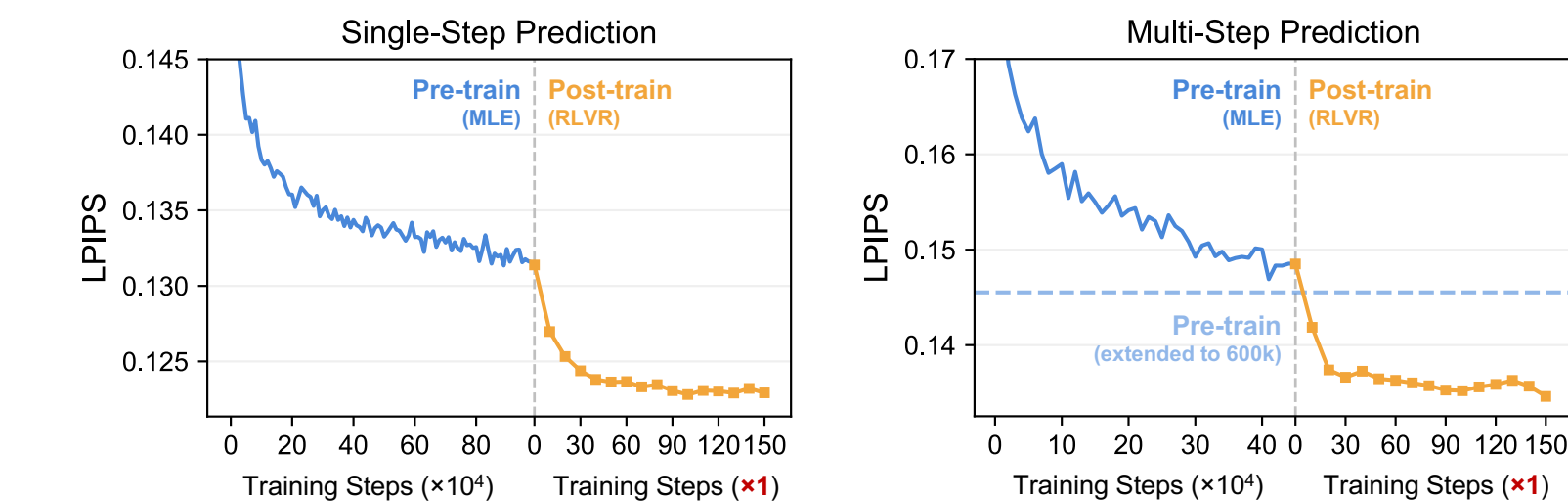
Language World Models

Beyond its success in math and coding, RLVR can also improve LLMs' performance on world modeling tasks involving verbal state transitions.

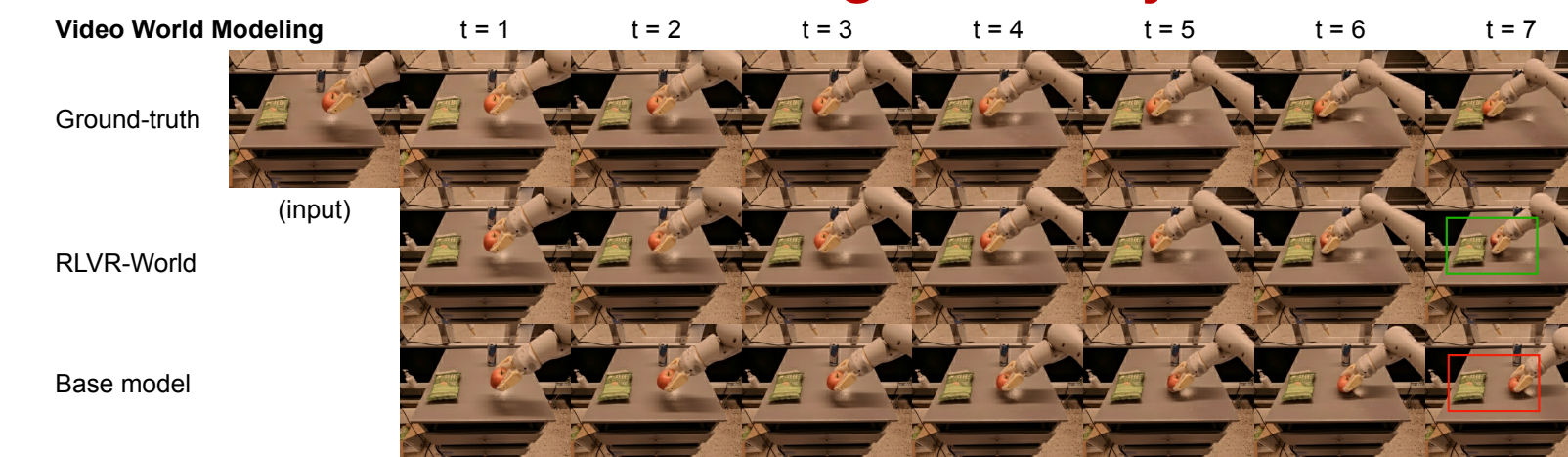


Video World Models

RLVR bridges the gap between pre-training objectives and visual prediction metrics, leading to more accurate predictions, improved training efficiency, and reduced artifacts such as repetition.

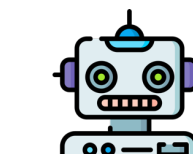


1000x Training Efficiency!



World Model Applications

RLVR-trained world models can improve downstream tasks, including policy evaluation and model predictive control.



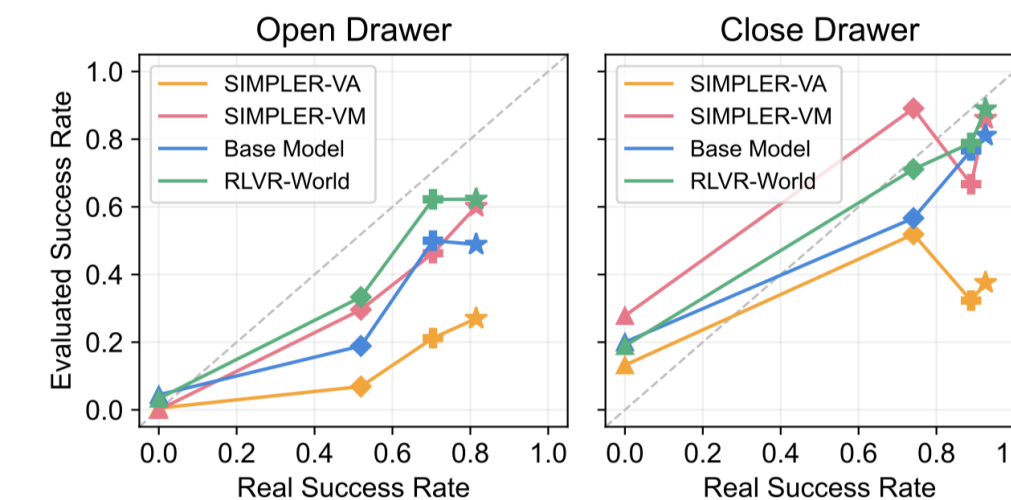
Real2Sim Evaluation of RT-1 Checkpoints

Success Rate:

SFT: 12.06%

RLVR: 14.29%

(relatively +18.4%)



First work on world model-based robot policy evaluation!