Skewer with Confidence: Adapting RT-1 for Uncertainty-Aware Food Acquisition

Bohan Yang, CS '26

Advisor: Tapomayukh Bhattacharjee

Research Problem

- Developing reliable robotic manipulation models require not just accuracy, but also the ability to know when the model is uncertain.
- Prior work, such as SPANet *[1]*, demonstrates strong performance on food manipulation tasks. However, it lacks the mechanisms for uncertainty estimation, especially in the low-level action parameters.

Key Contributions:

- We finetuned a Vision-Language-Action (VLA) transformer RT-1 for the food skewering task.
- Our model matches the previous bite acquisition benchmark in accuracy (e.g., MAE).
- This VLA model lays the foundation of token-level uncertainty quantification .

Acknowledgements

This research was supported by the Bowers Undergraduate Research Experience (BURE) program during the summer 2025.

Contact:

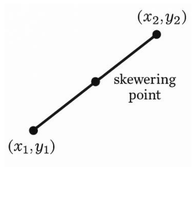
Bohan Yang
Email: by93@cornell.edu

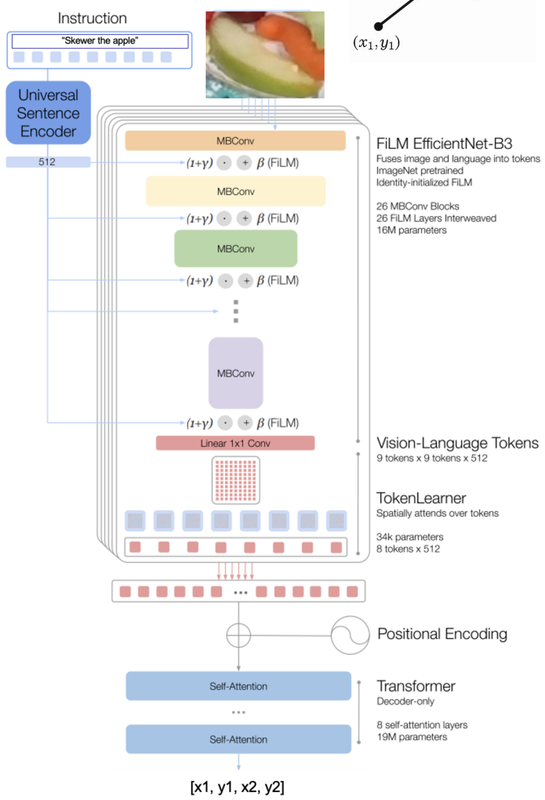
VLA Model & Training

Architecture: Modified RT-1 *[2]*

Dataset: 12,961 (Image, Text, Action Label)

- Finetuned from pretrained RT-1 with two-stage training (5 epochs in total)
- Output: [x1, y1, x2, y2] each discretized into 256 bins
- Loss: Cross-entropy



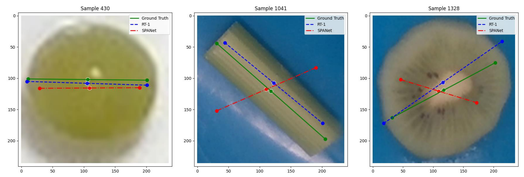


Modified RT-1 Architecture

Results

Metric	RT-1	SPANet
Skewer X Error	0.046 ± 0.053	0.039 ± 0.034
Skewer Y Error	0.051 ± 0.052	0.044 ± 0.037
Orientation Error (°)	56.2 ± 49.9	61.9 ± 43.5
Skewer Point Error	$\mathbf{0.078 \pm 0.064}$	$\mathbf{0.066 \pm 0.041}$

Prediction Error Summary for RT-1 and SPANet.



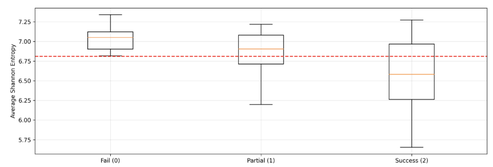
*The "Skewer X Error", "Skewer Y Error", and "Skewer Point Error" are normalized coordinates.

Uncertainty Quantification

We manually labeled ~500 RT-1 predictions into 3 classes: 0=Fail, 1=partial, 2=Success.

For each prediction, we compute the average Shannon entropy across the four coordinate token distributions.

$$H = -\sum_{i=1}^{k} p_i * \log_2(p_i)$$



We find the optimal threshold by maximizing the Youden's J.

$$\text{Youden's } J = \text{Sensitivity} + \text{Specificity} - 1 = \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1$$

Conclusion

We demonstrate that a Vision-Language-Action (VLA) transformer finetuned from RT-1 can match a prior benchmark in action accuracy while enabling token-level uncertainty quantification.

By analyzing entropy distributions over the model output token logits, we can estimate the uncertainty of the model's prediction on the spatial coordinates. This offers interpretable signals for failure detection and confidence-aware decision making.

Our current entropy-based uncertainty metric is preliminary; in future work, we plan to explore and evaluate alternative uncertainty quantification methods to further improve reliability and robustness.

Future Work

- Extend the current VLA model beyond skewering to other manipulation skills such as scooping, twirling, pushing.
- Explore and evaluate alternative uncertainty quantification methods for improved reliability

References

[1] Feng et al., "Robot-Assisted Feeding: Generalizing Skewering Strategies," arXiv:1906.02350, 2019. arXiv. https://doi.org/10.48550/arXiv.1906.02350

[2] Brohan et al., "RT-1: Robotics Transformer for Real-World Control at Scale," arXiv:2212.06817, 2022.