CARP: Visuomotor Policy Learning via Coarse-to-Fine Autoregressive Prediction

Zhefei Gong (Presenter), Pengxiang Ding, Shangke Lyu, Siteng Huang,
Mingyang Sun, Wei Zhao, Zhaoxin Fan, Donglin Wang

ICCV OCT 19-23, 2025 HONOLULU HAWAII

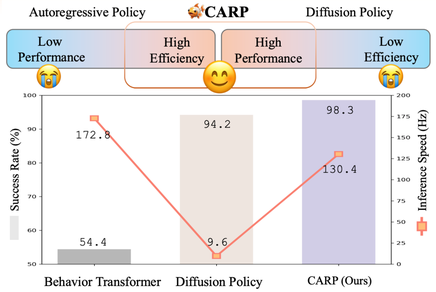
Homepage https://carp-robot.github.io/

Code https://github.com/ZhefeiGong/carp

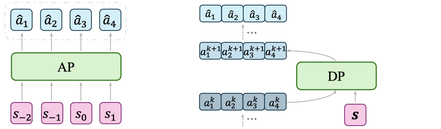
TL;DR

We propose **CARP**, a **coarse-to-fine**, **next-scale** autoregressive policy for visuomotor learning that progressively refines actions across scales, enabling **efficient** and **high-performance** generation.

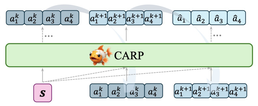
Motivation



- **Autoregressive Policy** struggles to capture long-range dependencies, global structures, and temporal coherence, resulting in suboptimal performance.
- **Diffusion Policy** relies on multi-step denoising, making it computationally expensive and inefficient for robotic tasks requiring efficient real-time inference.



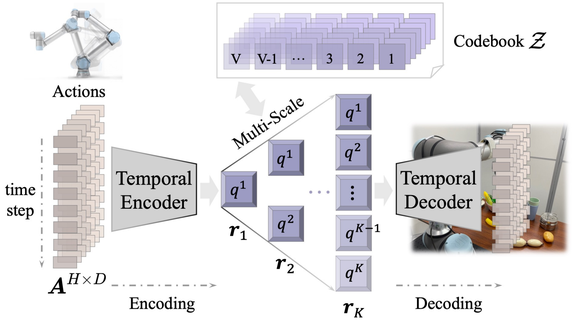
How can we design **a new action generation** paradigm that resolves **the trade-off between efficiency and high performance**?
— combining the strengths of autoregressive prediction with multi-stage action refinement?



Method

Coarse-to-Fine AutoRegressive Policy (CARP)

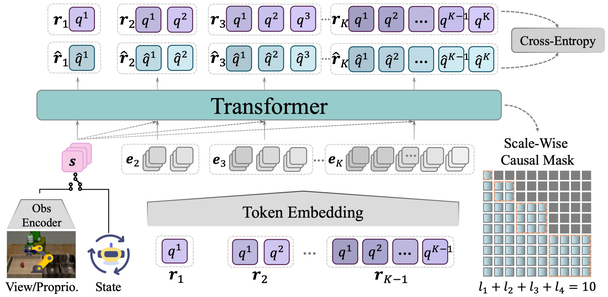
Multi-Scale Action Tokenization



A multi-scale action autoencoder extracts token maps $r_1, r_2, \ldots, r_K$ to represent the action sequence at different scales, trained using the standard VQVAE loss.

$$\mathcal{L} = \underbrace{\|A - \hat{A}\|_2}_{\mathcal{L}_{\text{recon}}} + \underbrace{\|\text{sg}(F) - \hat{F}\|_2}_{\mathcal{L}_{\text{quant}}} + \underbrace{\|F - \text{sg}(\hat{F})\|_2}_{\mathcal{L}_{\text{commit}}}$$

Coarse-to-Fine Autoregressive Prediction



The autoregressive prediction is reformulated into a **coarse-to-fine, next-scale paradigm**, where the sequence is progressively refined from a coarse token map r_1 to a fine-grained map r_K.

Each scale is trained with cross-entropy loss, and the multi-scale token maps $r_{1:K}$ are decoded into continuous actions.

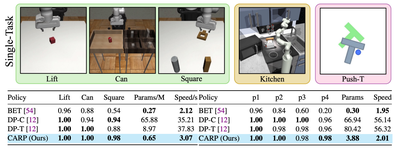
$$p(r_1, r_2, \ldots, r_K) = \prod_{k=1}^{K} p(r_k \mid r_1, r_2, \ldots, r_{k-1}; s_O)$$

$$\mathcal{L}_{\text{Cross-Entropy}} = \sum_{k=1}^{K}\sum_{i=1}^{l_k}\left[-\sum_{v=1}^{V} r_k^{i,v} \log \hat{r}_k^{i,v}\right]$$

Experiments

High Performance Comparable to Diffusion-Based Policies?

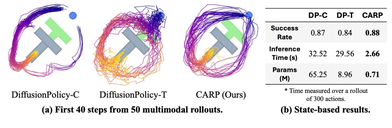
High Inference Efficiency Comparable to Autoregressive Policies?



Policy	Lift	Can	Square	Params/M	Speed/s
BET [54]	0.96	0.88	0.54	**0.27**	**2.12**
DP-C [12]	**1.00**	0.94	**0.94**	65.88	35.21
DP-T [12]	**1.00**	**1.00**	0.88	8.97	37.83
CARP (Ours)	**1.00**	**1.00**	**0.98**	0.65	3.07

Policy	p1	p2	p3	p4	Params	Speed
BET [54]	0.96	0.84	0.60	0.20	**0.30**	**1.95**
DP-C [12]	**1.00**	**1.00**	**1.00**	0.96	66.94	56.14
DP-T [12]	**1.00**	0.98	0.98	0.96	80.42	56.32
CARP (Ours)	**1.00**	**1.00**	0.98	**0.98**	**3.88**	**2.01**

Overall Performance and Efficiency



(a) First 40 steps from 50 multimodal rollouts.

	DP-C	DP-T	CARP
Success Rate	0.87	0.84	**0.88**
Inference Time (s)	32.52	29.56	**2.66**
Params (M)	65.25	8.96	**0.71**

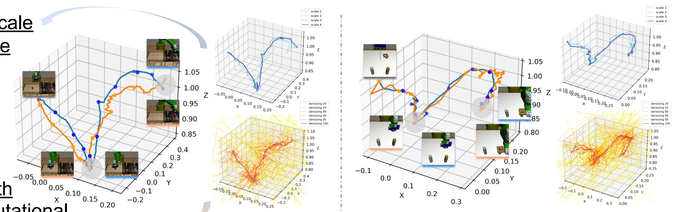
* Time measured over a rollout of 300 actions.

(b) State-based results.

Multi-Modal Behavior Results

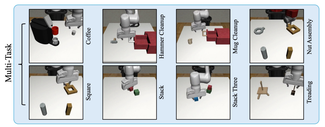
compact 4-scale coarse-to-fine steps

100-step denoising with heavy computational redundancy



Visualization of Trajectory Refinement

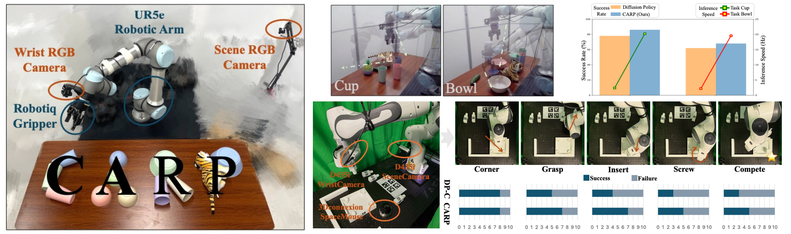
Flexibility Benefits of a GPT-Style Architecture in Multi-Task Scenarios?



Policy	Prams/M	Speed/s	Coffee	Hammer	Mug	Nut	Square	Stack	Stack three	Threading	Avg.
TCD [36]	156.11	107.15	0.77	0.92	0.53	0.44	0.63	0.95	0.62	0.56	0.68
SDP [63]	159.85	112.39	0.82	**1.00**	0.62	0.54	0.82	0.96	0.80	**0.70**	0.78
CARP (Ours)	**16.08**	**6.92**	**0.86**	0.98	**0.74**	**0.78**	**0.90**	**1.00**	**0.82**	**0.70**	**0.85**

Simply extending a single-task policy with a task embedding **outperforms** specially designed multi-task diffusion-based policies.

Real-World Results for Effectiveness Verification?



Conclusion / Takeaways

- A novel visuomotor policy learning paradigm through **coarse-to-fine**, **next-scale** autoregressive action generation.
- Achieve up to 10% **higher success rates** and 10× **faster inference** than state-of-the-art policies.
- Retain strong **multi-modal modeling** capability for complex action distributions.
- Offer high **architectural flexibility**, easily extensible to multi-task settings with simple scaling and conditioning.
- Demonstrated effectiveness across multiple **real-world tasks**, including pick-and-place, rotation, insertion, etc.



Zhefei Gong (龚哲飞)
Homepage (https://zhefeigong.github.io/)
WeChat (zhefeigong)

Looking for 2026Fall Phd Position

Robot Learning / Reinforcement Learning / Generative Modeling / Robotics

Experienced with real-world robots including UR5e, Franka Panda, and Tiago.



Always open to chatting about collaborations or PhD positions.

zhefeigong at gmail dot com