SpecPrune-VLA: Accelerating Vision-Language-Action Models via Action-aware Self-Speculative Pruning

Hanzhen Wang†, Jiaming Xu†, Yushun Xiang, Jiayi Pan, Yongkang Zhou, Yong-Lu Li and Guohao Dai*
Shanghai Jiao Tong University, Shanghai Innovation Institute, Infinigence AI

*Corresponding authors
†These authors contributed equally.

ICML International Conference On Machine Learning

A training-free two-level token pruning framework for Vision-Language-Action models that combines **global-temporal consistency**, local **layer-wise importance** and **action-aware** control.

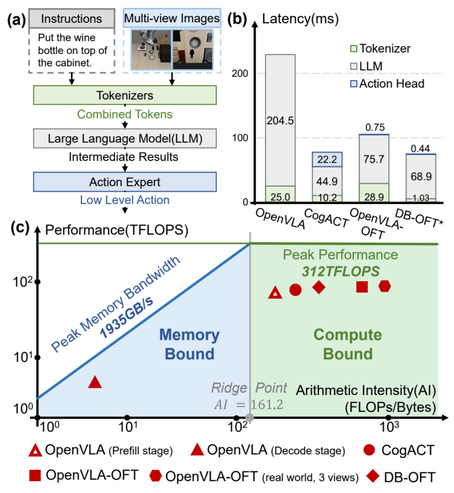
1.57×	**1.70×**	**Negligible**
speedup in simulation	speedup on real-world tasks	success-rate degradation

1. Background & Key Insight

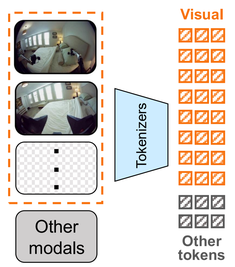
Motivation

- VLA inference is dominated by the LLM backbone (70% of end-to-end lateney in most models).
- Modern single-step VLA inference is compute-boundrather than memory-bound.
- Existing pruning methods rely mainly on loginformation and may incur >20% success-ratedrops or limited speedup.

Roofline Model on Nvidia A800 GPU



More than 80% visual input



Bottleneck:

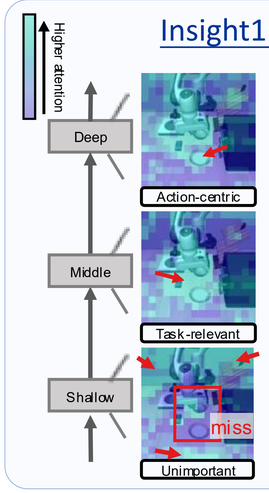
- Memory Bound: Data transfer
- Compute Bound: Busy computation

FFN and ATTN in VLA model are compute bound

- Latency(ATTN) ∝ L^2
- Latency(FFN) ∝ L

Key Insights

Insight1: Hierarchical attention pattern

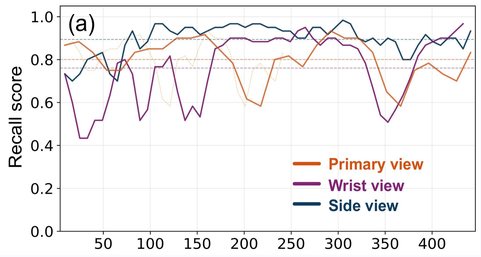


→ Focus on action-centric regions directly involved in execution

→ Focus on objects and regions relevant to task understanding

→ Attend broadly, including background and irrelevant areas.

Insight2: Spatial-temporal consistency

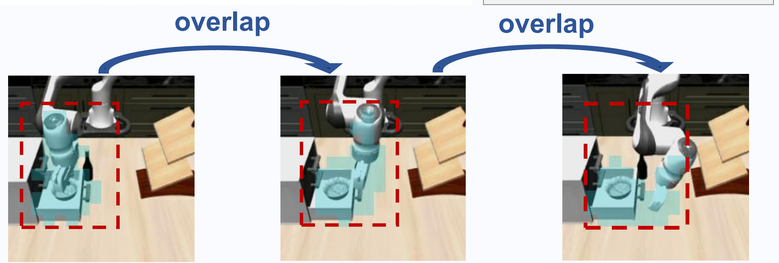


Globally important tokens largely overlap between consecutive inference steps

Overlap Rate

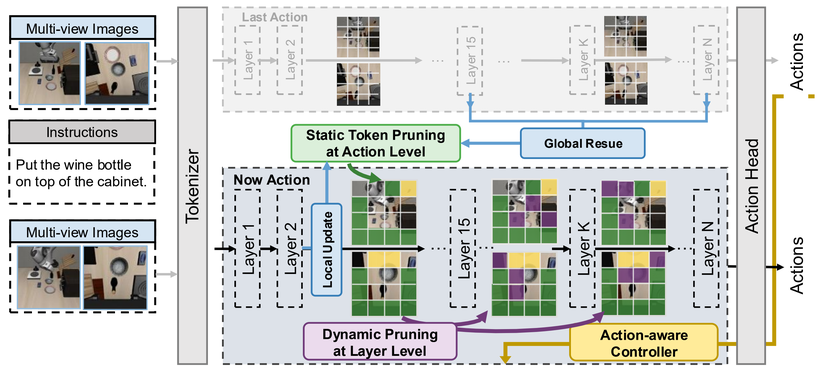
$$\frac{|V_{t-1} \cap V_t|}{|V_t|}$$

70%~90%



Reliable pruning should combine global context from previous steps with local cues from current step.

2. Method: Action-aware two-level pruning



1 Action-level static pruning

- Last inference result serves as draft results. Reuse global attention from middle and deep layers of last step.
- Enhance it with dynamic elements through speed-based frame comparison.
- Add current-step local speculation from the first two layers.
- **Results: prune 60%-70% visual tokens before LLM forward.**

2 Layer-level dynamic pruning

- Dynamically update token importance scores across layers
- Use entropy-based layer confidence and exponential moving averaging.
- Remove redundant tokens progressively through the model.
- **Result: further reduce about 20% computation.**

3 Lightweight action-aware controller

- Distinguish coarse-grained vs. fine-grained actions using end-effector speed.
- Apply more aggressive pruning for coarse actions and conservative pruning for fine actions.
- Maintain robustness during manipulation-critical stages.

2. Method: Action-aware two-level pruning

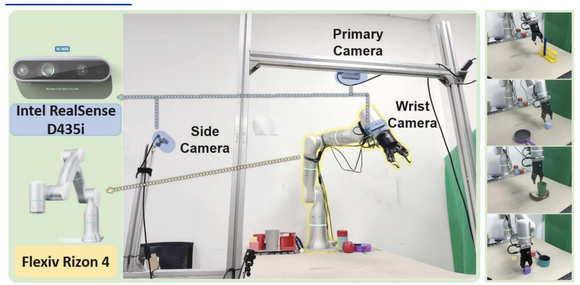
- Three models, three different robot arms
- Two simulation benchmarks and four real world tasks

Simulation

LIBERO benchmark	Success Rate (%)				Avg. Speedup	Avg. SR(%)	FLOPs
Method	Spatial	Object	Goal	Long			
OpenVLA-OFT	97.6%	96.5%	97.9%	94.5%	1.00×	96.6%	100%
+ FastV [ECCV24]	94.6%	95.8%	94.0%	88.8%	1.44×	93.3%	57%
+ DivPrune [CVPR25]	92.4%	91.2%	89.0%	84.8%	1.46×	89.4%	54%
+ SparseVLM [ICML25]	96.8%	94.2%	97.6%	93.6%	1.28×	95.6%	77%
+ VLA-Cache [NIPS25]	99.0%	97.7%	97.4%	93.6%	1.07×	96.9%	83%
+ EfficientVLA [NIPS25]	96.5%	91.1%	96.0%	72.1%	1.52×	88.9%	35%
+ Ours (α=0.8)	97.4%	95.8%	97.7%	93.4%	1.46×	96.1%	43%

SimplerEnv benchmark	Success Rate (%)				Avg. Speedup	Avg. SR(%)	FLOPs
Method	PutCarrot	PutSpoon	StackCube	PutEggplant			
DB-OFT	64.1%	89.2%	30.8%	97.5%	1.00×	70.4%	100%
+ FastV [ECCV24]	59.8%	83.5%	24.0%	94.0%	1.40×	65.3%	56%
+ DivPrune [CVPR25]	60.7%	85.0%	24.8%	96.3%	1.43×	66.7%	54%
+ SparseVLM [ICML25]	62.2%	86.4%	26.8%	96.3%	1.28×	67.9%	75%
+ Ours (α=0.8)	63.8%	88.7%	30.0%	98.0%	1.44×	70.1%	42%

Real robot



Flash

Safe

Easy to employ

Real world task	Success Rate (%)				Avg. Latency (ms)	Avg. Speedup
Method	Pick&Place	TransferTube	PickUpCup	MultiCubeTask		
OpenVLA-OFT	96.7%	85.0%	91.7%	95.0%	187.7	1.00×
Ours	96.7%	82.0%	90.0%	95.0%	110.2	1.70×