VINROBOTICS

VINUNIVERSITY

dfki ai Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

imprs-is
International Max Planck Research School for Intelligent Systems

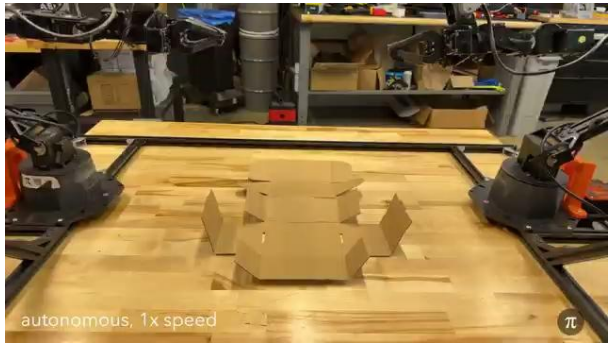
FOCA: Future-Oriented Conditioning for Data-Efficient Vision-Language-Action Adaptation

Duc Minh Nguyen*, Nghiem Tuong Diep*, Binh Gia Nguyen*, et al.

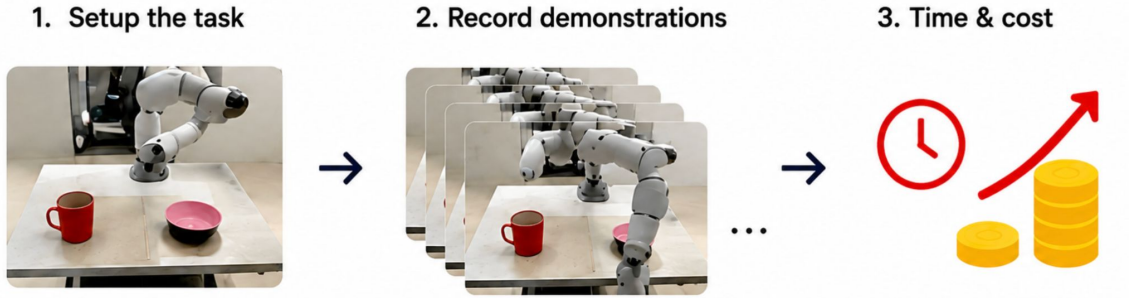
1. Motivation

Large-scale pretraining

Pi-0, Gr00t, OpenVLA

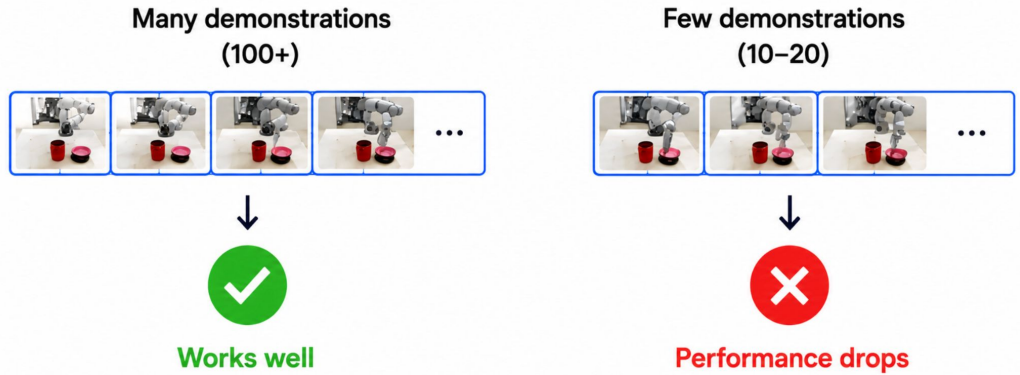


In real-world deployment, collecting robot demonstrations remains expensive and time-consuming.



1. Motivation

Can pretrained VLA models adapt effectively with only **10 to 20 demonstrations?**



Unfortunately, the answer is often no.

2. Stress Test of Existing VLA Models

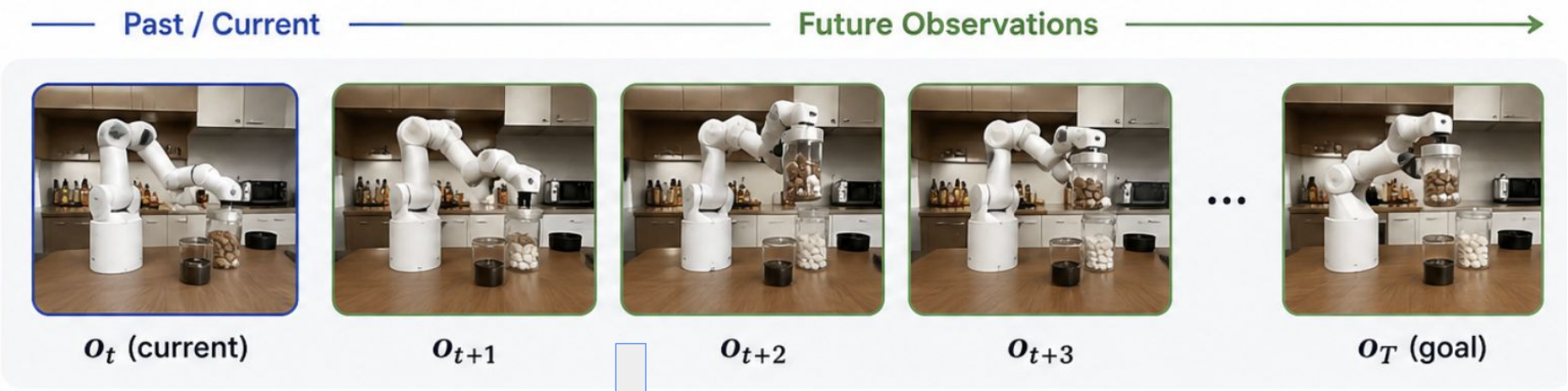
(a) Sensitivity to few-shot imitation learning rate

Method	Demonstration Ratio											
	100%				40%				10%			
	Avg	10	Obj.	Spa.	Avg	10	Obj.	Spa.	Avg	10	Obj.	Spa.
π_0	94.6	90.0	98.2	94.6	89.9	82.0	95.2	89.6	77.6	59.0	80.6	83.4
Groot-N1.5	94.6	**92.8**	98.4	94.4	91.4	84.5	98.9	90.6	78.2	62.6	85.7	85.7
EO-1	94.1	91.4	96.6	89.8	91.0	**88.4**	96.0	86.8	82.2	65.0	89.6	83.0
SmolVLA	92.5	82	99	93	90.3	80	96.0	90.0	77.3	51.3	86	81.0
FOCA	**96.6**	92.4	**99.8**	**97.0**	**94.0**	88.0	**99.6**	**93.6**	**85.3**	**69.4**	**90.4**	**89.4**

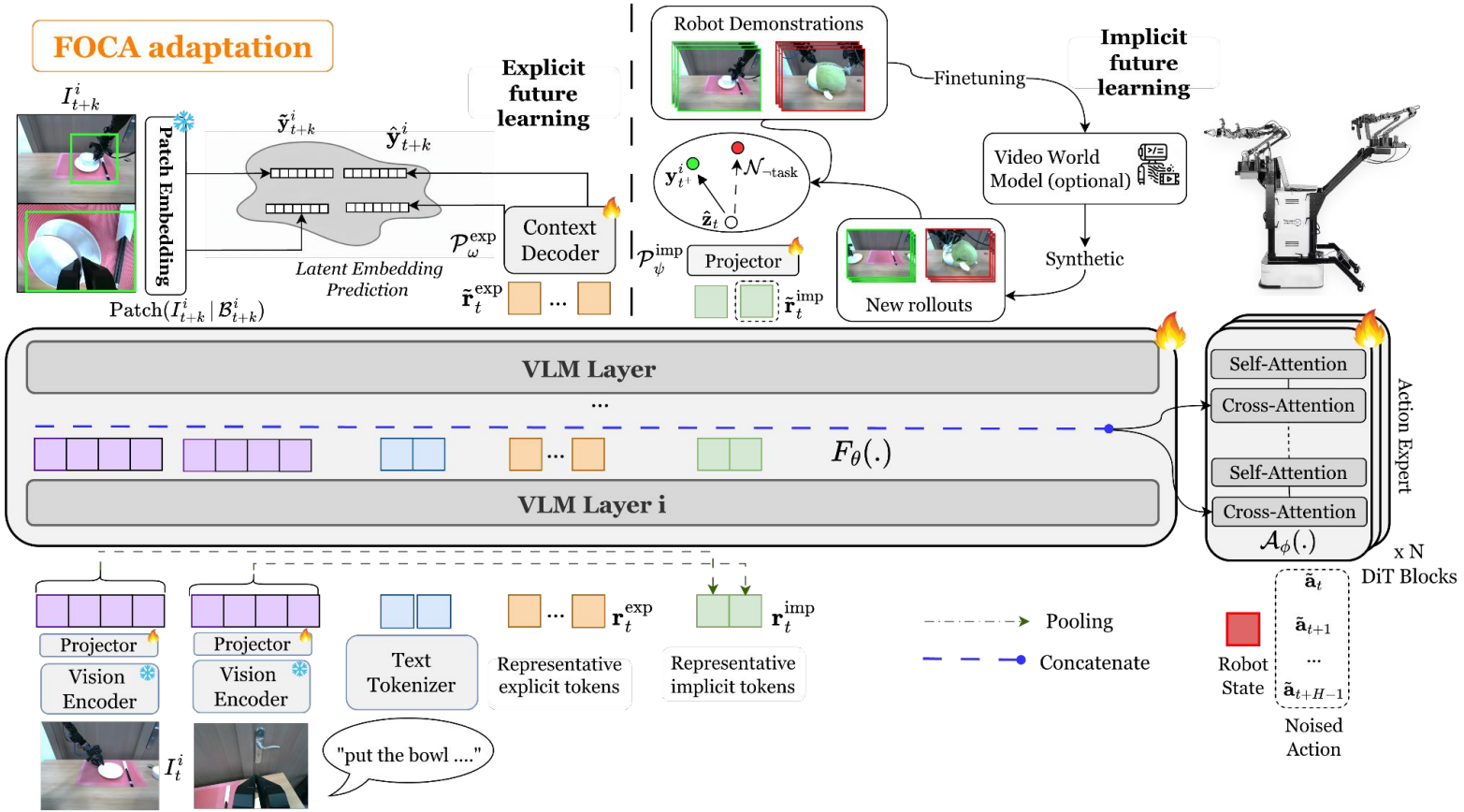
Question: What information are we failing to exploit during adaptation?

3. Key Observation

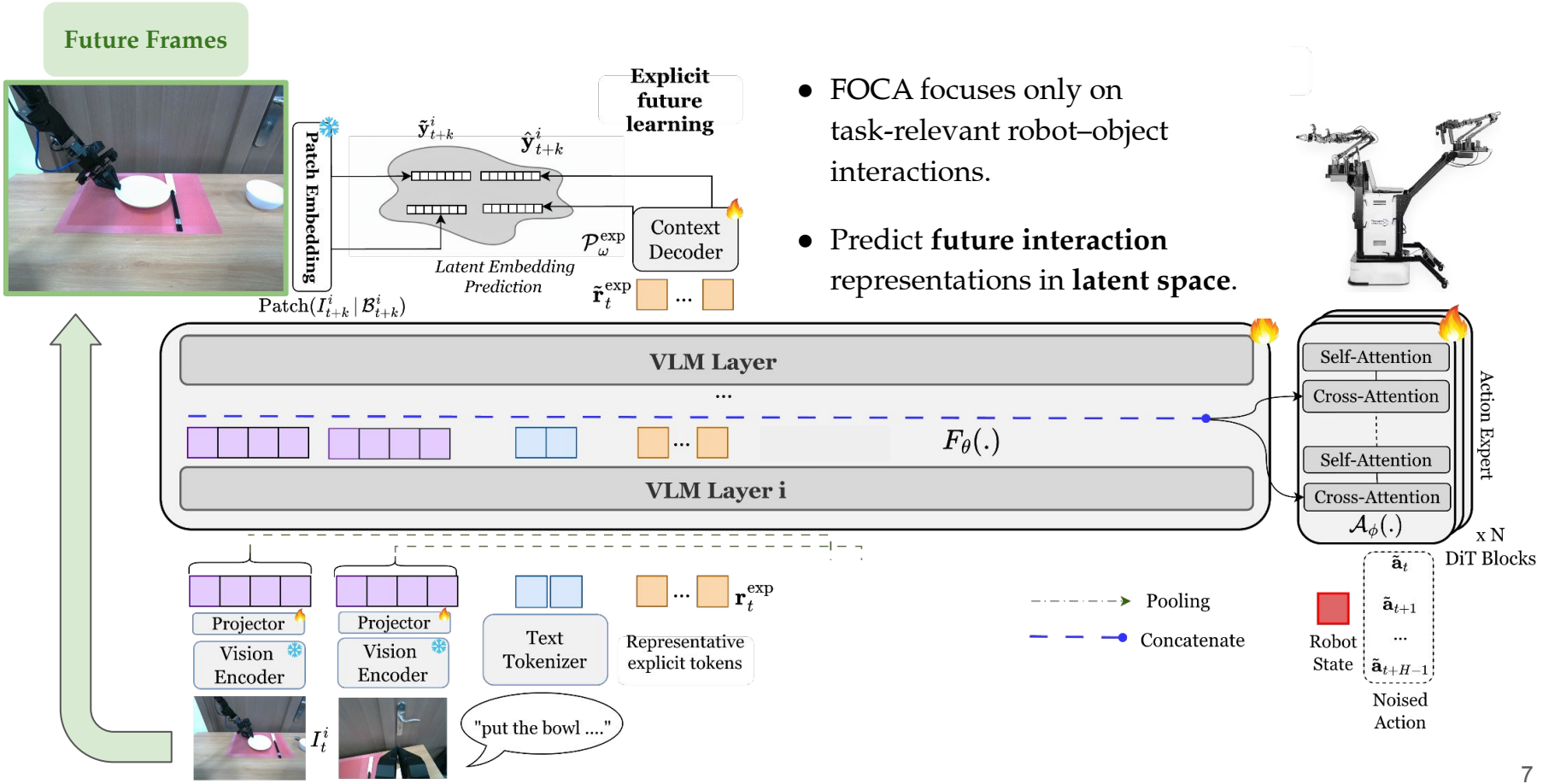
A demonstration contains much more information than just action labels, one of them is future information.



4. FOCA: Future-Oriented Conditioning for VLA Adaptation

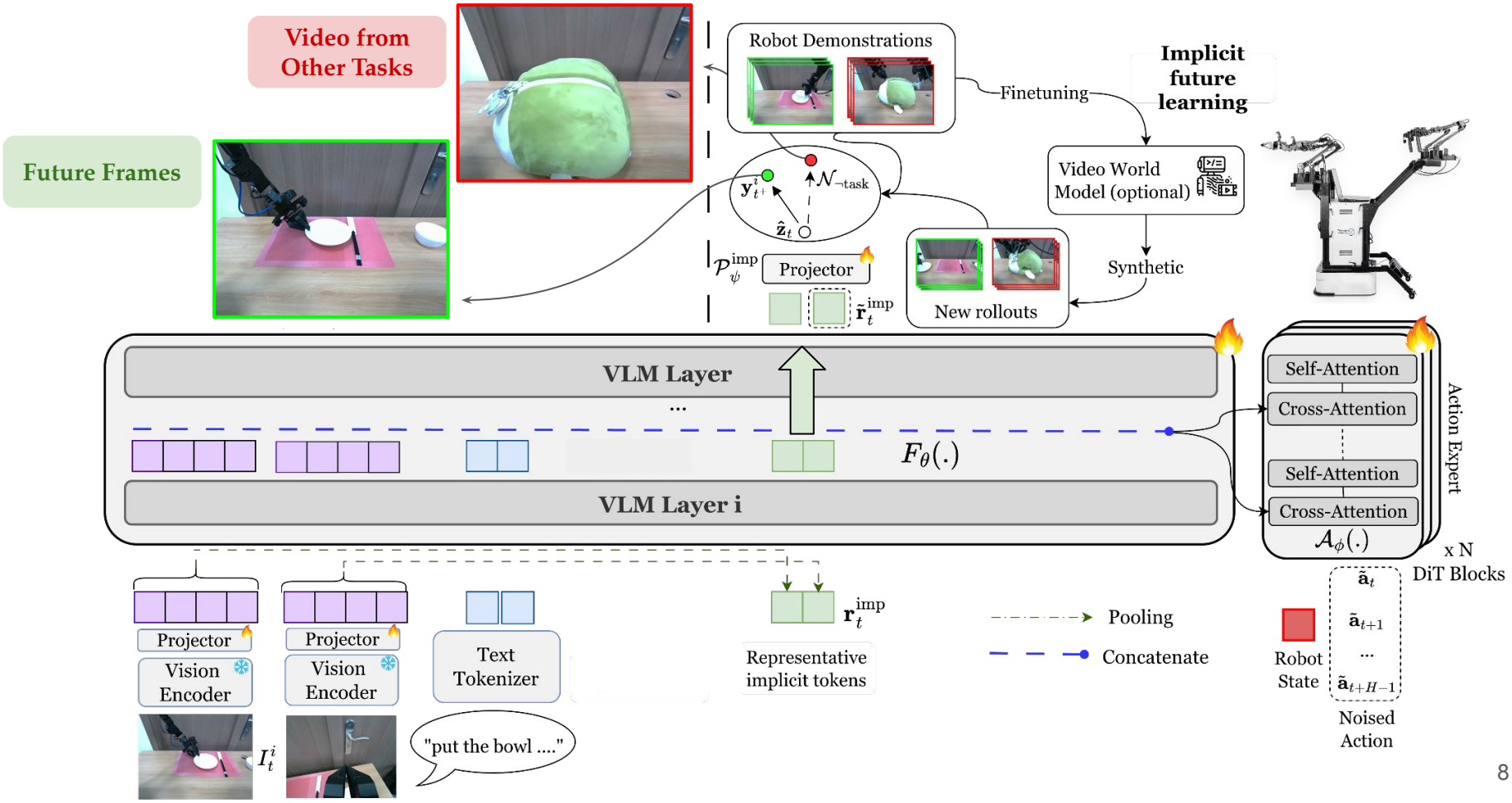


4. FOCA: Future-Oriented Conditioning for VLA Adaptation



- FOCA focuses only on task-relevant robot–object interactions.
- Predict **future interaction** representations in **latent space**.

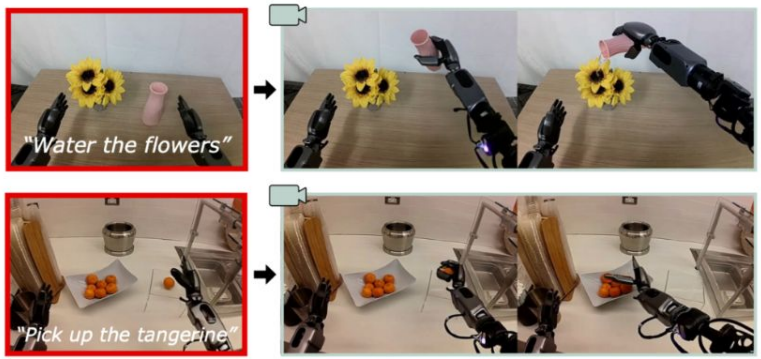
4. FOCA: Future-Oriented Conditioning for VLA Adaptation



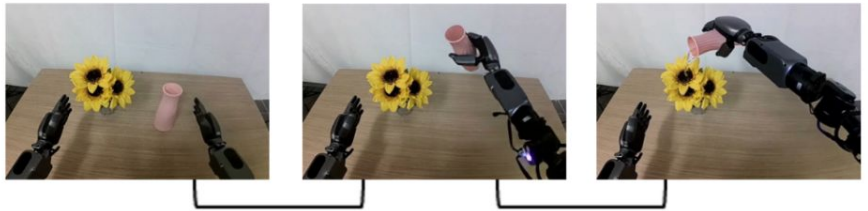
5. Learning from Synthetic Videos

Video World Model

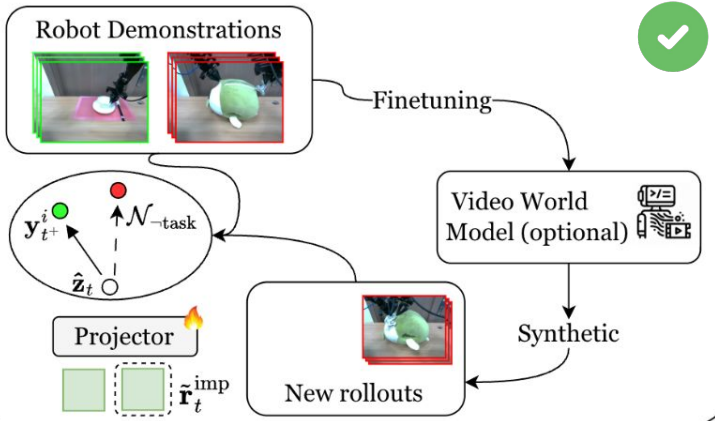
1. Rollout



2. Existing: Extracting pseudo-actions by IDM/LAPA



2. FOCA: Implicit Future Alignment (No need pseudo-actions)



3. VLA Adaptation

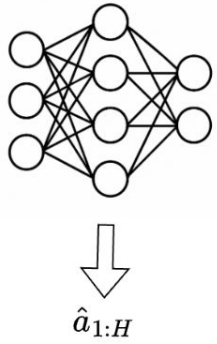


Table 2. Performance comparison between FOCA variants and pseudo-actions learned via inverse generative modeling (IGM) from DreamGen-generated synthetic videos.

Method	Demonstration Ratio														
	100%					40%					10%				
	Avg	10	Go.	Obj.	Spa.	Avg	10	Go.	Obj.	Spa.	Avg	10	Go.	Obj.	Spa.
π_0	94.6	90.0	95.4	98.2	94.6	89.9	82.0	92.6	95.2	89.6	77.6	59.0	87.4	80.6	83.4
IGM	94.3	90.0	95.2	98.2	93.6	90.2	82.8	95.6	94.8	87.8	76.8	61.8	84.0	79.8	81.4
FOCA(Imp.)	95.8	91.0	97.2	98.4	96.6	93.0	87.0	94.4	97.6	92.9	83.6	65.6	91.0	90.0	87.6
FOCA(D.G.)	**96.7**	**92.6**	**96.8**	**99.2**	**98.0**	**95.7**	**89.4**	**96.6**	**98.6**	**98.0**	**86.4**	**73.2**	**91.4**	**91.8**	**89.2**

- Directly use videos
- Simpler pipeline
- No pseudo-action extraction
- Avoid noise from imperfect actions

5. Reinforcement learning perspective

Current State

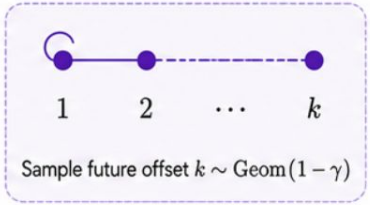
s_t

Robot observes the current state.

Contrastive Learning

+

Geometric Future Sampling



Theorem 4.1

The contrastive objective estimates the goal-conditioned discounted occupancy:

$$\rho^{\mu}(g \mid x_t) := (1-\gamma) \sum_{k=1}^{\infty} \gamma^{k-1} \Pr(g_{t+k} = g \mid x_t)$$

h is the goal extractor.
x_t is the current state.

Future Goal

g_{t+k}

Learned Similarity Score

$$f(s,g) \approx$$

Value-like Reachability Function

Measures how likely a future goal representation is reachable from the current state.

Implicit alignment learns which future outcomes are achievable, providing value-like supervision for long-horizon reasoning.

6. Main Results on simulation benchmark: LIBERO and Robocasa

- **FOCA** consistently **outperforms** existing **VLA baselines** at **100% data** (**96.6** Avg.).
- Remarkably, **FOCA** trained with only **40% demonstrations** achieves **94.0** Avg., **close** to **full-data performance** of baselines.

➔ These results highlight FOCA's strong **data efficiency** and **robustness.**

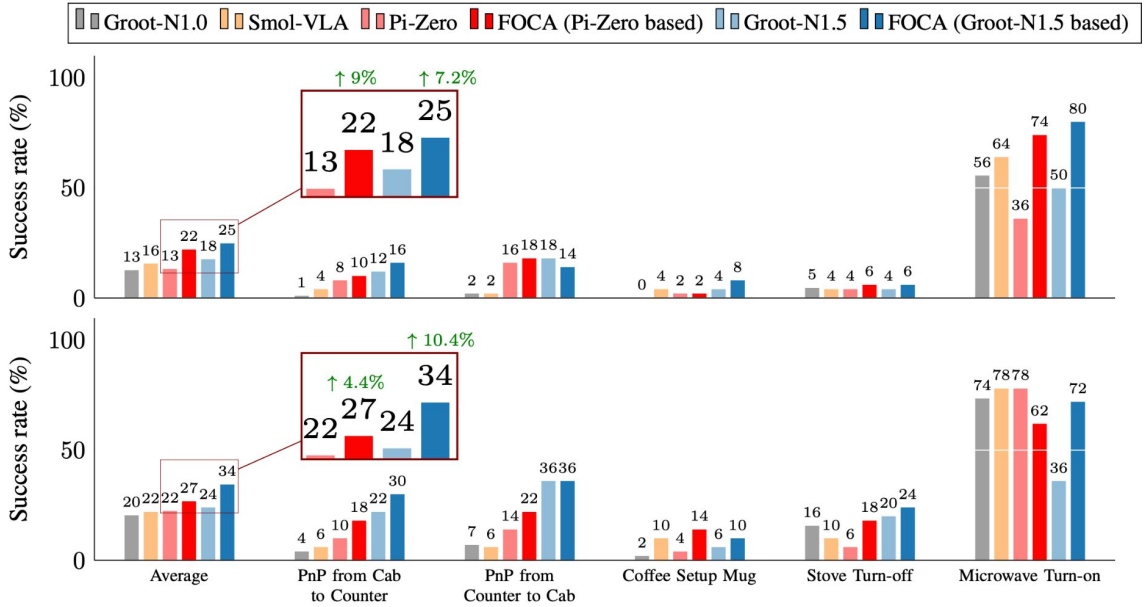
(a) Sensitivity to few-shot imitation learning rate

Method	Demonstration Ratio											
	100%				40%				10%			
	Avg	10	Obj.	Spa.	Avg	10	Obj.	Spa.	Avg	10	Obj.	Spa.
π_0	94.6	90.0	98.2	94.6	89.9	82.0	95.2	89.6	77.6	59.0	80.6	83.4
Groot-N1.5	94.6	**92.8**	98.4	94.4	91.4	84.5	98.9	90.6	78.2	62.6	85.7	85.7
EO-1	94.1	91.4	96.6	89.8	91.0	**88.4**	96.0	86.8	82.2	65.0	89.6	83.0
SmolVLA	92.5	82	99	93	90.3	80	96.0	90.0	77.3	51.3	86	81.0
FOCA	**96.6**	92.4	**99.8**	**97.0**	**94.0**	88.0	**99.6**	**93.6**	**85.3**	**69.4**	**90.4**	**89.4**

Pi0 original on LIBERO

Pi0 + Foca on LIBERO

6. Main Results on simulation benchmark: LIBERO and Robocasa



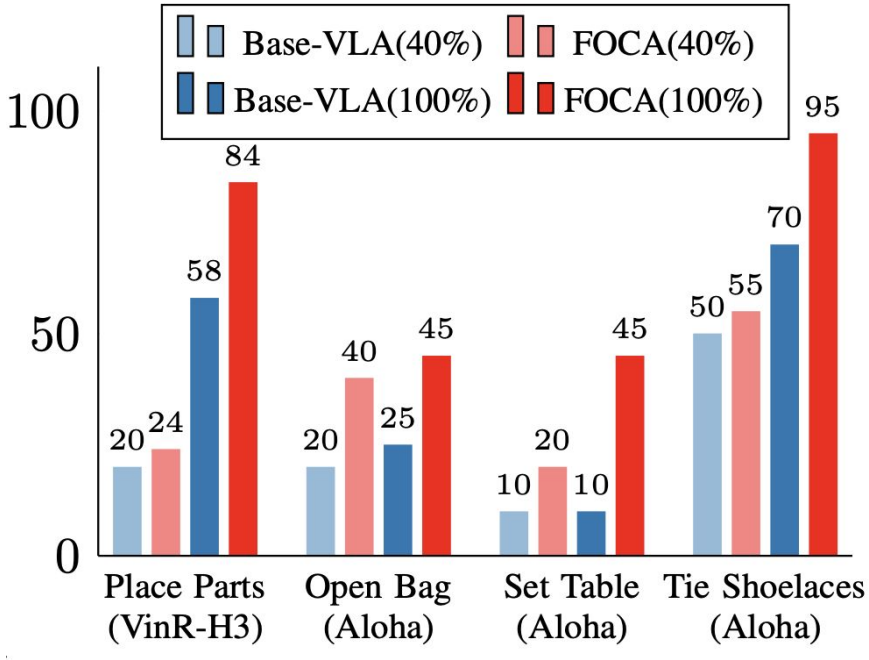
Groot-N1.5 + Foca on Robocasa

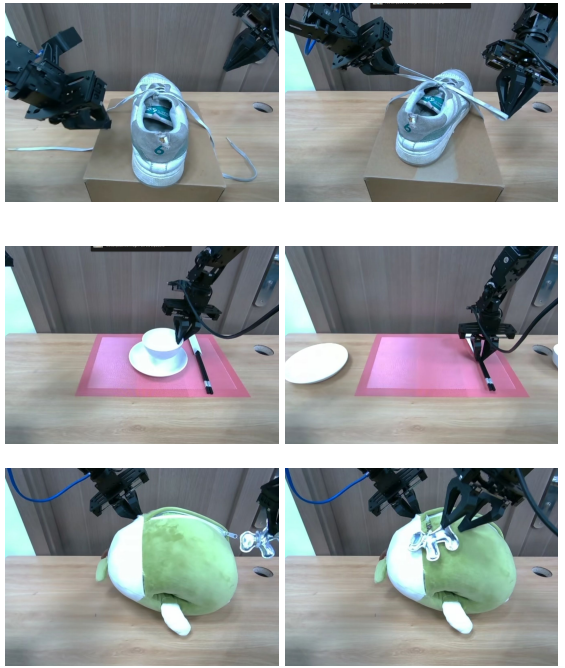
Figure 10. FOCA's generalizations performance across Pi-Zero and Groot N1.5 on Robocasa, with 30 demos (top) and 100 demos (bottom) on most 5 challenge tasks

- **FOCA** reliably **boosts downstream success rates** for both **Pi-Zero** and **Groot-N1.5** on **Robocasa**.
- **Improvements** reach **+7.2%** (30 demos) and **+10.4%** (100 demos) on **average challenge-task performance** with **Groot-N1.5**

9. Generalization Across Realworld Robot

FOCA consistently outperforms Base-VLA in both 40% and 100% data settings across all tasks: Tie Shoelaces, Set Table, Open Bag.





9. Conclusion

- **FOCA improves few-shot VLA adaptation** by leveraging richer supervision from future visual outcomes.
- **Combines explicit and implicit future supervision** in latent space for data-efficient learning.
- **Action-free adaptation** of pretrained VLA models without additional action labels.
- **Consistent gains across simulation and real robots** with strong cross-embodiment generalization.
- **Robust to imperfect synthetic rollouts**, enabling practical deployment.
- **Future-conditioned representation learning** is a promising direction for scalable robot adaptation.