Preference Goal Tuning: Post-Training as Latent Control for Frozen Policies

Guangyu Zhao[1], Kewei Lian[2], Haoxuan Ru[1], Borong Zhang[1], Haowei Lin[1], Zhancun Mu[1], Haobo Fu[3], Qiang Fu[3], Shaofei Cai[1], Zihao Wang[1], Yitao Liang[1]

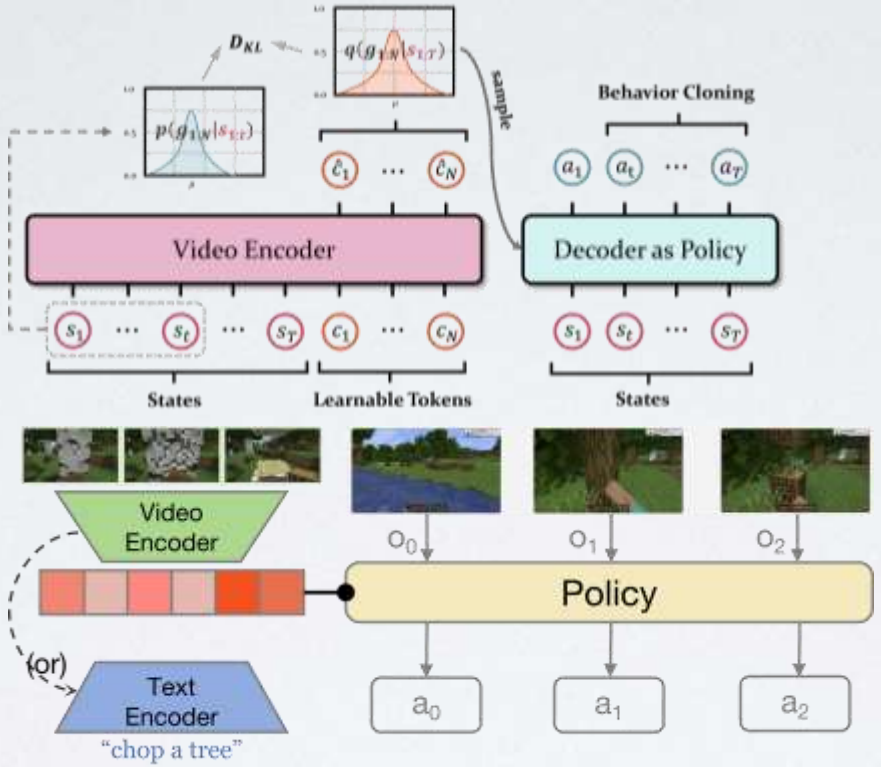
[1] Institute for Artificial Intelligence, Peking University [2] School of Computing, National University of Singapore [3] Tencent AI Lab

Motivation

- Sequential decision models: **sensitive** to “prompts”.
  - Language
  - Image
  - Video
  - ...
- Prompts: modulate the behavior of a frozen policy.
- Bypass prompts?
- Post-train the “prompt-policy” interface.

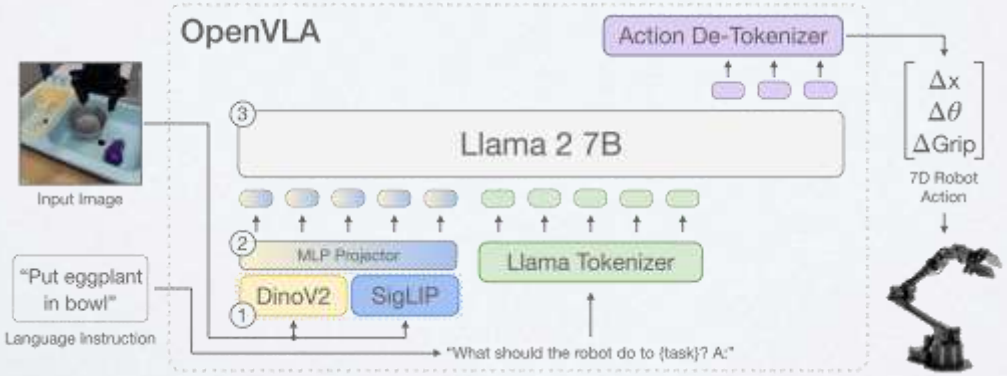
Problem Formulation

Conditional Variational Autoencoder (CVAE)



Goal-embedding modulation

Vision-language-action models

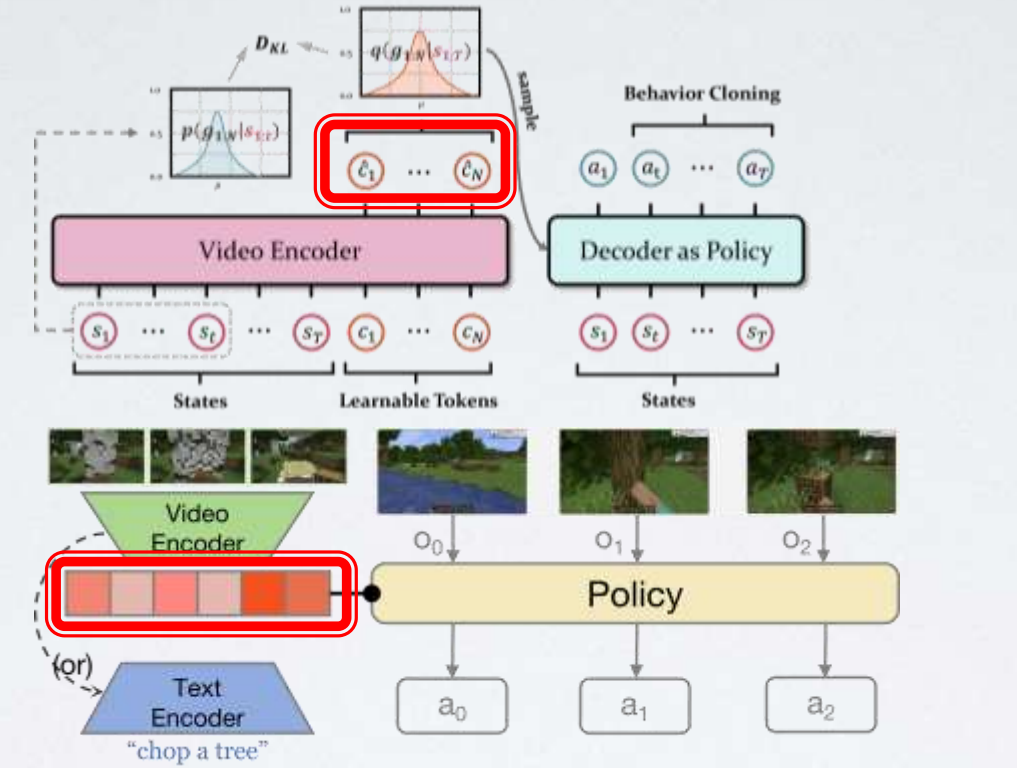


Problem Formulation

Conditional Variational Autoencoder (CVAE)

Goal-embedding modulation

Vision-language-action models



latent goal

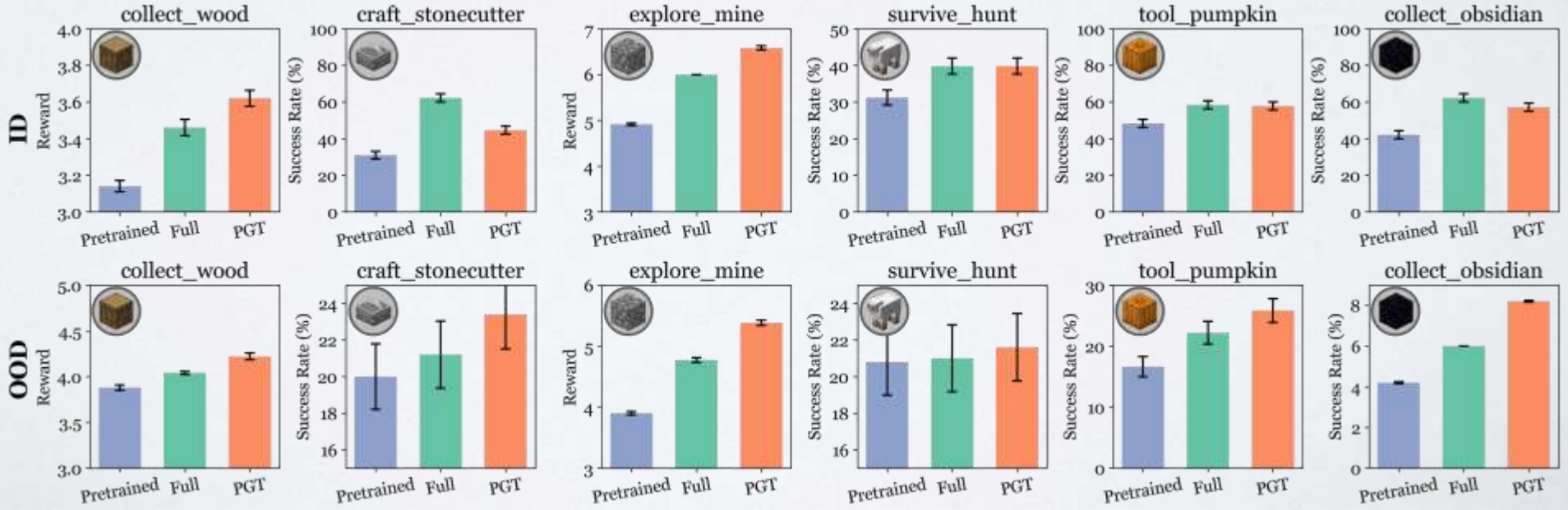
as a control interface

The conditional representation through which a policy selects and modulates behavior.

Design Choices

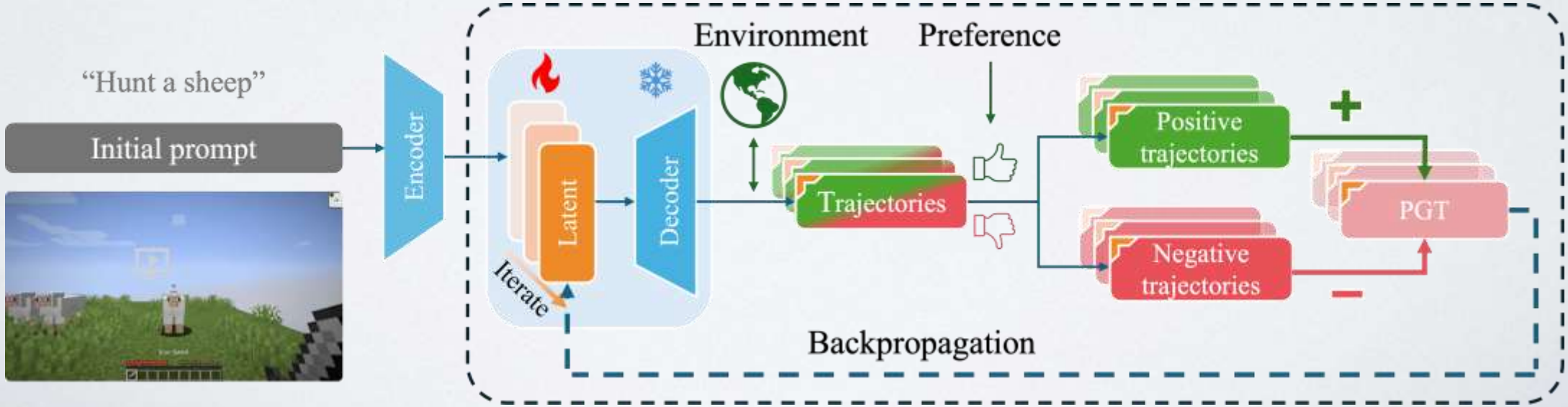
- Preferences, not only positives.
- Latent goal, not full policy.
- Preserve the frozen policy's generalization.

Task	Latent-goal-only			Full Fine-Tuning		
	Pretrained	BC	DPO	Pretrained	BC	DPO
	3.14	3.28	**3.62**	3.14	3.26	**3.46**
	42.0	18.2	**57.2**	42.0	15.0	**62.2**
	4.91	4.76	**6.58**	4.91	4.80	**6.00**
	48.3	45.4	**57.8**	48.3	48.6	**58.4**



Methodology: Preference Goal Tuning

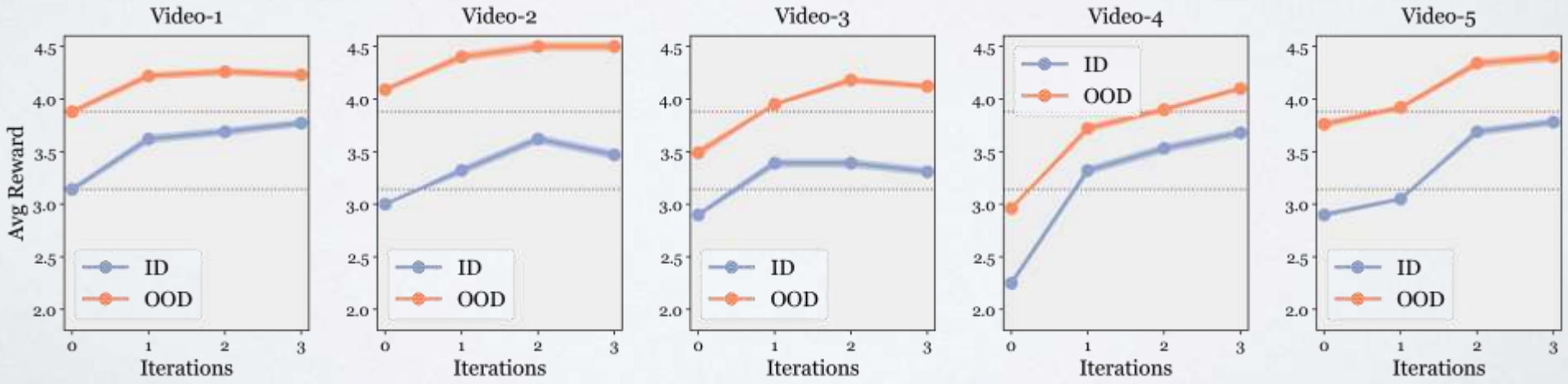
- Start from an initial prompt.
- Roll out the frozen policy.
- Rank trajectories by preference.
- Backpropagate only to the latent goal.



Experiment: Beyond Prompt Engineering

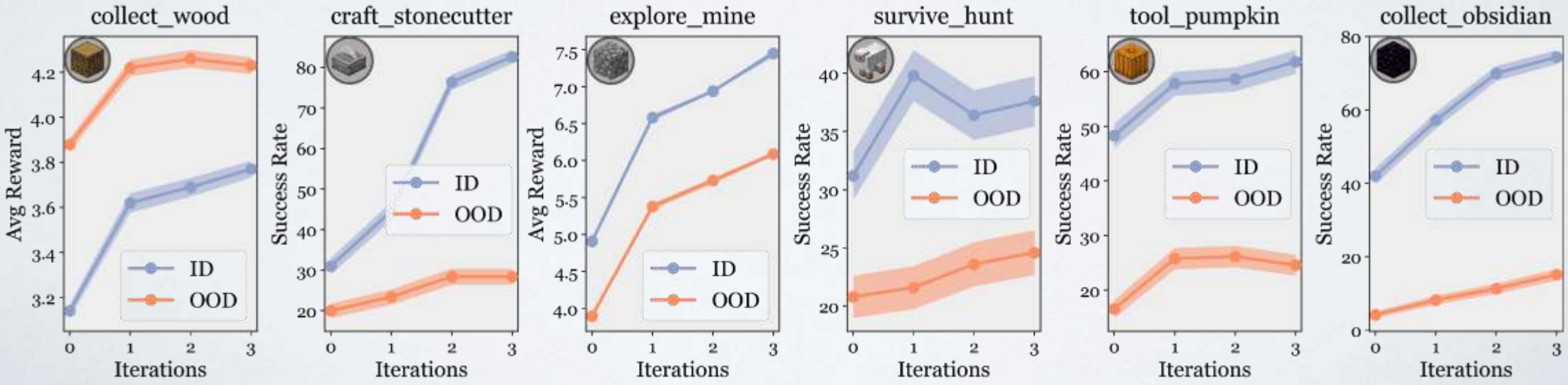
- 2 foundation policies.
- 17 Minecraft tasks.
- ID and OOD evaluation.
- Consistent gains beyond prompt search.

Task	In Distribution						Out of Distribution					
	GROOT	GROOT+	Δ	STEVE	STEVE+	Δ	GROOT	GROOT+	Δ	STEVE	STEVE+	Δ
	3.14	**3.62**	15.3%	3.73	**3.90**	4.6%	3.88	**4.22**	8.8%	4.22	**4.29**	1.7%
	27.0	**62.8**	132.6%	16.3	**36.4**	123.3%	15.4	**54.6**	254.5%	30.4	**48.0**	57.9%
	30.4	**40.8**	34.2%	43.3	**56.6**	30.7%	34.0	**41.6**	22.4%	45.6	**60.2**	32.0%
	20.2	**20.8**	3.0%	4.2	**21.8**	419.0%	7.8	**9.4**	20.5%	41.4	**49.0**	18.4%
	31.0	**44.6**	43.9%	14.1	**19.0**	34.8%	20.0	**23.4**	17.0%	36.2	**48.4**	33.7%
	• • •	(17 tasks)										



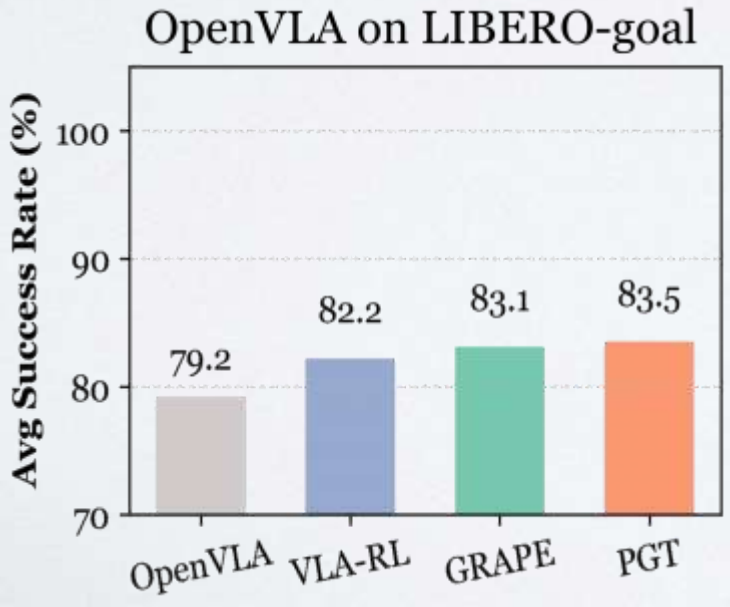
Experiment: Iterative Training

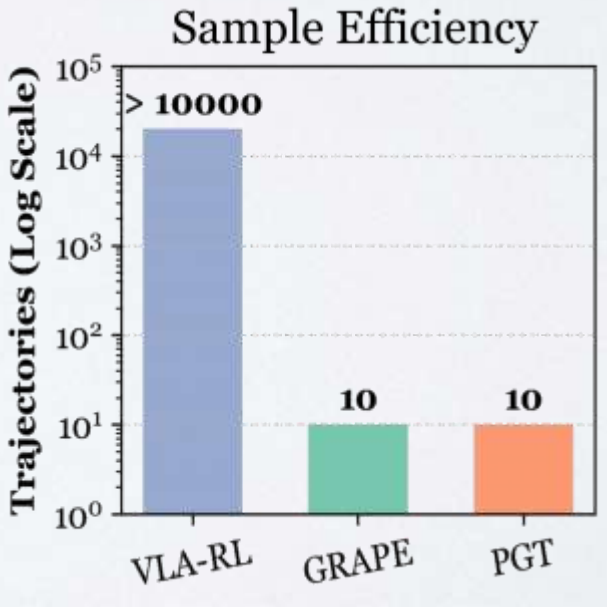
- Updated goals collect better rollouts.
- New rollouts provide new preferences.
- PGT forms a lightweight data flywheel.
- Performance improves across rounds.



Experiment: Cross-domain

- Same principle beyond Minecraft.
- OpenVLA on LIBERO-goal.
- Token embeddings as the interface.
- Improve behavior without changing the backbone.





Thank You For Your Attention!