CS229: Extending on Speculative Decoding for Vision-Language-Action Models in Robotics

Darrow Hartman, Hadil Owda, Josh Bowden

{darrowrh, hadilo12, jjosh }@stanford.edu

Stanford Computer Science

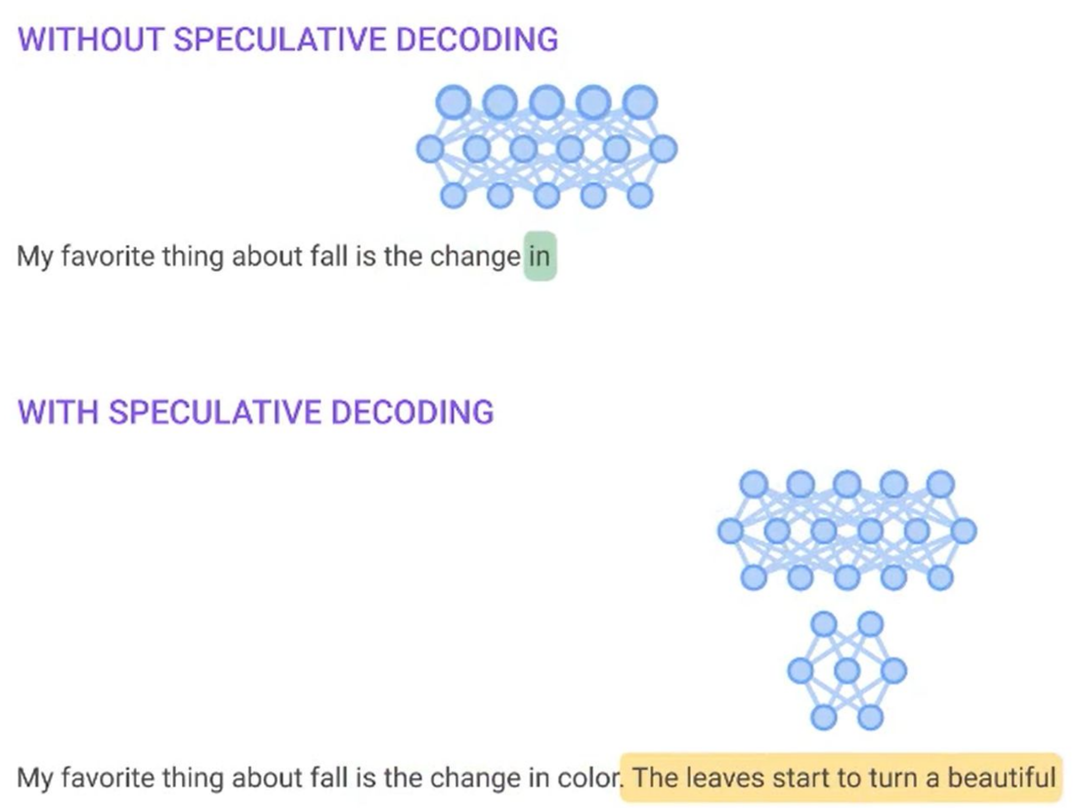
Project Overview

- Motivation: Vision-Language-Action Model (VLA) performance in robotics has improved through the application of speculative decoding and applying state-space models. We were curious about applying both of these techniques to further improve VLA performance.
- Project: We trained a Mamba draft model and a modified roboMamba model, and benchmarked them against SpecVLA's Llama draft model.
- Results: Original paper failed to replicate and Mamba draft model achieved no significant speedup over Llama draft model on average

Background

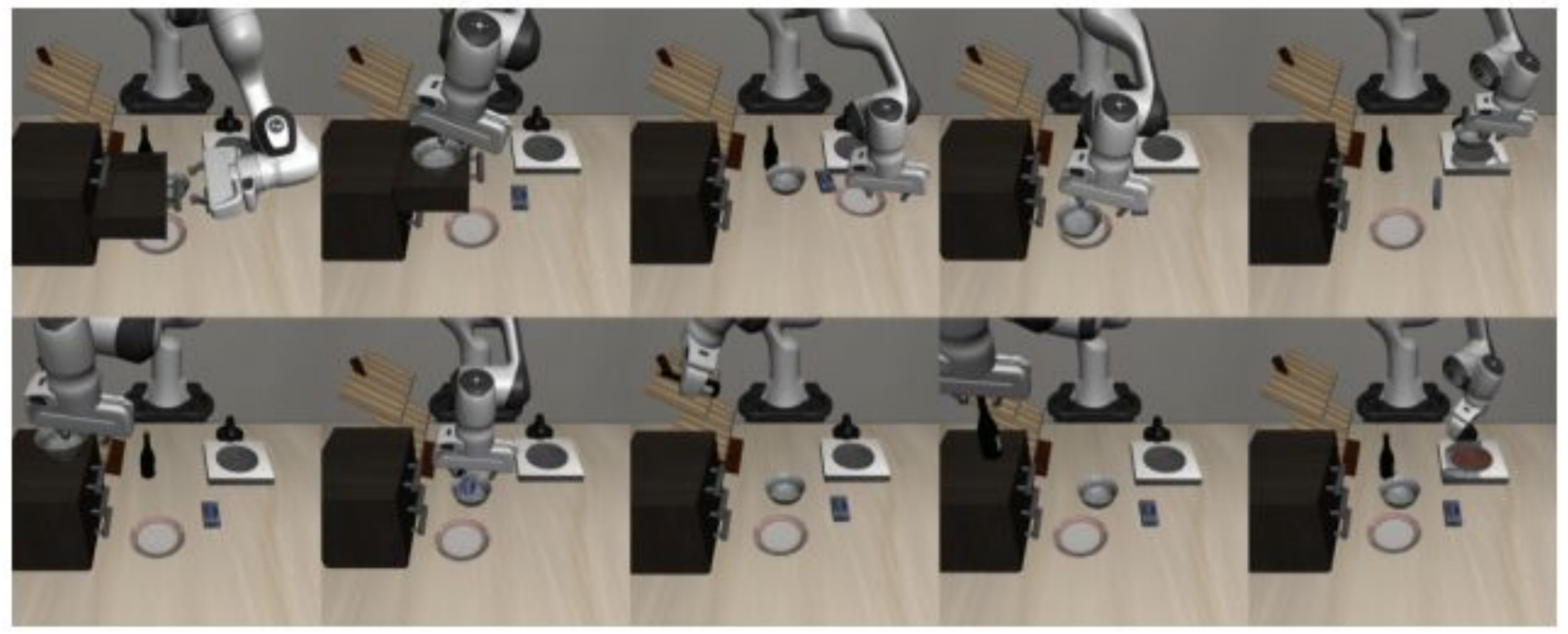
Speculative Decoding:

- Very effective in LLMs with 2-3x speedup on same hardware with identical outputs (Google)
- Small draft model generates candidate sequences during serial autoregressive generation
- Large verifier model processes entire sequence in one forward pass, accepting multiple tokens or falling back to producing one token
- Newly applied to VLAs in Sept 2025 conference paper (SpecVLA, EMNLP 25). They had low success and had to allow the output to shift; ideally the output is the same
- Language / LLM has lots of redundancy; robotics / VLA is more complicated



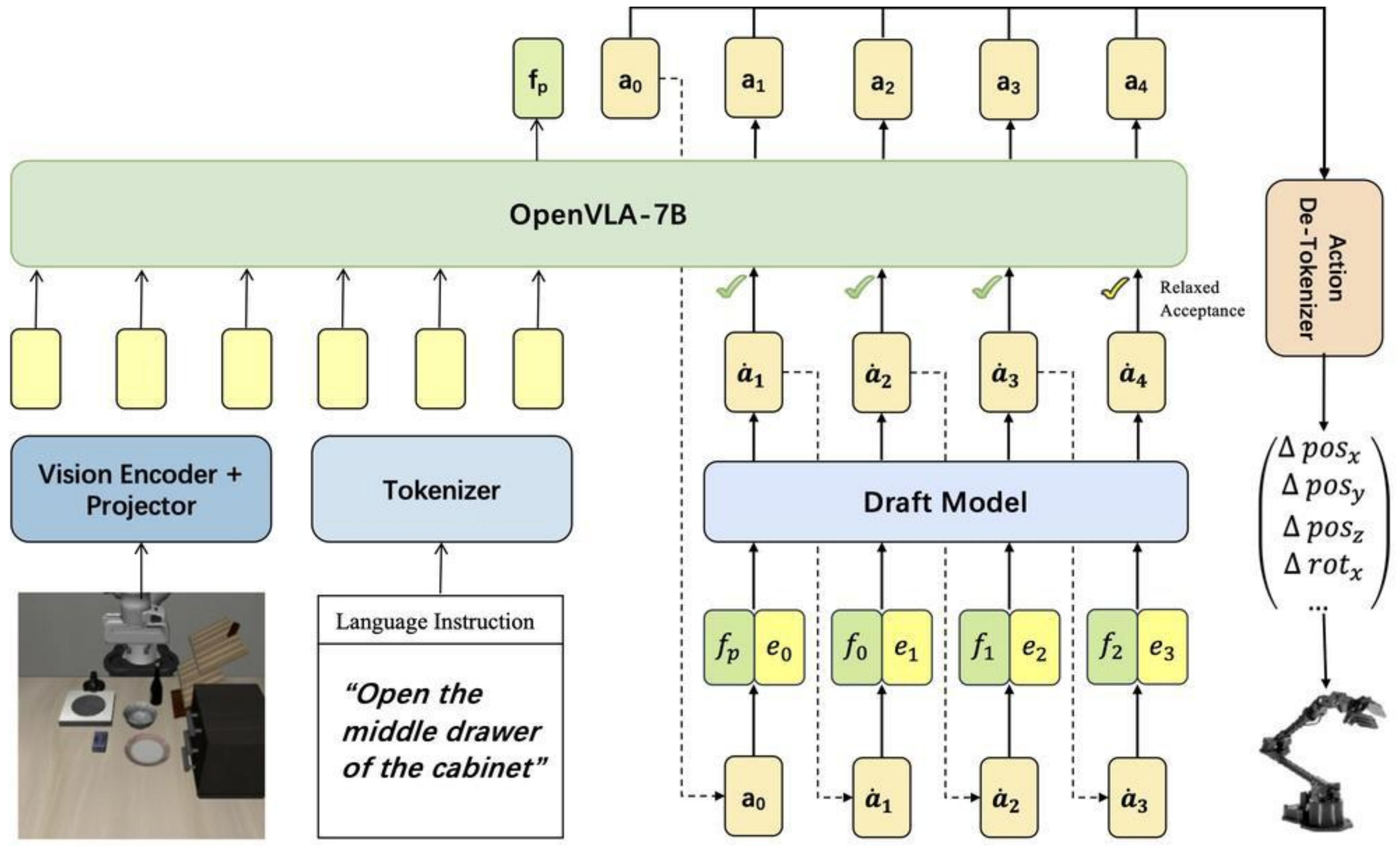
Datasets & Metrics

- We use the LIBERO-GOAL benchmark, which is a kitchen environment with 10 different tasks
- We run OpenVLA, producing observation–language–action triplets. We filter out no-ops
- Each example contains a 256×256 RGB image, a natural-language instruction, OpenVLA vision-encoder hidden states (sequence length × 4096), token embeddings, and a ground-truth discrete action label (7D end-effector control quantized into 256 bins per dimension).
- Our main metric is seconds per episode. Since strict speculative decoding rejects errors from the draft model, overall speed represents the speed and accuracy of the draft model.



Methods & Experiments

- The verifier model is OpenVLA, a 7B-parameter vision-language-action model that autoregressively predicts discrete robot action tokens.
- We evaluate two draft models: (1) a LLaMA-based transformer decoder and (2) a Mamba state-space model, both trained to predict future hidden states
- We train the Mamba model from scratch with model dimension of 4096, state
- dimension of 16, expansion factor of 2, and bf16 instead of fp16 to avoid RNN overflow.
- We integrate our draft model into existing speculative decoding and EAGLE sampling frameworks



Results

Mamba draft model vs replication of SpecVLA autoregressive and draft model approaches

Model / Approach	Train Acc.	Avg. Time (s)	Std. Dev. (s)	Accuracy (%)
Autoregressive	-	23.38	13.01	69%
Speculative (Llama Draft)	97.63%	22.61	12.15	71%
Speculative (Mamba Draft)	96.86%	23.29	12.49	71%

Discussions & Future Research

Discussions:

- We failed to replicate SpecVLA's reported 1.09x speedup, achieving only 1.03x with their Llama draft model.
- Our Mamba draft model was in the same ballpark as their Llama draft model.
- Robot actions are more information-dense than language tokens, making accurate speculation challenging.
- Suggests architectural choice matters less than the fundamental difficulty of predicting robot actions
- Training accuracy may be high due to low diversity of LIBERO task suite

Future Research:

- Analyze Mamba's acceptance length patterns to better understand accuracy-speed trade offs
- Investigate the high variance in per-task performance across models.
- Try finetuning Mamba model that is already pretrained on VLAs (RoboMamba)
- Developing a simplified testing harness that can isolate and benchmark draft models independently of the complex VLA codebase

References

[1] Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces.

[2] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. Openvla: An open-source vision-language-action model.

[3] Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR.

[4] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024. Eagle-2: Faster inference of language models with dynamic draft trees.

[5] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning.

[6] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. 2024. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation.

[7] Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. 2025. Large vlm-based vision-language-action models for robotic manipulation: A survey.

[8] Songsheng Wang, Rucheng Yu, Zhihang Yuan, Chao Yu, Feng Gao, Yu Wang, and Derek F. Wong. 2025. Spec-vla: Speculative decoding for vision-language-action models with relaxed acceptance.