Moto: Latent Motion Token as the Bridging Language for Learning Robot Manipulation from Videos

Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge,
Mingyu Ding, Ying Shan, Xihui Liu
The University of Hong Kong, ARC Lab, Tencent PCG,
The Chinese University of Hong Kong, University of California, Berkeley

ICCV HONOLULU HAWAII OCT 19-23, 2025

香 港 大 學
THE UNIVERSITY OF HONG KONG
ARC Applied Research Center
Berkeley UNIVERSITY OF CALIFORNIA
香 港 中 文 大 學
The Chinese University of Hong Kong

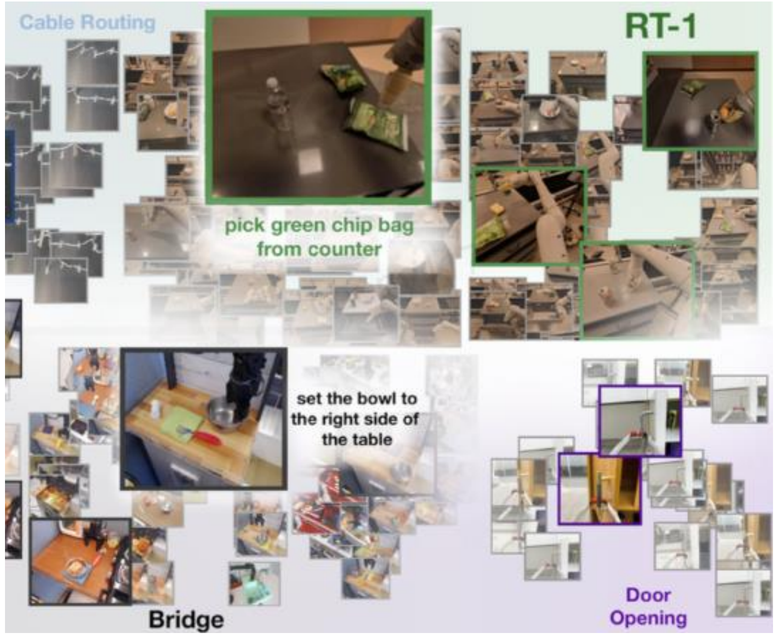
Background

- Robot data collection is slow and sparse, with varying action spaces across embodiments.

Zheng J., et al. Universal actions for enhanced embodied foundation models. In *CVPR*, 2025.

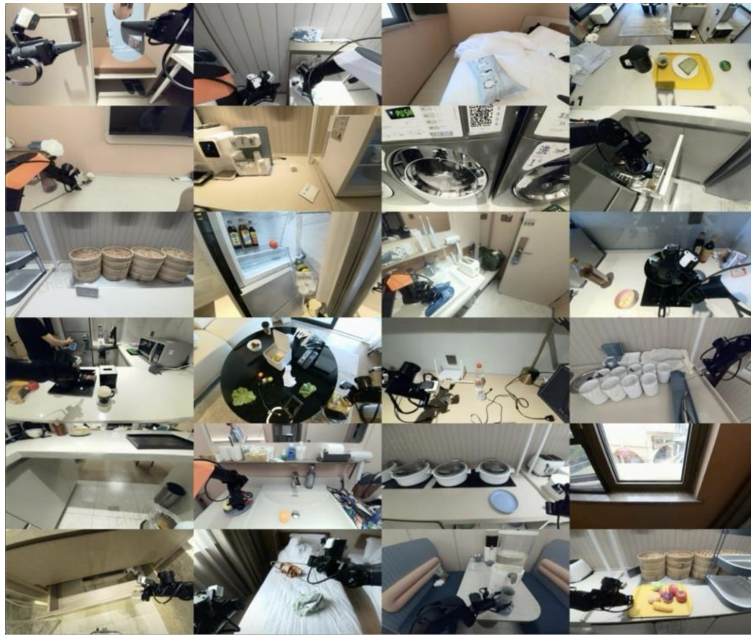
Background

- Video data is more diverse and scalable, which contains rich motion-related knowledge.



Open-X-Embodiment

Something-Something-V2



Galaxea Open-World Dataset

Motivation

- Large Language Models (LLMs) pre-trained on extensive corpora have shown significant success in various natural language processing (NLP) tasks with minimal fine-tuning.

- This success offers new promise for robotics, which has long been constrained by the high cost of action-labeled data.

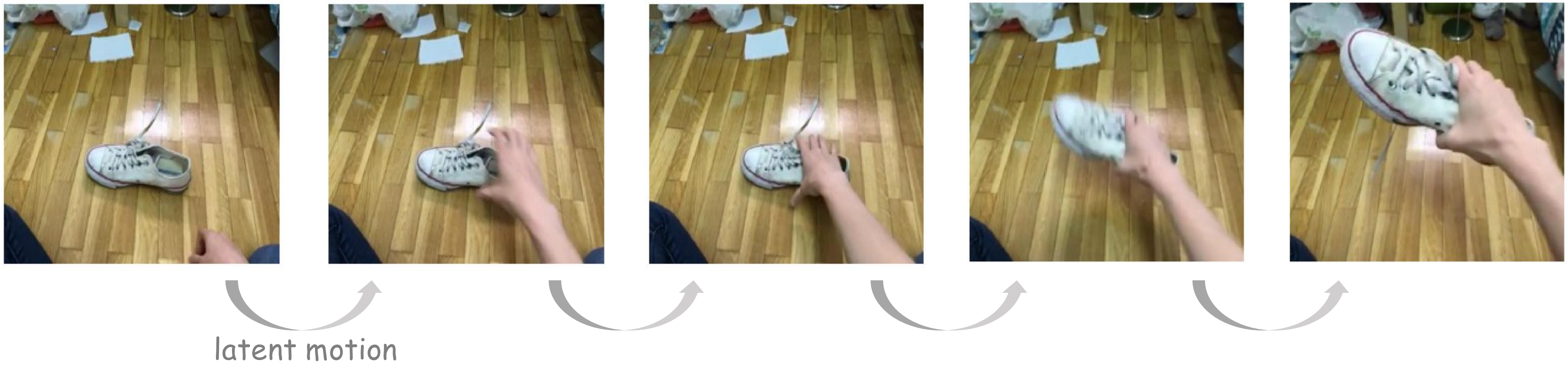
- ***Given the abundant video data containing interaction-related knowledge available as a rich "corpus", can we apply a similar generative pretraining approach to enhance robot learning?***

Motivation

- The key challenge is to identify an effective representation for autoregressive pre-training that benefits robot manipulation tasks.

Motivation

- Inspired by the way humans learn new skills through observing dynamic environments, we propose that, we propose that **effective robotic learning should emphasize motion-related knowledge**, which is **closely tied to low-level actions** and is **independent of hardware**, facilitating the transfer of learned motions to actual robot actions.



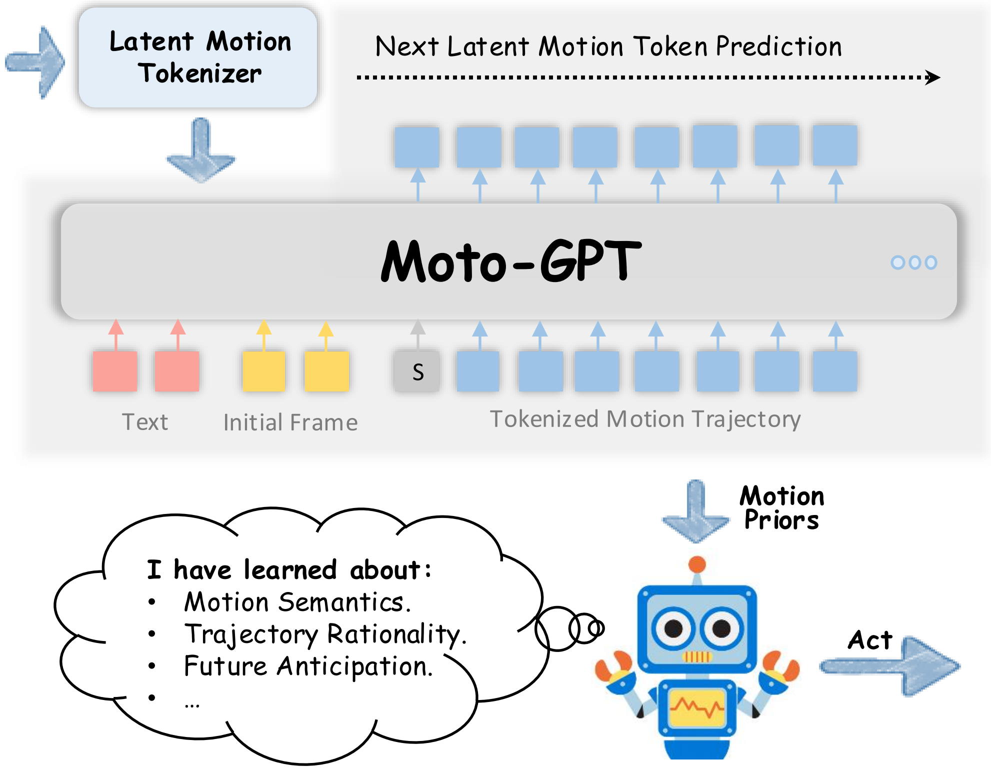
Motivation

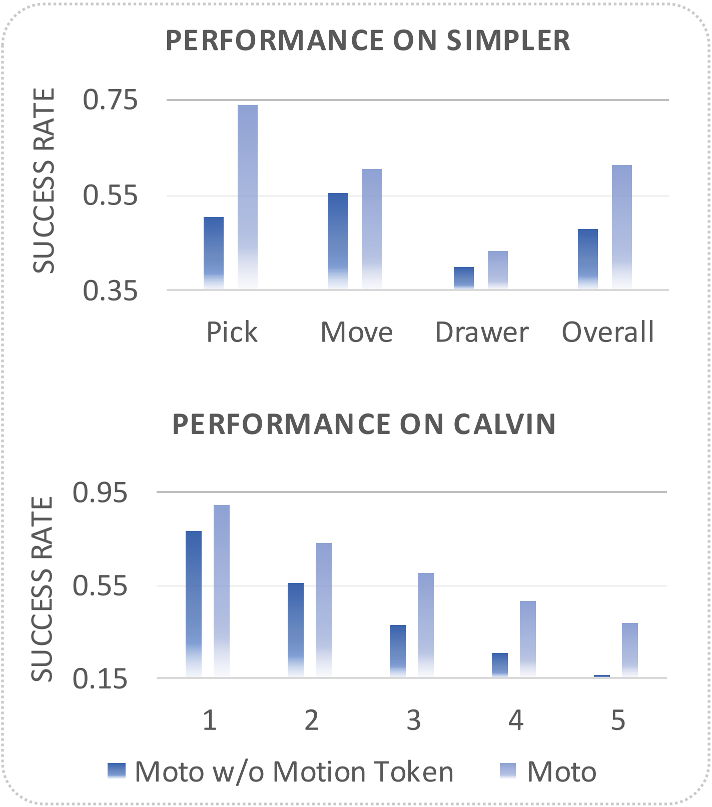
- Genie is the first generative interactive environment trained in an unsupervised manner from unlabelled Internet videos using latent actions.

Bruce, J., et al. Genie: Generative Interactive Environments. In *ICML*, 2024.

Overview of Moto

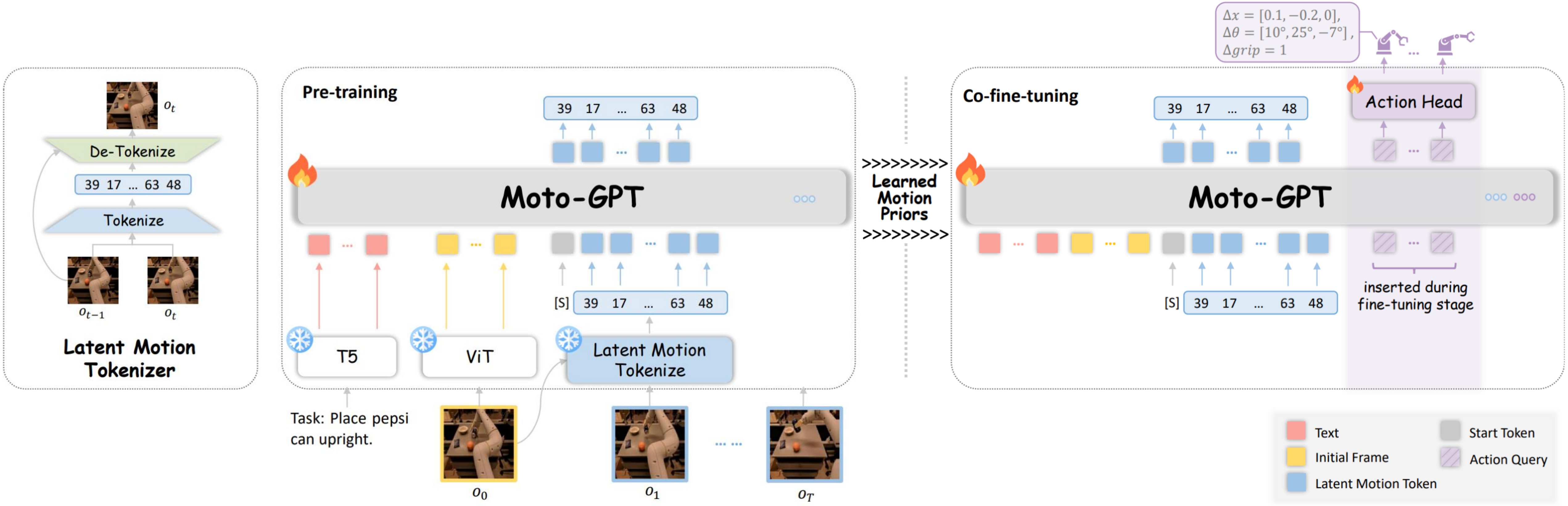
- **Moto** utilizes latent **Mo**tion **To**kens as a “language” interface to bridge generative pre-training on video data with precise robot control.



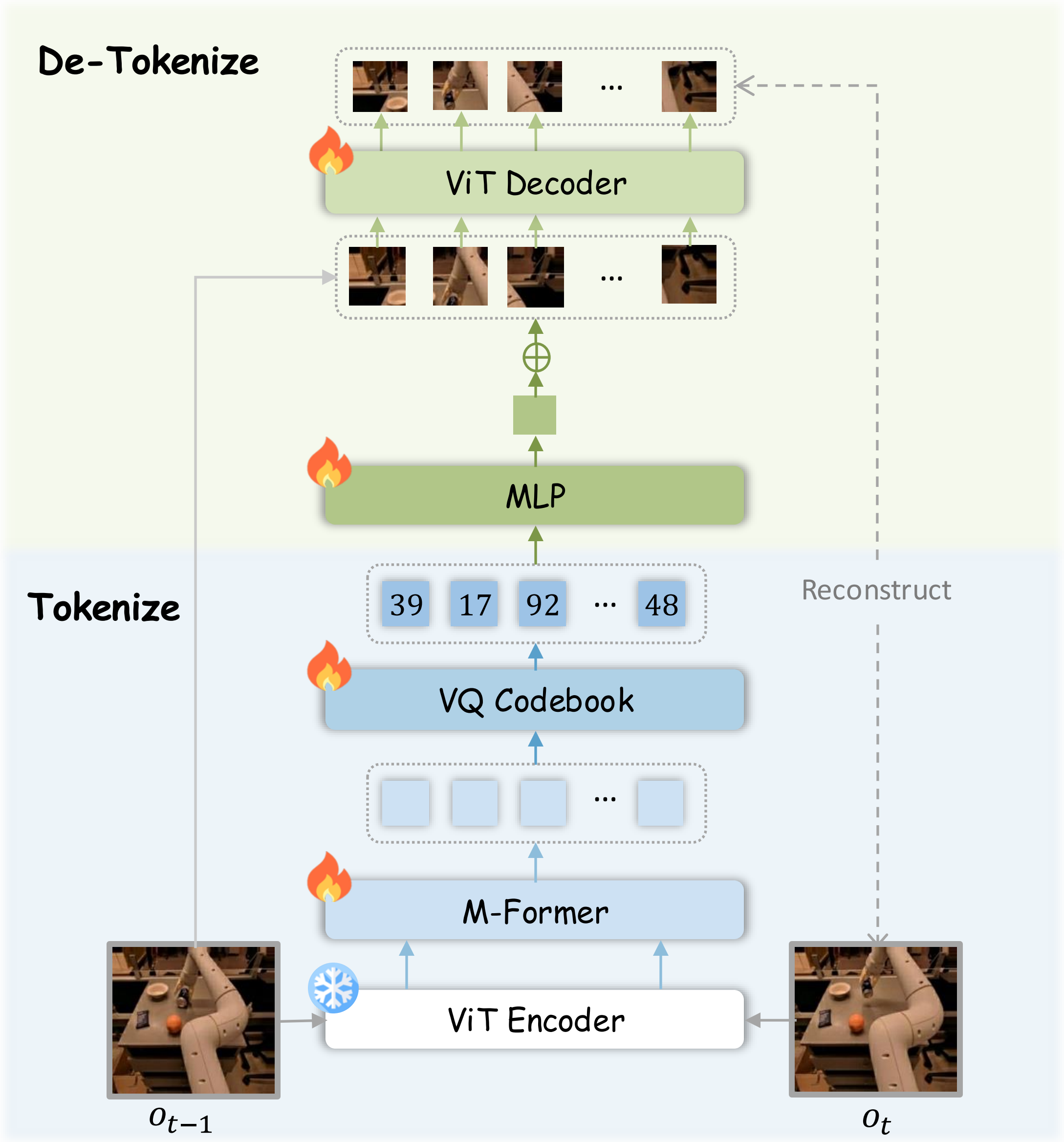


Training Procedures of Moto

- Moto consists of three stages: 1) unsupervised training of the Latent Motion Tokenizer, 2) pre-training of the generative model, and 3) co-fine-tuning for robot policy adaptation.

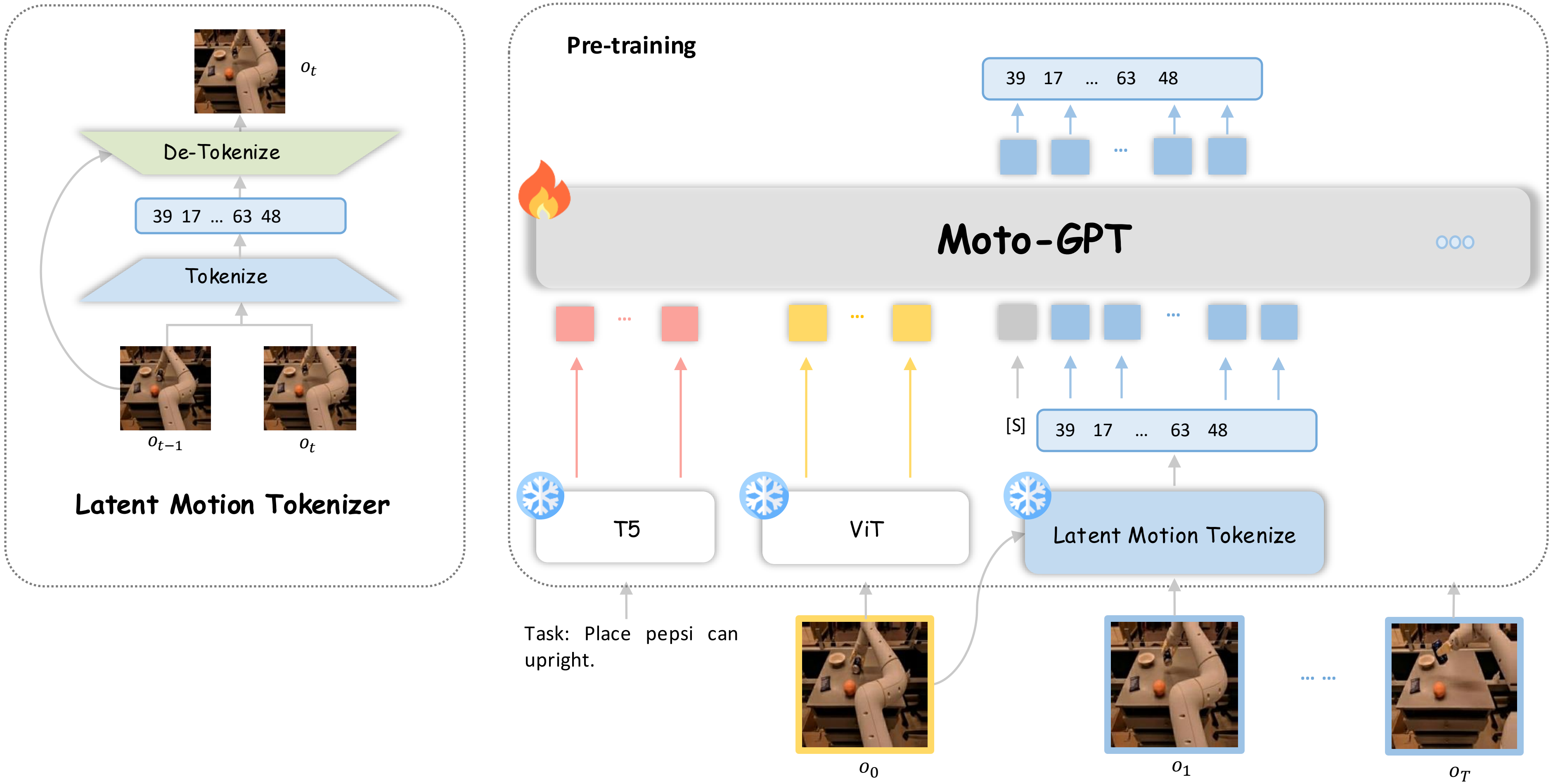


Stage-1: Latent Motion Tokenizer

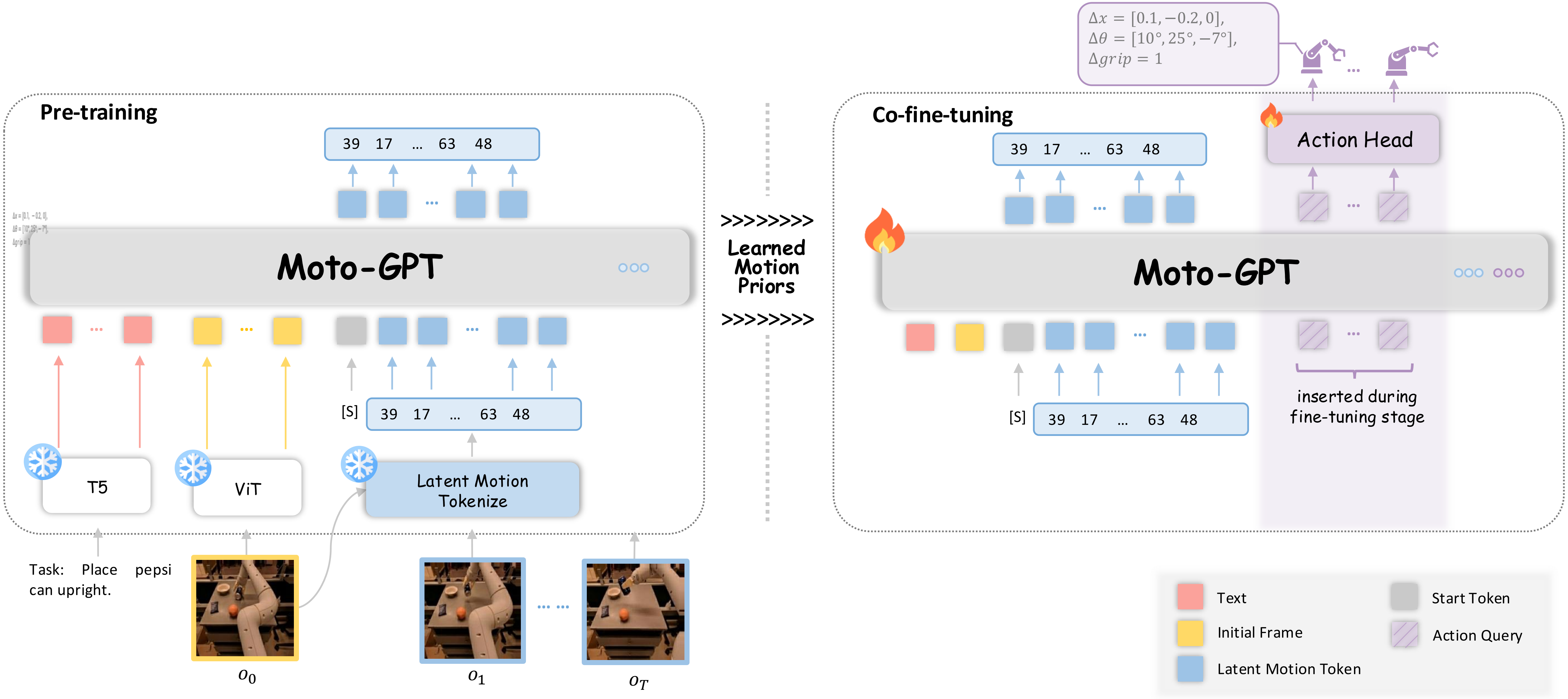


- The Latent Motion Tokenizer produces discrete latent motion tokens from two consecutive video frames.
- It regularizes the decoder to reconstruct the second frame based on the first frame and the discrete tokens, capturing essential visual motion between frames.

Stage-2: Motion Token Autoregressive Pre-training



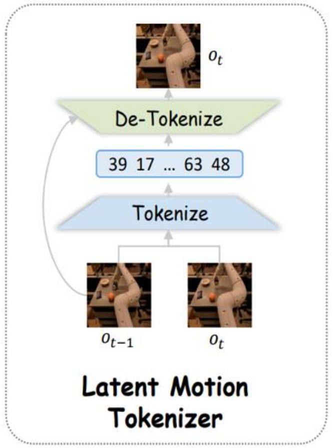
Stage-3: Co-fine-tuning for Robot Manipulation



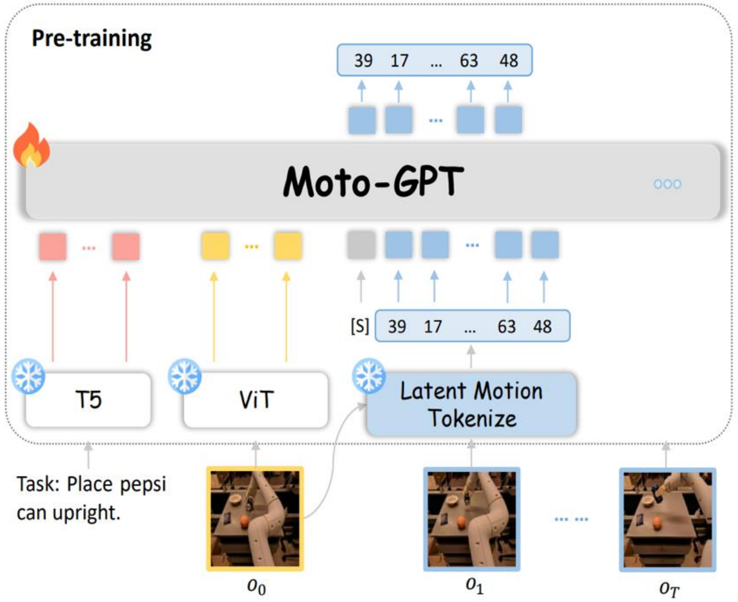
Experiments

To comprehensively evaluate the effectiveness of Moto, we study three key questions:

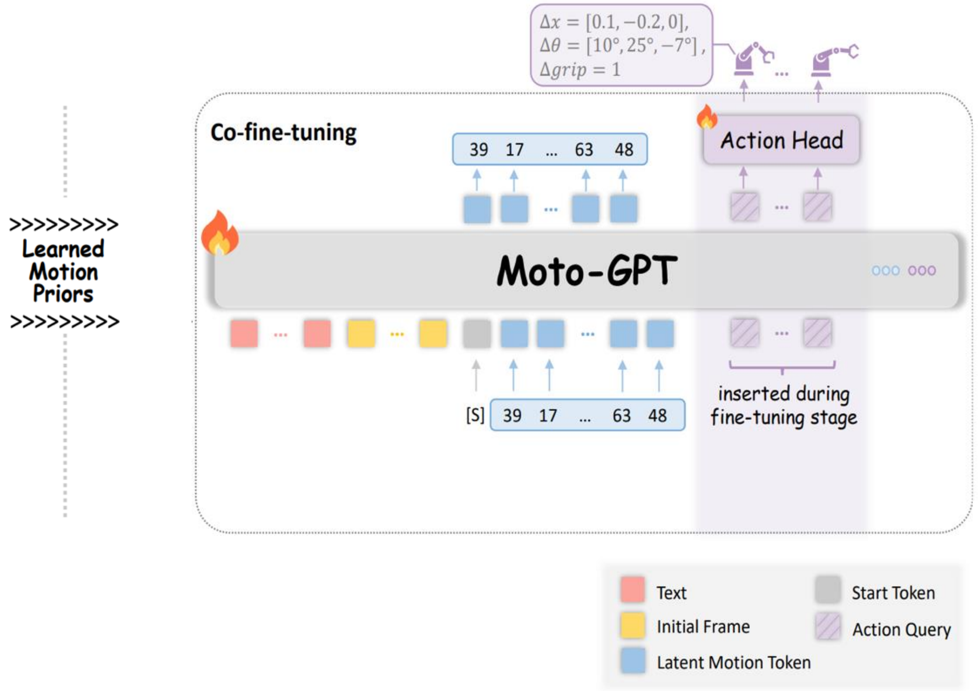
- **Q1 (Interpretability):** Do latent motion tokens represent meaningful visual motions?
- **Q2 (Motion Priors):** Does Moto-GPT learn useful priors about trajectories?
- **Q3 (Performance):** Can these priors be effectively transferred to real robot policies?



Q1 (Interpretability)



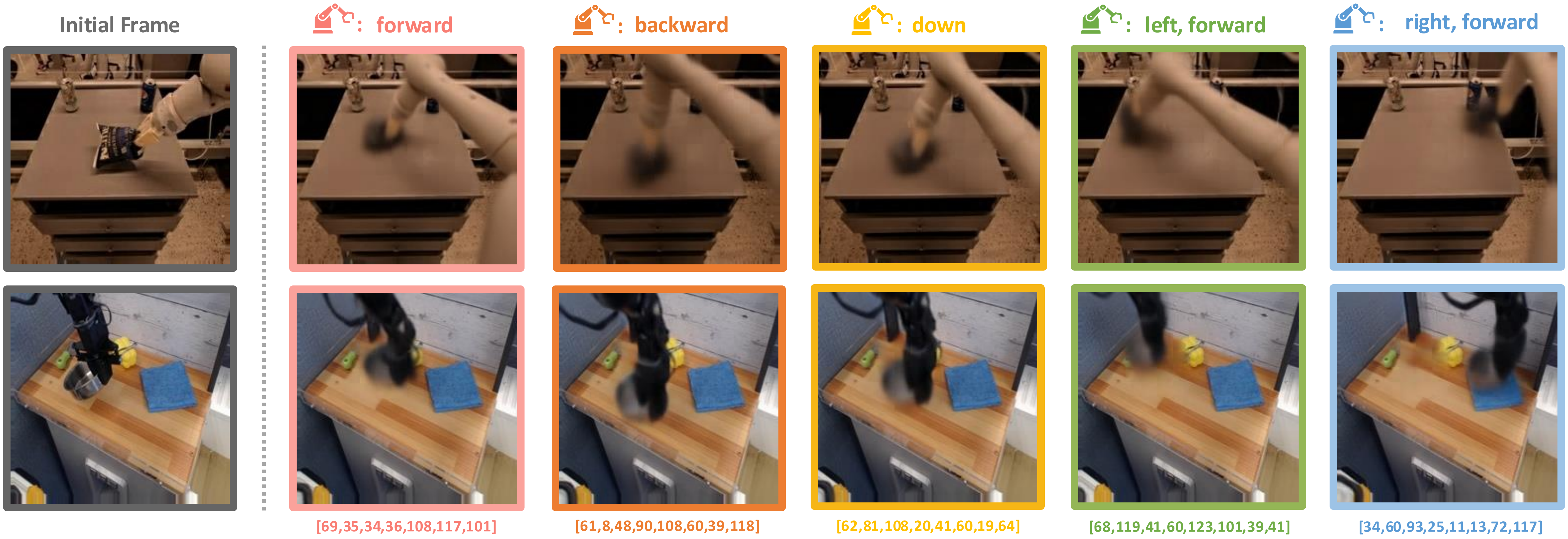
Q2 (Motion Priors)



Q3 (Performance)

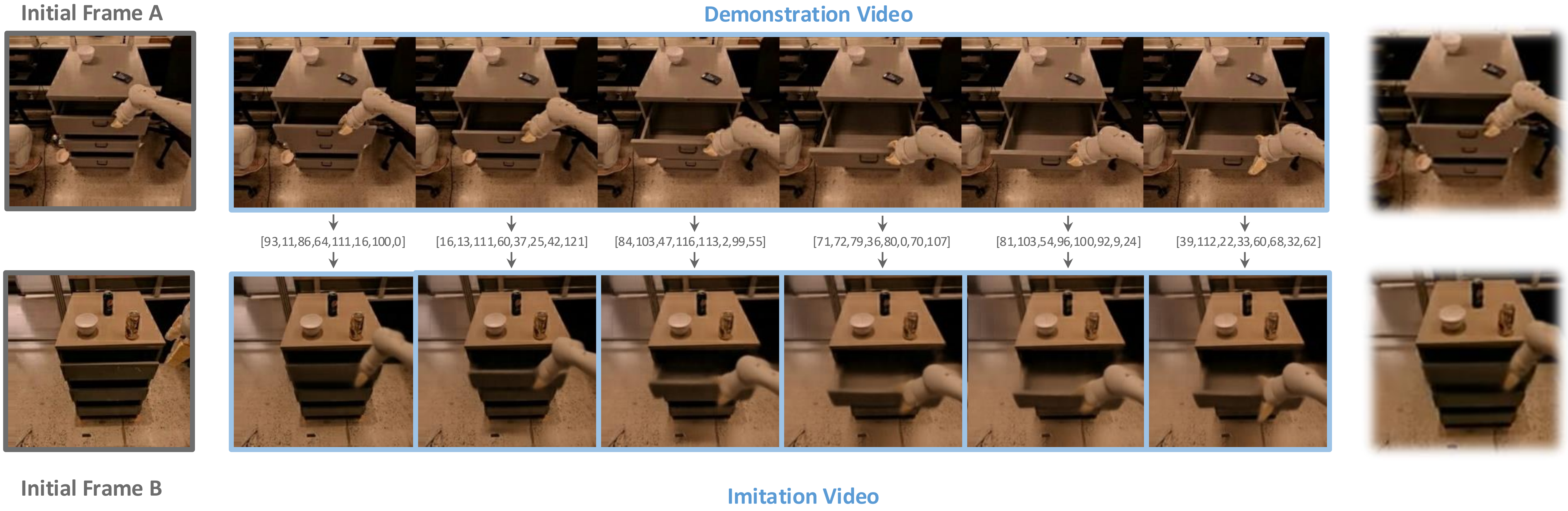
Latent Motion Token as an Interpretable Motion Language (Q1)

- Visualization of latent motion token interpretability



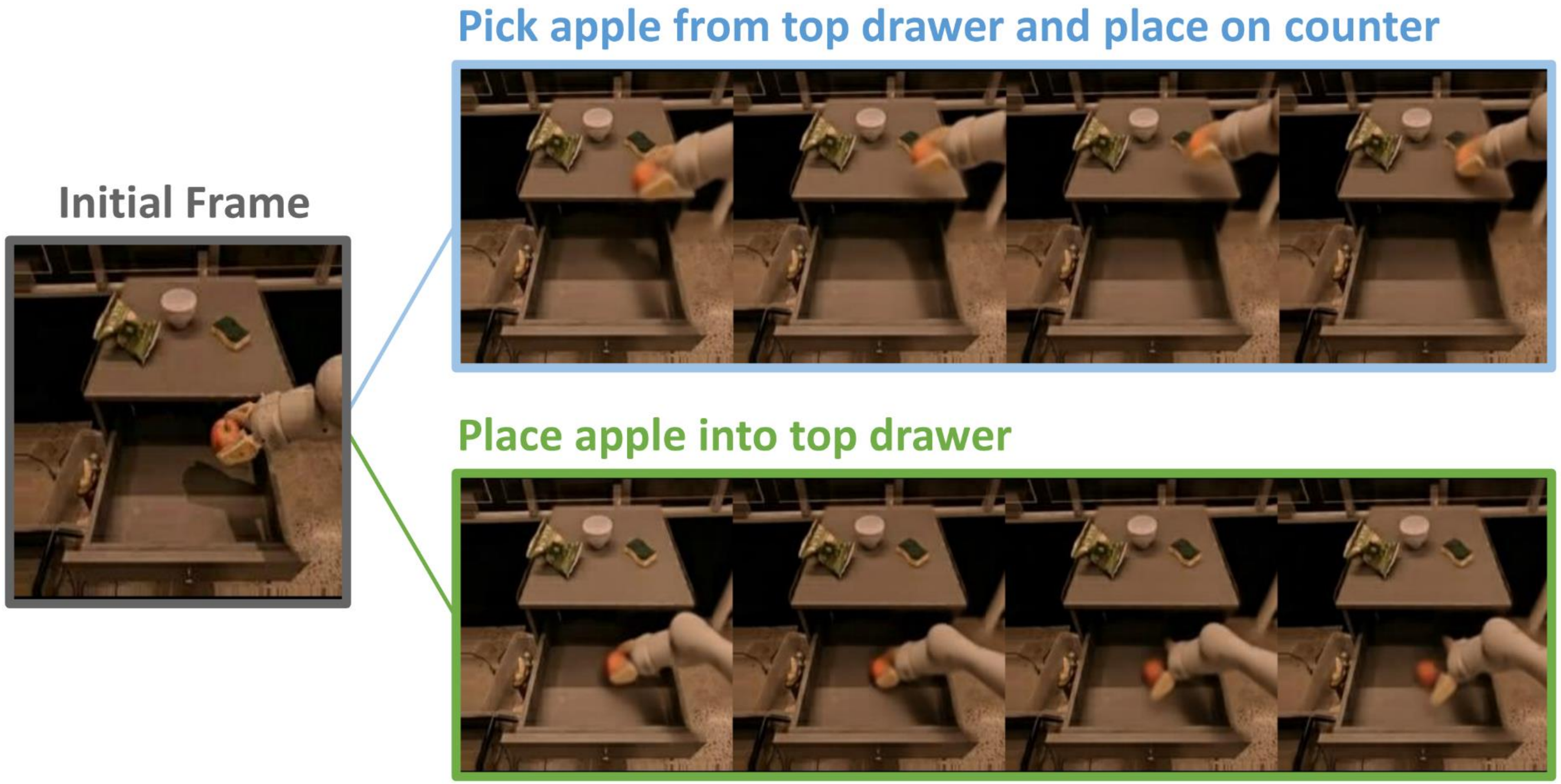
Latent Motion Token as an Interpretable Motion Language (Q1)

- Video imitation generation via latent motion tokens



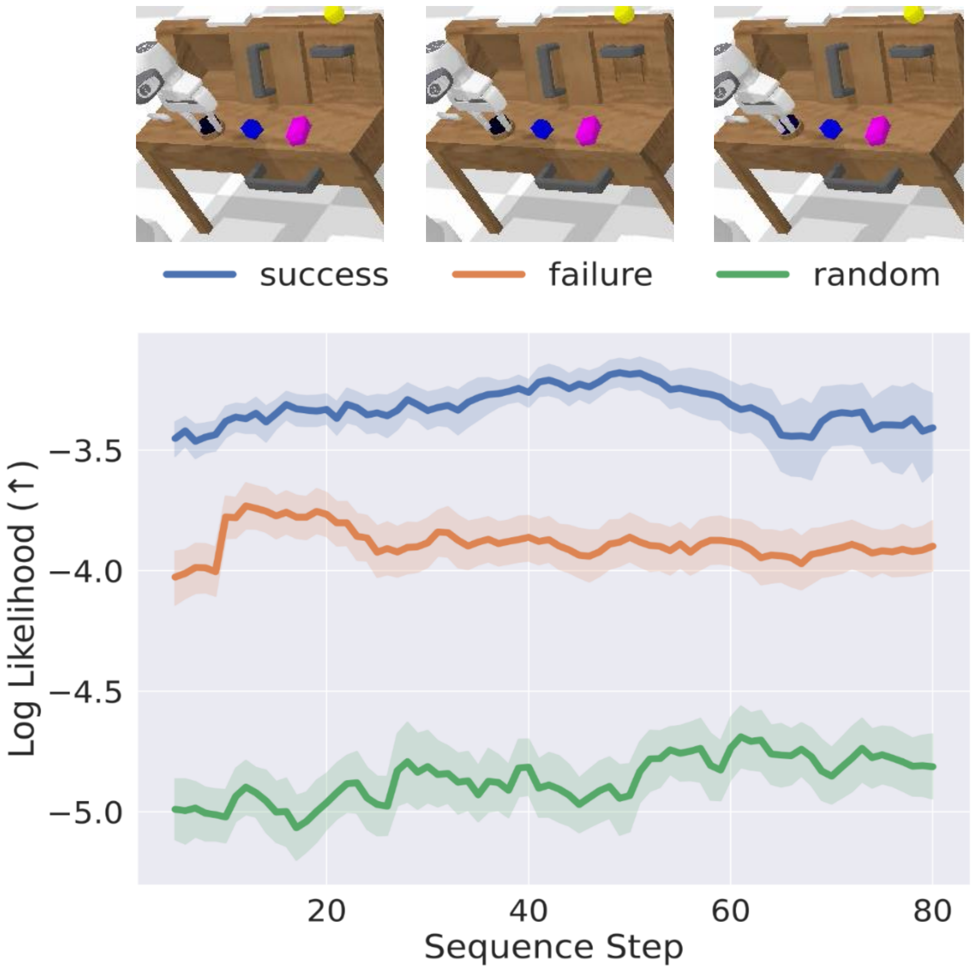
Pre-trained Moto-GPT as a Useful Prior Learner (Q2)

- Visualization of video trajectories generated from a sequence of latent motion tokens, which are predicted by the pre-trained Moto-GPT given different language instructions.



Pre-trained Moto-GPT as a Useful Prior Learner (Q2)

- Moto-GPT distinguishes successful, failed, and random trajectories using log-likelihoods, enabling effective assessment of robot trajectory rationality and potential reward signals.



Fine-tuned Moto-GPT as an Effective Robot Policy (Q3)

- **Performance on SIMPLER**

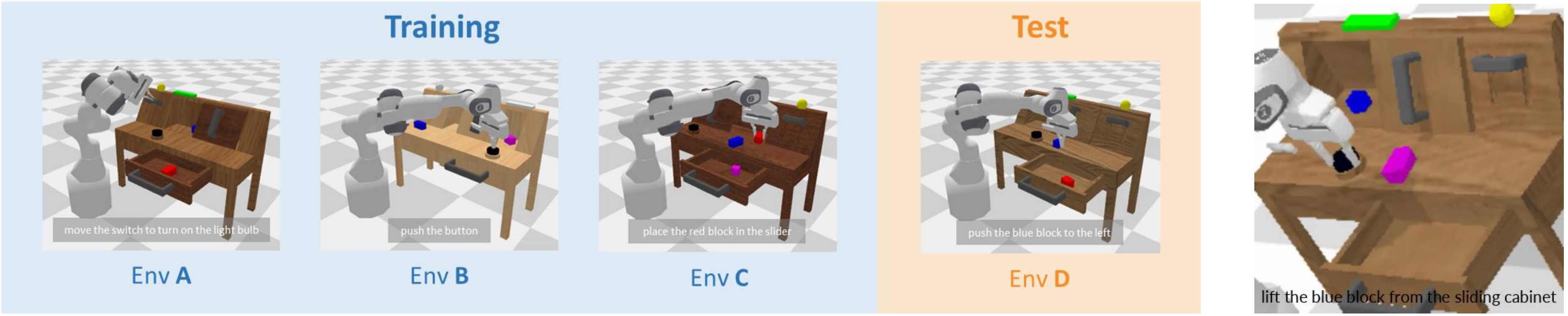
  Moto-GPT achieves competitive performance with larger models like RT-2-X (PaLI-X 55B) and OpenVLA (Prismatic 7B), despite having only 98M parameters for the GPT-style backbone.

Method	Pick Coke Can				Move Near	Open / Close Drawer			Overall
	Horizontal	Vertical	Standing	Average	Average	Open	Close	Average	Average
RT-1-X [4]	**0.820**	0.330	0.550	0.567	0.317	**0.296**	**0.891**	**0.597**	0.534
RT-2-X [62]	0.740	**0.740**	0.880	**0.787**	**0.779**	0.157	0.343	0.250	0.607
Octo-Base [41]	0.210	0.210	0.090	0.170	0.042	0.009	0.444	0.227	0.169
OpenVLA [27]	0.270	0.030	0.190	0.163	0.462	0.194	0.518	0.356	0.248
OpenVLA (fine-tuned) [27]	0.470	0.080	0.540	0.363	0.542	0.102	0.361	0.231	0.349
Moto	**0.820**	0.500	**0.900**	0.740	0.604	0.130	0.732	0.431	**0.614**
Moto w/o Motion Token	0.600	0.190	0.740	0.503	0.554	0.000	0.796	0.398	0.480

Fine-tuned Moto-GPT as an Effective Robot Policy (Q3)

- **Performance on CALVIN (ABC→D)**

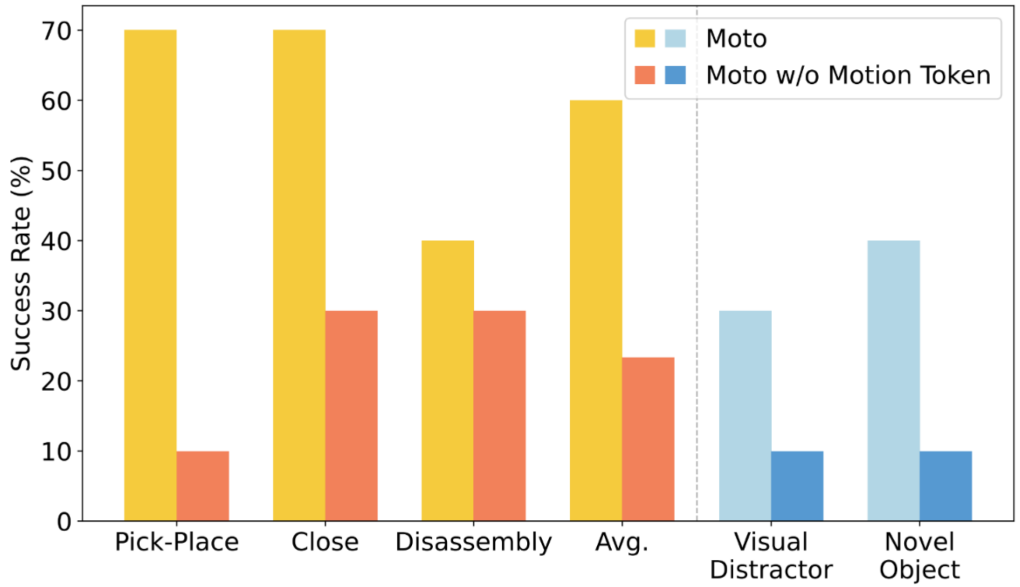
  Moto-GPT shows strong zero-shot generalization ability in the unseen CALVIN environment, despite relying solely on RGB images from a static camera.

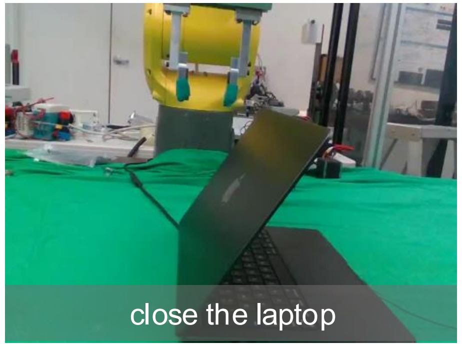


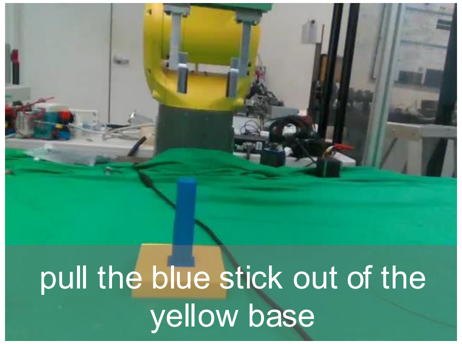
Model	Observation Space	Tasks competed in a row (1000 chains)					
		1	2	3	4	5	Avg. Len.
SuSIE [2]	Static RGB	0.870	0.690	0.490	0.380	0.260	2.69
RoboFlamingo [28]	Static RGB + Gripper RGB	0.824	0.619	0.466	0.331	0.235	2.47
MT-R3M [49]	Static RGB + Gripper RGB + Proprio	0.529	0.234	0.105	0.043	0.018	0.93
GR-1 [49]	Static RGB + Gripper RGB + Proprio	0.854	0.712	0.596	**0.497**	**0.401**	3.06
Moto	Static RGB	**0.897**	**0.729**	**0.601**	0.484	0.386	**3.10**
Moto w/o Motion Token	Static RGB	0.779	0.555	0.380	0.256	0.167	2.14

Fine-tuned Moto-GPT as an Effective Robot Policy (Q3)

- **Performance in Real-World Environment**

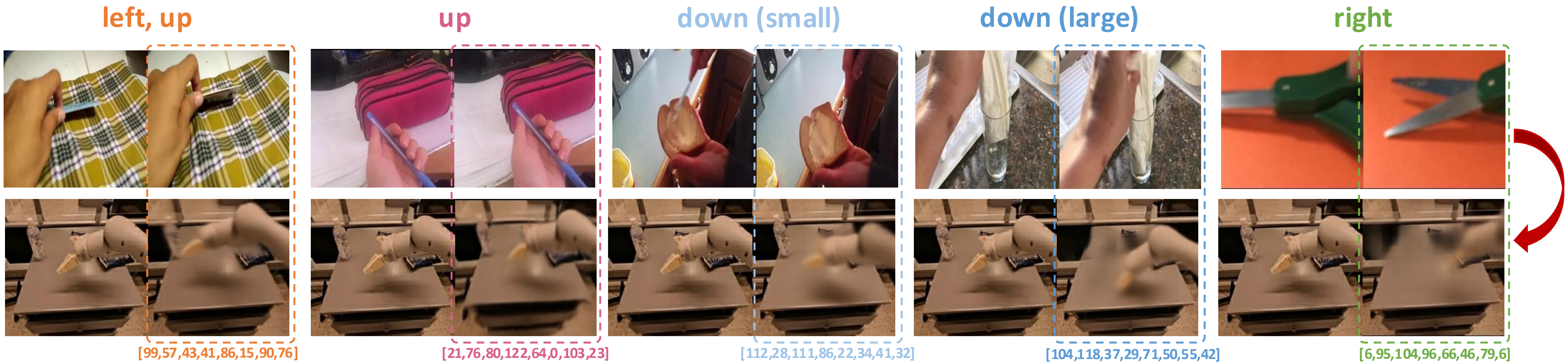


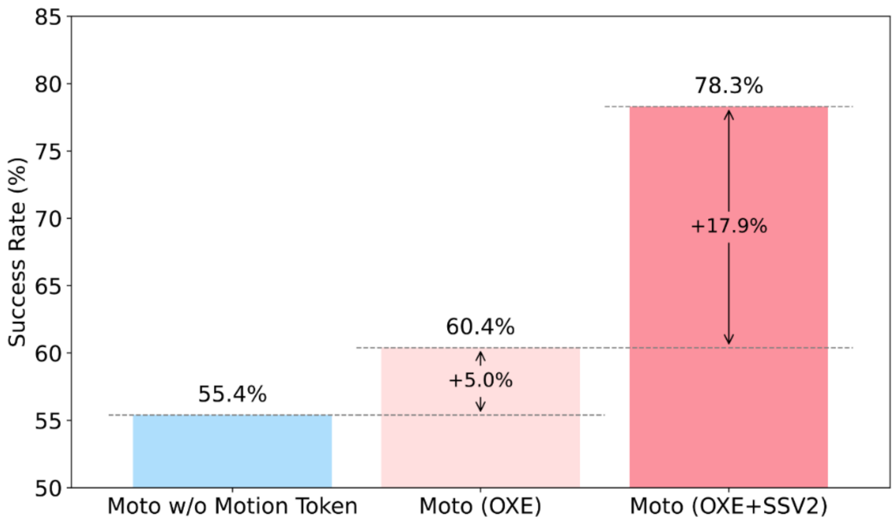




Fine-tuned Moto-GPT as an Effective Robot Policy (Q3)

- **Learning from Human Videos**

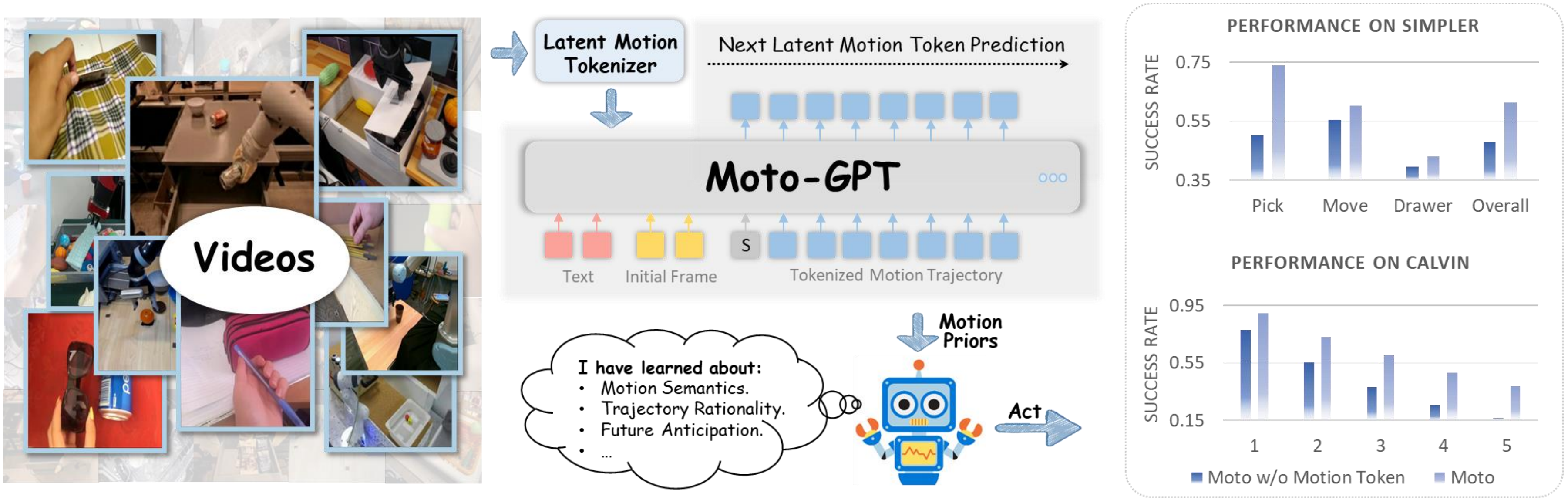




Conclusion

- We present Moto, a novel method that utilizes latent motion tokens as a “language” interface to bridge generative pre-training on video data with precise robot control.

- By learning motion-related priors from videos without the need for hardware-specific action labels, Moto effectively translates learned motions into precise robot actions.



Future Directions

- Learning from large-scale in-the-wild human videos
  - Decoupling camera motion and hand movements.
- Application to more robot embodiments and tasks
  - Dual-arm robots, dexterous manipulation, whole-body control
- Improve the Latent Motion Tokenizer
  - e.g., incorporating 3D information, combining ground-truth action labels
- Retrieval augmented generation / In-context learning with few-shot demonstrations

See our project page for more details!