Seer: Predictive Inverse Dynamics Models are Scalable Learners for Robotic Manipulation

World Model Seminar

Presented by Mostafa Kotb

http://www.informatik.uni-hamburg.de/WTM/

Motivation

Q: How to learn ***scalable*** and ***generalizable*** policies for robotic manipulation?

A: Utilize a large-scale pre-training phase

Action-centric Pre-training

- ***Objective:*** Behavior Cloning Policy from large-scale robotic datasets.
- ***Datasets:*** DROID / X-Embodiment ***(Robotic)***.
- ***Approaches:*** RT-X / Octo / OpenVLA.
- **Limitation:** Naive BC in both pre-train and fine-tune phases.

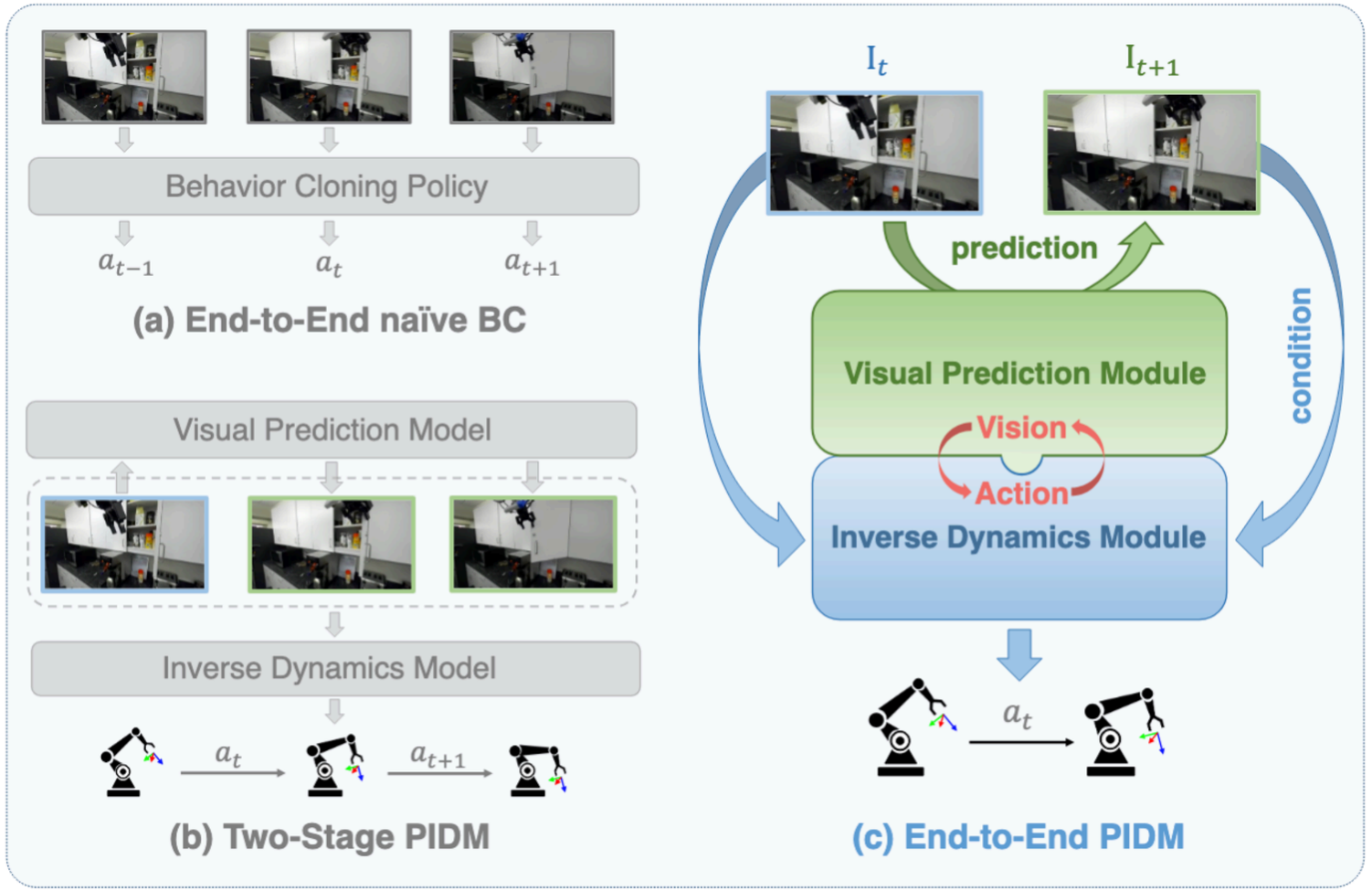
Vision-centric Pre-training

- ***Objective:*** Learn representations through discriminative or generative ways.
- ***Datasets:*** Ego4D ***(Videos)***.
- ***Approaches:*** R3M / MVP (Discriminative)
  UniPi / Susie (Generative)
- **Limitation:** Two stages; decouple vision and action.

KNOWLEDGE TECHNOLOGY

Motivation

- This work proposes that vision and action should be integrated together in a closed loop end-to-end manner.
- This integration is natural and necessary, as humans typically coordinate their hands and eyes to manipulate objects.
- Therefore, closing the loop during training and inference are both necessary for a better scalable action learner.
- ***Limitation:*** *rely only on expensive robot data during pre-training and fine-tuning.*



KNOWLEDGE TECHNOLOGY

Seer: End-to-End PIDM

- First, we explain the end-to-end training of vision and action.

Vision: Conditional Visual Foresight

- ***Seer takes as input:***
  - A goal *g* in the form of *language instructions* or *robot states*.
  - Historical observations *ht* consists of *the last m* RGB frames and robot states.
- ***Seer predicts:***
  - The RGB image at time step *t+n*

$$\hat{o}_{t+n} = f_{\text{fore}}(g, h_t).$$

- ***Loss:***
  - The mean squared error (MSE) at the pixel level

$$\mathcal{L}_{\text{fore}} = \| f_{\text{fore}}(g, h_t) - o_{t+n} \|_2^2.$$

Seer: End-to-End PIDM

- First, we explain the end-to-end training of vision and action.

Action: Inverse Dynamics Prediction

- ***Seer takes as input:***
  - A goal *g* in the form of *language instructions* or *robot states*.
  - Historical observations *ht* consists of *the last m* RGB frames and robot states.
  - the predicted image in the latent space.
- ***Seer predicts:***
  - Action sequence *t : t+n-1*

$$\hat{a}_{t:t+n-1} = f_{\text{inv}}(g, h_t, \hat{o}^l_{t+n}).$$

  - predicting multiple steps provides temporal action consistency and robustness

- ***Loss:***
  - Smooth-L1 arm loss and Binary Cross Entropy (BCE) gripper loss

$$\mathcal{L}_{\text{inv}} = \mathcal{L}_{\text{arm}} + \lambda \mathcal{L}_{\text{gripper}},$$

Seer: End-to-End PIDM

- First, we explain the end-to-end training of vision and action.

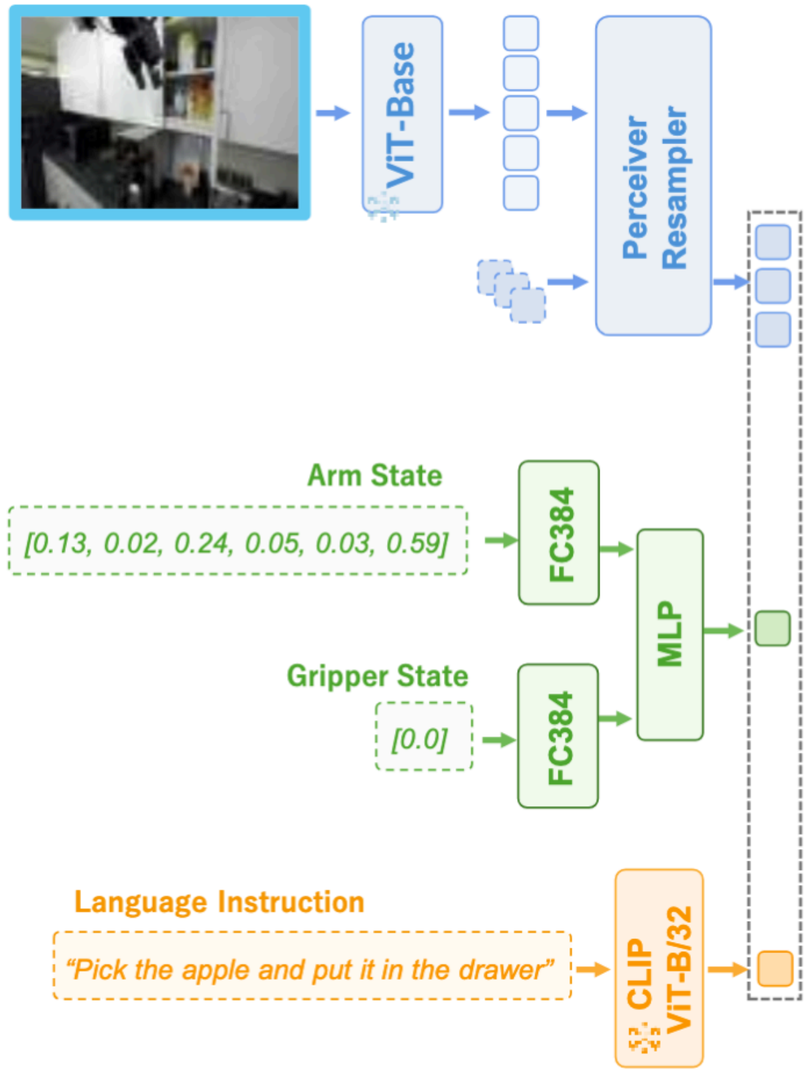
Vision + Action: Close the loop

- ***Total Loss:***
  - visual foresight loss + IDM loss

$$\mathcal{L} = \alpha \mathcal{L}_{\text{fore}} + \mathcal{L}_{\text{inv}}$$

Seer: Architecture

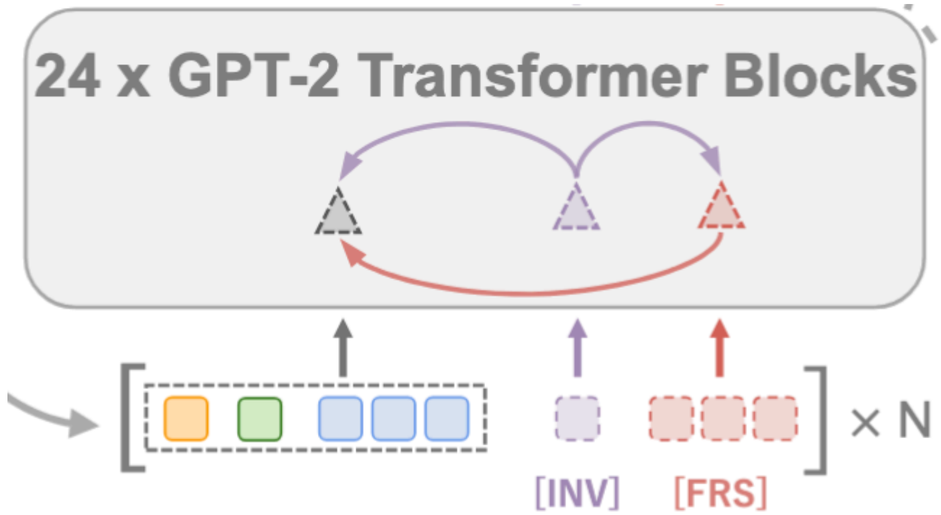
- **Input Tokenizers:**
  - **Language: Pre-trained CLIP text encoder**
  - **Robot states: MLPs**
  - **RGB frames: Pre-trained ViT + Perceiver Resampler**



KNOWLEDGE TECHNOLOGY

Seer: Architecture

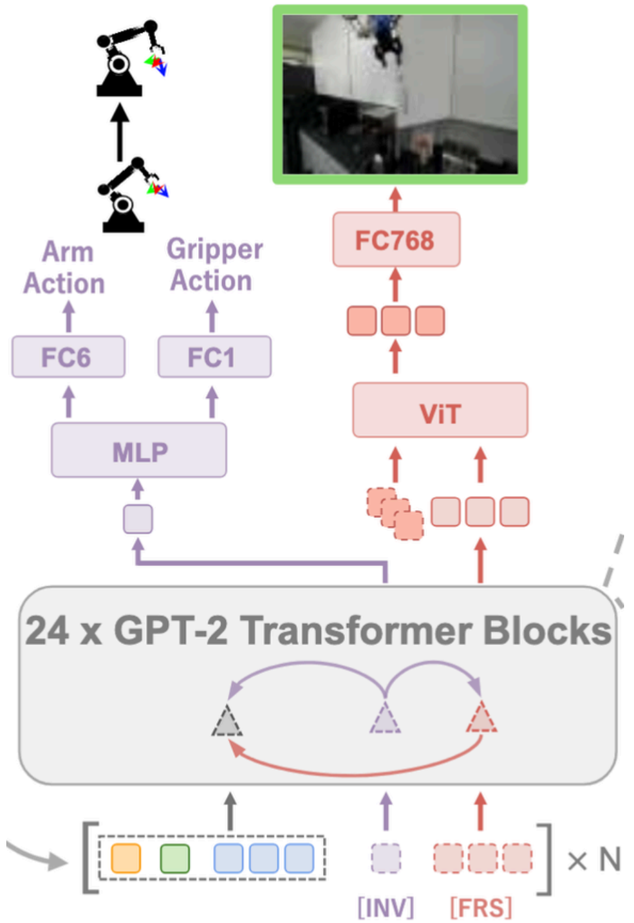
- **Multi-Modal Encoder:**
  - GPT-2 style Transformer
- **Readout Tokens:**
  - Append [FRS] and [INV] tokens to generate image and action latents



KNOWLEDGE TECHNOLOGY

Seer: Architecture

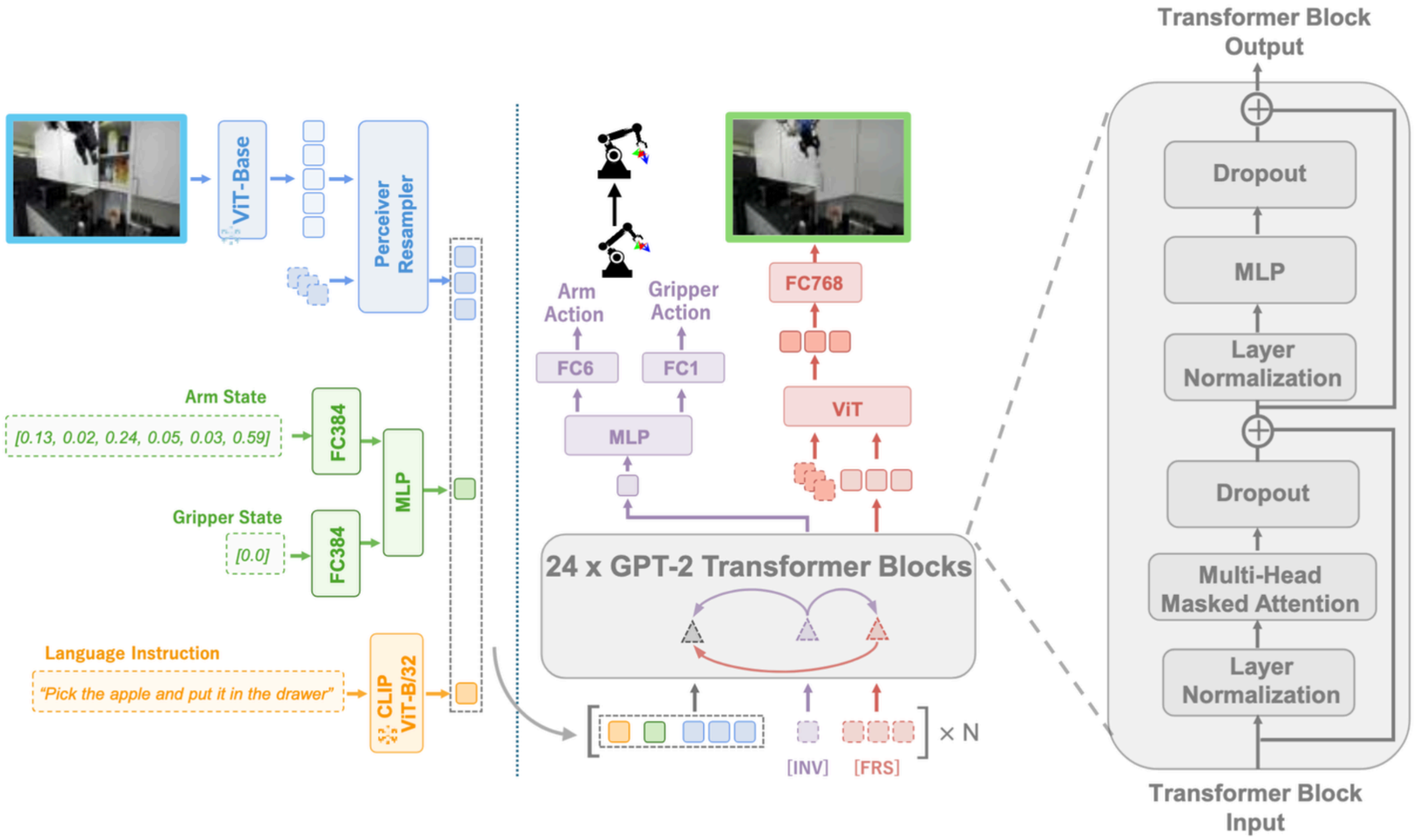
- **Decoders:**
  - The action and image latents generated by the [INV] and [FRS] tokens are input to the action decoder and image decoder to predict actions and images.
  - **Action decoder: MLPs**
  - **Image decoder: ViT**



KNOWLEDGE TECHNOLOGY

Seer: Architecture

- **Model size:**
  - Standard Seer: 65 million *(pre-trained vision and text encoders are kept frozen)*
  - Seer-Large: 315 million.



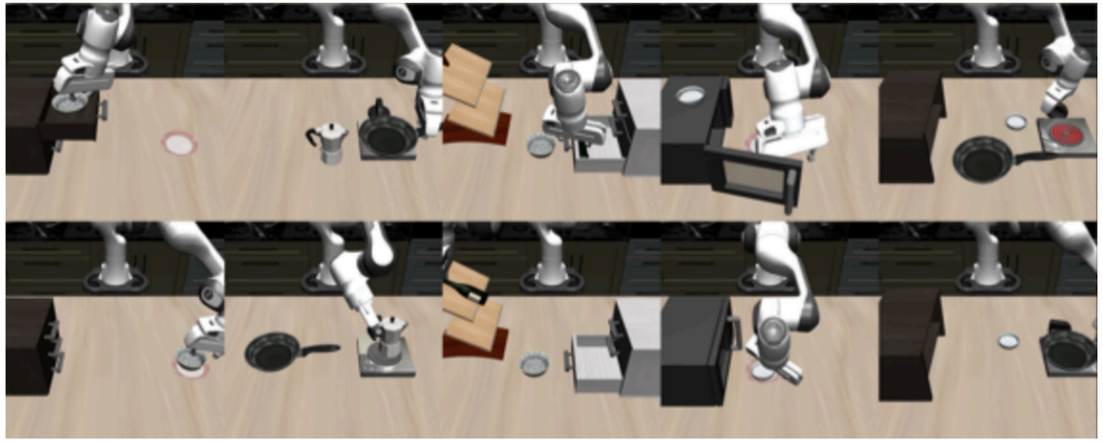
KNOWLEDGE TECHNOLOGY

Simulation Experiments

- **LIBERO Environment:**
  - *Pre-training:* LIBERO-90 dataset, which includes demonstrations for 90 short-horizon tasks with full annotations
  - *Fine-tuning:* LIBERO-LONG, which features long-horizon tasks

Method	Avg. Success ↑	Put soup and box in basket	Put box and butter in basket	Turn on stove and put pot	Put bowl in drawer and close it
MTACT	41.0	30.0	50.0	75.0	85.0
MVP	68.2	83.3	90.0	80.0	88.3
MPI	77.3	66.6	86.6	96.6	95.0
OpenVLA	54.0	35.0	**95.0**	65.0	45.0
Seer (scratch)	78.7	80.0	90.0	91.7	81.7
Seer	**87.7**	**91.7**	90.0	**98.3**	**100**

Put mugs on left and right plates	Pick book and place it in back	Put mug on plate and put pudding to right	Put soup and sauce in basket	Put both pots on stove	Put mug in microwave and close it
20.0	75.0	0.00	0.00	10.0	65.0
46.7	63.3	45.0	78.3	60.0	46.7
83.3	83.3	56.6	86.6	40.0	**78.3**
40.0	80.0	60.0	45.0	20.0	55.0
85.0	65.0	**86.7**	**88.3**	51.7	66.7
91.7	**93.3**	85.0	**88.3**	**61.7**	71.7

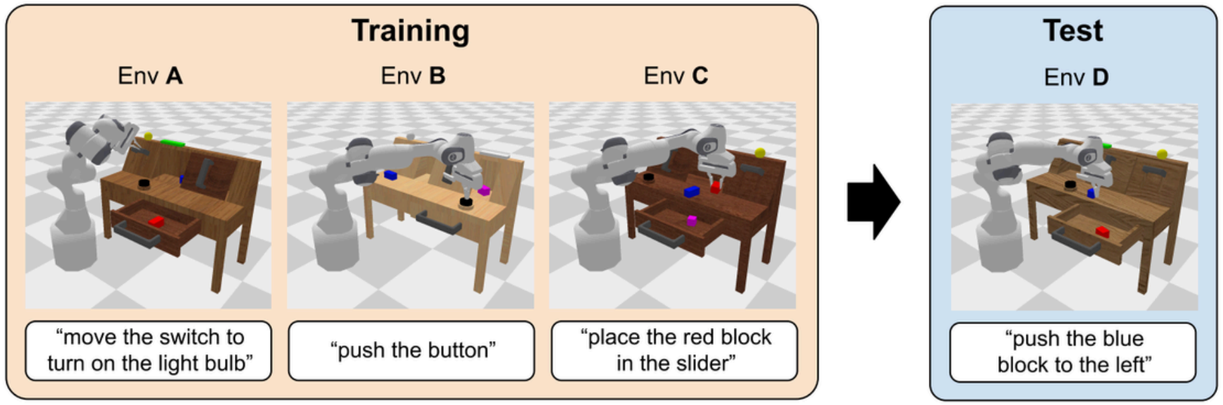


KNOWLEDGE TECHNOLOGY

Simulation Experiments

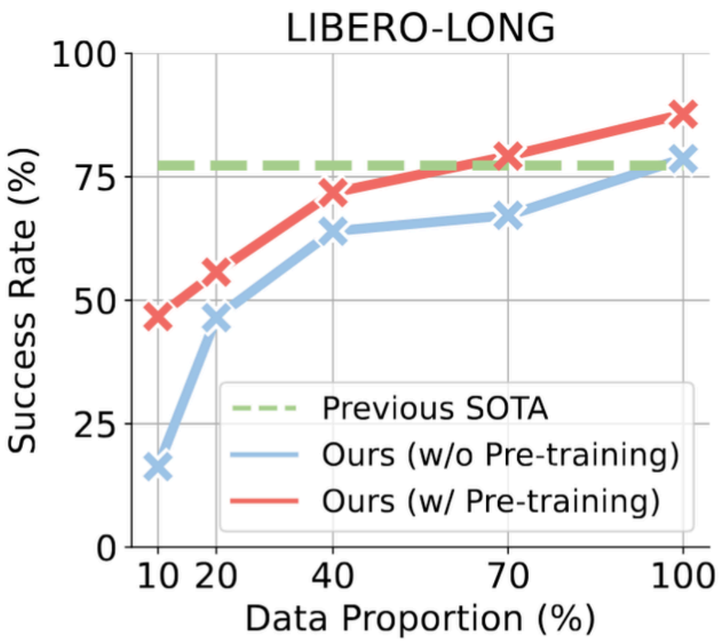
- **CALVIN ABC-D Environment:**
  - *Pre-training:* robot play data with no language instructions
  - *Fine-tuning:* 1% annotated data

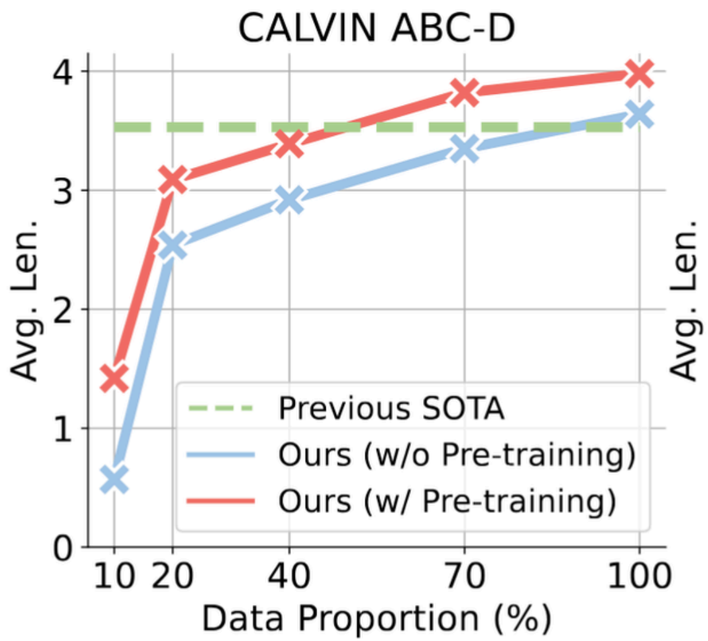
Method	Task completed in a row					
	1	2	3	4	5	Avg. Len. ↑
Roboflamingo	82.4	61.9	46.6	33.1	23.5	2.47
Susie	87.0	69.0	49.0	38.0	26.0	2.69
GR-1	85.4	71.2	59.6	49.7	40.1	3.06
3D Diffusor Actor	92.2	78.7	63.9	51.2	41.2	3.27
CLOVER	96.0	83.5	70.8	57.5	45.4	3.53
Seer (scratch)	93.0	82.4	72.3	62.6	53.3	3.64
Seer	94.4	87.2	79.9	72.2	64.3	3.98
Seer-Large (scratch)	92.7	84.6	76.1	68.9	60.3	3.83
Seer-Large	**96.3**	**91.6**	**86.1**	**80.3**	**74.0**	**4.28**

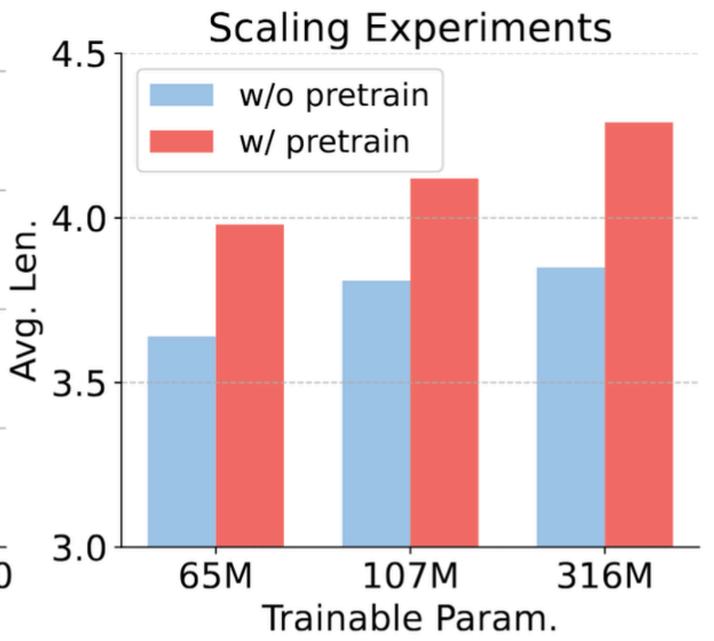


KNOWLEDGE TECHNOLOGY

Data Efficiency and Scalability







KNOWLEDGE TECHNOLOGY