Ask Before You Act: Token-Level Uncertainty for Intervention in Vision-Language-Action Models

Ulas Berk Karli
Department of Computer Science
Yale University
New Haven, USA

Tetsu Kurumisawa
Department of Computer Science
Yale University
New Haven, USA

Tesca Fitzgerald
Department of Computer Science
Yale University
New Haven, USA

***Abstract*—Robots operating in open-ended environments must be able to recognize the limits of their understanding and know when to rely on human input. While vision-language-action (VLA) models such as π_0-FAST provide scalable and expressive policies through next-token prediction, they often lack mechanisms for introspection or fallback under uncertainty. We present a system that leverages token-level uncertainty from a fine-tuned π_0-FAST model to enable *uncertainty-aware human intervention* during robotic manipulation. When prediction uncertainty exceeds a threshold, the robot halts execution and explicitly requests a one-step corrective action from a human operator. We evaluate our system against two baselines—*random intervention* and *no intervention*—and demonstrate that using uncertainty as a trigger improves success and reliability across manipulation tasks. As a supplementary investigation, we also explore a shared control variant that blends human joystick input with model actions based on uncertainty, illustrating an alternate use of model introspection. Our results suggest that token-level uncertainty in VLA models provides meaningful signals for decision arbitration, failure prediction, and adaptive human-robot collaboration.**

I. Introduction

Robots operating in unstructured environments must be able to recognize when their knowledge is insufficient and know how to appropriately seek human input [24, 5]. This ability is critical for them to operate safely and reliably, especially as behavior models increase in complexity and are deployed in increasingly open-ended tasks. Recent vision-language-action (VLA) models offer a promising direction for general-purpose robotic policies, using autoregressive token prediction to flexibly map from observations and language instructions to action sequences. However, these models typically lack mechanisms for introspection [26, 6, 18]; they not only predict the next token without signaling when they are uncertain or likely to fail, but also do not provide feedback on how well they have learned from training data, making it difficult for the robot's human operator to understand what training data they should provide in order to improve the model's performance [17, 4, 3, 2, 22, 15].

Our work takes the first step toward a lifelong learning paradigm for VLA models: instead of collecting training data all at once, we aim to support frequent iteration across (i) training the model, (ii) deploying the model in the wild, (iii) using introspection and intervention to selectively query a human teacher for assistance when it is uncertain, and (iv) using this assistance to improve both immediate task completion and future model updates. In this paper, we focus on the introspection and intervention step. We instantiate token-level uncertainty metrics that are commonly applied to LLMs (entropy and perplexity) in a VLA model, analyze their suitability and reliability in robotics tasks, and demonstrate that intervention at high-uncertainty moments can meaningfully improve model performance. To our knowledge, this is the first use of such metrics for VLA models.

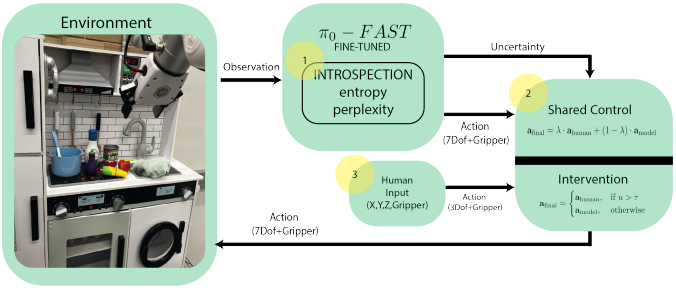

Fig. 1. Overview of our uncertainty-aware VLA system. A fine-tuned π_0-FAST model introspects on its own predictions by computing token-level uncertainty metrics—entropy and perplexity—during action generation (1). When uncertainty is high, the robot either (2) blends human joystick input with model predictions using a continuous shared control scheme, or (3) requests a one-step human intervention via an intuitive Cartesian interface (X, Y, Z, gripper). This design allows both continuous and discrete collaboration modes based on model confidence. Yellow circles highlight key system contributions.

We apply these metrics in two uncertainty-aware control strategies. In the **intervention** setting, the robot pauses execution and explicitly requests help when uncertainty exceeds a threshold, prompting the human to provide a one-step corrective action. In the **shared control** setting, the robot blends its actions with human joystick input in real time, using uncertainty as a dynamic confidence weight. Through empirical evaluation across a range of manipulation tasks, we find that uncertainty-based methods improve task success and reduce unnecessary interventions—but only when the model has seen many training examples on the task. In low-data regimes, uncertainty metrics behave similarly to random triggers, underscoring the importance of model calibration.

Our contributions are:

- An uncertainty-aware intervention system using token-level introspection from a fine-tuned π_0-FAST model.

- A comparative evaluation of uncertainty-based, random, and no-intervention approaches.
- A supplementary shared autonomy framework illustrating broader applications of uncertainty-aware VLA inference.
- A novel xArm manipulation dataset for fine-tuning vision-language-action models.

Together, these results show that token-level uncertainty offers more than just introspection—it is an actionable signal that can inform control, structure collaboration, and guide future learning in robot systems.

II. Related Work

A. Vision Language Action Models

Vision-Language-Action models (VLAs) such as RT-1 [4], RT-2 [3], Octo [13], and OpenVLA [17], are a class of embodied AI systems that integrate visual perception, natural language understanding, and action generation into a single framework. They enable robots to interpret high-level language instructions, perceive the environment through vision, and output low-level control commands to accomplish tasks. The current best performing open-source models are π_0 [2] and π_0-FAST (Frequency-space Action Sequence Tokenization) [22]. Our work is based on the π_0-FAST model due to its performance level, as well as the autoregressive decoding scheme which allows us to employ token level entropy measurements, which are standard model uncertainty metrics in LLM literature [26, 25].

By pre-training on extensive web data and vast amounts[1] of training episodes, VLAs have achieved impressive performance across a variety of sensory inputs and action spaces; yet, even the most advanced models still require post-training fine-tuning in novel Out-of-Distribution (OOD) settings [2]. Therefore, improving task performance through human supervision when using these models in real-world contexts is an active and crucial topic in this field.

B. Uncertainty Metrics in Foundation Models

Entropy captures the spread of a model's output distribution, enabling confidence ranking via differential entropy over the vocabulary [19]. At the token level, high entropy can indicate low-confidence predictions that are likely to be incorrect or hallucinated; for example, Fadeeva et al. [9] use it to detect unreliable spans in LLM outputs. *Semantic* entropy clusters responses by their meanings (using bi-directional entailment) and computes entropy across clusters to reveal confabulations [29, 10]. In training, entropy-regularized objectives, such as masked MLE with self-distillation, help reduce epistemic uncertainty and improve OOD robustness [20].

Perplexity, the exponentiated average negative log-likelihood per token, quantifies model "surprise":

$$\mathrm{PPL}(x) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log_2 p(x_i \mid x_{<i})\right) \tag{1}$$

[1]OpenVLA [17] was trained on 970k real-robot demonstrations drawn from the Open X-Embodiment dataset.

It is directly linked to cross-entropy loss and is a standard proxy for model confidence [12]. Lower perplexity correlates with more coherent outputs and better human-judged quality. It is also effective for OOD detection; Ren et al. [24] show that thresholding sequence perplexity enables abstention from poor outputs and improves summarization and translation performance.

Overall, entropy and perplexity offer complementary quantitative measures of uncertainty of foundation models: entropy provides fine-grained, distribution-aware confidence at the token or sequence level, while perplexity delivers a holistic "surprise" metric over full text. However, comprehensive metrics of uncertainty for robotic foundation models have yet to be developed, due to the higher dimension inputs, model complexity, as well as less well-defined token spaces [11, 14]. In this study, we address this gap by investigating whether uncertainty metrics used for LLMs also apply to VLAs.

C. Shared Control

Shared control systems, which arbitrate human control and robots' autonomous movements and provide assistance based on the operator's inferred goals, improve teleoperation by speeding up task completion and requiring less input [16, 1, 21]. Control *arbitration* determines how authority is divided between the human and robot, and is key to performance and user experience in shared control [21]. Without clear arbitration, mixed-initiative systems risk conflicts, suboptimal actions, or unstable control. Dragan and Srinivasa [8] modeled arbitration as policy blending, showing that the blending parameter α significantly affects stability and task success, with poor choices causing deadlock or failure. Here, we investigate using VLA uncertainty metrics as the arbitration signal.

III. Approach: Uncertainty-Based Policy Intervention and Shared Control

Our key insight is that the entropy and perplexity of predicted action tokens in autoregressive models like π_0-FAST can serve as reliable indicators of confidence. We develop an end-to-end system in which the robot either (1) requests a one-step correction when uncertainty exceeds a threshold, or (2) blends its actions with real-time human input using uncertainty as a dynamic weighting factor. To support this, we compute token-level uncertainty at each inference step of the π_0-FAST model. Since π_0-FAST generates discrete action tokens conditioned on image and language input, these tokens are then decoded into continuous joint-space actions via a learned compression scheme. At each decoding step, the model outputs a distribution over its vocabulary of tokens. We compute the Shannon **entropy** of the token distribution at each step (reflecting distributional spread) and **perplexity** of the selected token (Eq. 1), capturing the model's surprise at its own choice. Because action decoding occurs over a sequence of tokens that do not correspond directly to individual robot actions, we summarize uncertainty across each inference sequence. Specifically, we compute the *mean* entropy and perplexity across all non-padding tokens in a predicted sequence. These

summary values are emitted alongside the decoded actions and serve as introspective signals of model confidence.

A. Threshold Selection

To identify meaningful thresholds for intervention, we analyze uncertainty statistics under in-distribution conditions. We run inference on the entire training set by sampling images at 30hz from each episode in our training set, collecting entropy and perplexity for each. This yields a distribution of expected uncertainty values for the fine-tuned model over our training data. We set thresholds for intervention at the *mean plus one standard deviation* for each metric. This provides a principled yet lightweight approximation of confidence intervals: values above this threshold indicate that the model's confidence is significantly lower than during training. This heuristic ensures that interventions are triggered only when the model deviates from its typical, well-calibrated behavior.

B. Intervention Policy

During execution, we continuously monitor the model's sequence-level uncertainty. If the mean entropy or perplexity of a predicted token sequence exceeds its respective threshold, the robot pauses and explicitly requests assistance from the human operator. The intervention policy is defined as:

$$\mathbf{a}_{\text{final}} = \begin{cases} \mathbf{a}_{\text{human}}, & \text{if } u > \tau \\ \mathbf{a}_{\text{model}}, & \text{otherwise} \end{cases} \tag{2}$$

where u is the model's sequence-level uncertainty (e.g., mean entropy or perplexity), and τ is the threshold for intervention. To facilitate this interaction, we use an Xbox controller through which the human operator (in our experiments, one of the authors) provides a single corrective action in Cartesian space, primarily controlling the end-effector's 3-DoF position and the gripper. When an intervention is triggered, we activate haptic feedback on the controller to notify the operator, and temporarily slow down the control loop to ensure accurate input. The human intervention is recorded and reflected in real-time by the robot, after which the model resumes inference from the updated state. This policy enables targeted assistance at moments of high uncertainty while minimizing unnecessary interruptions to autonomous execution.

C. Shared Control

In addition to discrete interventions, we implement a shared control variant as an exploratory use of model uncertainty. Previous research [21] has shown that a crucial component of shared control is control arbitration. In this mode, we treat the model as the intention inference engine, and then use the model's uncertainty of its output as the arbitration parameter. The robot fuses its predicted action $\mathbf{a}_{\text{model}}$ with a real-time human joystick action $\mathbf{a}_{\text{human}}$ using an uncertainty-aware blending rule:

$$\mathbf{a}_{\text{final}} = \lambda \cdot \mathbf{a}_{\text{human}} + (1 - \lambda) \cdot \mathbf{a}_{\text{model}} \tag{3}$$

where λ is computed as a normalized function of sequence-level uncertainty (e.g., scaled entropy). High model uncertainty increases human influence, whereas low uncertainty preserves autonomous execution. This shared control system is used to explore continuous human-model arbitration as a supplementary alternative to binary intervention.

TABLE I
SUMMARY OF TRAINING DATASET BY TASK TYPE.

Task Type	Count	Percentage
Lift	173	48.3%
Put	156	43.6%
Knock	14	3.9%
Stir	8	2.2%
Wipe	7	2.0%

IV. EXPERIMENTS

A. Preliminaries: Fine-Tuning a VLA Model for Real-World Uncertainty Evaluation

To investigate token-level uncertainty in VLAs during robotic manipulation, we first fine-tuned a pre-trained VLA policy on our own robot setup. We selected π_0-FAST as the base model due to its demonstrated performance across diverse manipulation tasks [23]. To create a realistic yet tractable testbed, we used a toy kitchen environment (shown in Fig. 1) inspired by prior VLA datasets such as BridgeV2 and Open-X [27, 7].

Our robot platform consists of an xArm7 manipulator, which was not part of the original π_0-FAST training distribution. To adapt the model, we collected a new dataset using a custom GELLO controller [28] that provides kinesthetic teaching and joystick teleoperation for xArm. The dataset includes demonstrations across five high-level task types: `Lift`, `Put`, `Knock`, `Wipe`, and `Stir`. All actions were recorded at 30 Hz and tokenized using the original π_0-FAST discretization pipeline.

Table I summarizes the distribution of task. In total, our dataset consists of 80,419 action steps spanning 5 unique task categories and 17 unique tasks. We performed full-parameter fine-tuning on the pre-trained π_0-FAST model using this dataset and selected a checkpoint with low training loss and strong performance during physical robot testing. This checkpoint serves as the foundation for our uncertainty-aware control policies.

B. Experimental Goals and Setup

We evaluate our intervention and shared control systems on a diverse set of robot manipulation tasks to investigate three key questions: (1) Can token-level uncertainty metrics such as entropy and perplexity reliably trigger helpful human interventions? (2) How do these policies compare to random and no-intervention baselines? (3) Can shared control offer a practical, continuous alternative for leveraging uncertainty in human-robot collaboration?

Tasks were selected to span a range of difficulties and training coverage. Some, like `knock the tomato sauce can`, were rarely observed during training and represent undertrained scenarios. Others, such as `lift the corn` and

`put the corn in the pot`, were consistently challenging for the base model. Well-represented tasks like `lift the eggplant` and `put the pot in the sink` serve as high-confidence benchmarks.

Our policy evaluation includes entropy- and perplexity-based interventions, where the robot halts and requests a one-step corrective action when uncertainty exceeds a predefined threshold. These are compared against two baselines that request interventions at random intervals: a conservative policy with probability $p = 0.1$ and an aggressive policy with $p = 0.5$. We also evaluate a shared control strategy where robot actions are blended with human joystick input based on uncertainty, and a no-intervention baseline where the model operates fully autonomously.

C. Results

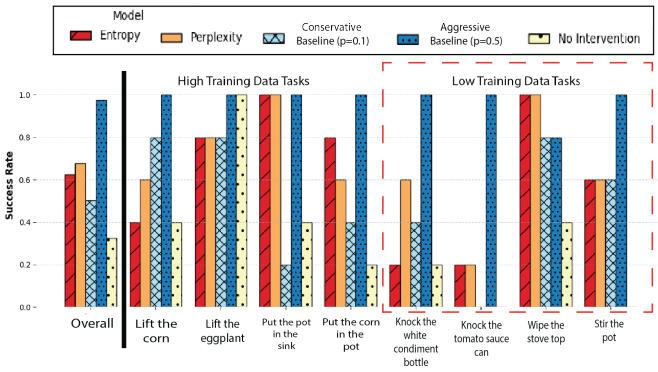

Fig. 2. Success rate per task and intervention policy for low-data and challenging tasks.

a) Performance on Low-Data and Challenging Tasks.: Several tasks have low representation in the data set: `knock`, `wipe`, and `stir`. We observed that some tasks were more challenging for the model: `lift the corn`, `put the corn in the pot`, and `put the pot in the sink`. We have one easy task as a control (`lift the eggplant`). Figure 2 reports task success rates across low-data and challenging settings. Uncertainty-based interventions outperform the conservative random baseline ($p = 0.1$) and often match the aggressive random baseline ($p = 0.5$), while requiring fewer interventions on well-represented tasks. Overall, the aggressive random baseline reaches near perfect success rate due to excessive intervention request. For highly underrepresented tasks, however, performance of conservative random and uncertainty-based methods converge. This indicates that when the model does not have enough training examples, its uncertainty metrics act like random signals.

To evaluate intervention efficiency, we define a metric as the ratio of task success to the number of interventions. Results are shown in Figure 3.

b) Validation on High-Data Tasks.: Given that entropy and perplexity appear most informative in settings with sufficient training data, we also tested their effectiveness on tasks that are well-represented in our training data. As shown in

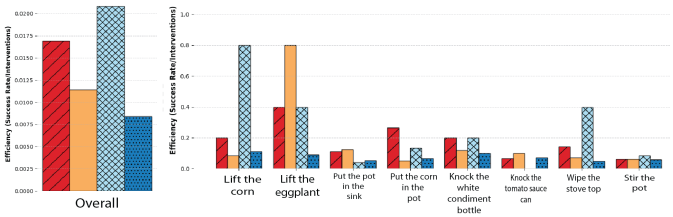

Fig. 3. Efficiency per task and intervention policy for low-data and challenging tasks.

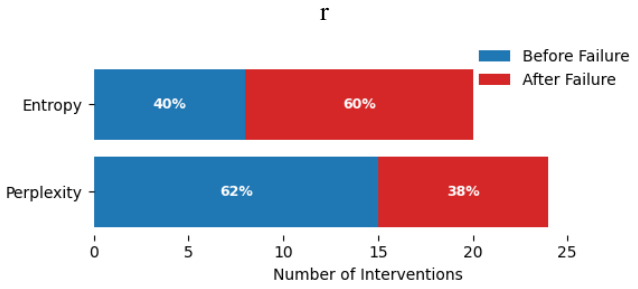

Fig. 4. Timing of first intervention in an episode: before vs. after failure.

Figure 5-Left, both uncertainty metrics maintain high success rates while reducing unnecessary interventions.

Efficiency results for the same tasks are shown in Figure 5-Right. Uncertainty-based policies, especially entropy, maintain strong performance while using fewer interventions.

c) Intervention Timing.: To understand when interventions occur, we examined the first intervention in each episode and labeled it as either *before failure* or *after failure*. We found that entropy tends to trigger interventions after failures, while perplexity is more likely to request assistance beforehand (see Fig. 4). These results suggest that perplexity may serve as a more proactive indicator of model uncertainty during execution.

d) Shared Control vs. No Intervention: We also tested whether token-level uncertainty can support continuous human-robot collaboration without discrete pauses. Figure 6 compares the shared control strategy—using either entropy or perplexity as a blending weight—against the no-intervention baseline. Shared control improves performance across nearly all tasks, without increasing task completion time, demonstrating the feasibility of using uncertainty not just for failure recovery, but also for proactive safety during execution.

These results show that token-level uncertainty enables a flexible spectrum of confidence-aware control strategies, including both discrete interventions and continuous shared autonomy.

V. Discussion

a) Uncertainty helps—but only when calibrated.: Our experiments reveal both the strengths and limitations of using token-level uncertainty metrics—entropy and perplexity—as control signals for robot intervention. In tasks where the

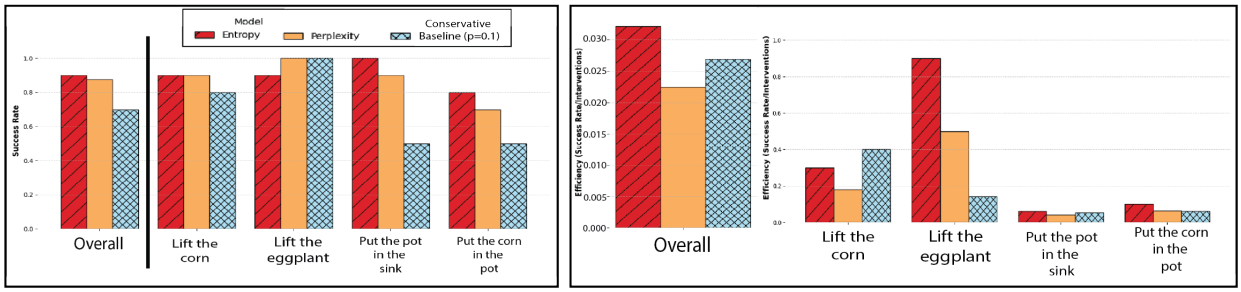

Fig. 5. Left: Success rate per well-trained task and intervention policy. Right: Efficiency per task and intervention policy for high-data tasks.

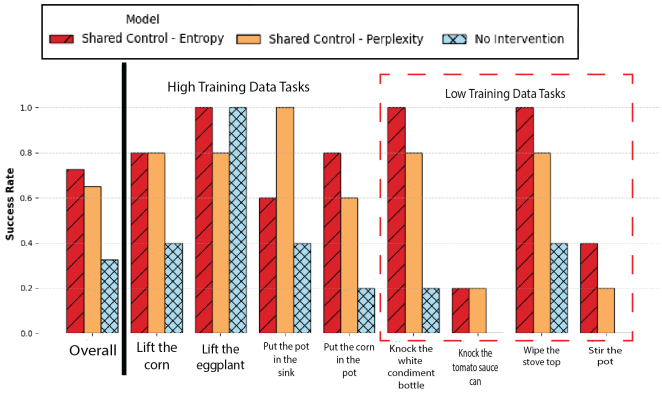

Fig. 6. Success rate per task for shared control and no-intervention baseline.

model had moderate or high training exposure, these uncertainty metrics improved performance by selectively requesting human input. However, in low-data regimes, such as rare or underrepresented task-object combinations, entropy and perplexity behaved similarly to random intervention policies. This highlights a key limitation: when the model is poorly calibrated, its uncertainty estimates lose discriminative power and cannot reliably signal recoverable states.

b) Entropy and perplexity behave differently over time.: During deployment, we also noticed consistent behavioral differences between entropy- and perplexity-based interventions. Entropy tended to rise later in execution, often triggering after an error had begun unfolding, while perplexity sometimes spiked earlier, ahead of failure. While not rigorously quantified, this observation suggests that entropy may reflect distributed indecision across a sequence, while perplexity reacts more to localized surprise. Understanding and formalizing this difference remains an interesting open question.

c) Uncertainty as a stabilizing signal.: Uncertainty-based intervention not only corrected specific errors but also helped reposition the robot into more confident areas of its action distribution. Human input guided the robot out of ambiguous regions and enabled more stable continuation. In shared control, we observed similar stabilizing effects without needing to pause execution: the robot increasingly deferred to the human in high-uncertainty states. Together, these mechanisms demonstrate the broader utility of model introspection as a tool for adaptive control.

d) Toward active learning from intervention.: While our current system treats intervention as a purely reactive process, it also opens the door to active learning. Each human correction offers a targeted demonstration for a high-uncertainty state. Future extensions could convert these interventions into labeled training examples, enabling the robot to incrementally improve its policy in regions where it is least confident.

VI. Conclusion

We introduced a system for introspective robot control that leverages token-level uncertainty—specifically entropy and perplexity—from a fine-tuned vision-language-action (VLA) model. By monitoring these metrics during inference, our system dynamically modulates robot behavior through either discrete intervention or continuous shared control. Our experiments show that when the model has been sufficiently trained on relevant tasks, these uncertainty signals can improve task success while reducing the frequency of unnecessary human involvement. While we observed that entropy tends to trigger reactive interventions and perplexity more often anticipates failure, this distinction merits further quantitative study.

Beyond control, our system also offers a natural entry point for active learning. Each time the robot expresses high uncertainty and receives a corrective demonstration from a human, it uncovers a region of its policy space where it lacks confidence. These moments could be used to automatically generate new training data, enabling continual fine-tuning targeted at the model's weakest regions. By closing the loop between introspection and adaptation, future systems could use uncertainty not only to stay safe at test time, but also to improve autonomously over time.

This work lays the foundation for scalable, model-agnostic uncertainty-aware control in robotic systems. Future directions include formalizing the entropy/perplexity dynamics across tasks, learning adaptive thresholds for intervention, and integrating intervention data into continual or curriculum-based fine-tuning pipelines. In doing so, we move closer to collaborative robots that know when they need help—and learn from it.

References

[1] Reuben M. Aronson and Elaine Schaertl Short. Intentional user adaptation to shared control assistance. pages

4–12, Boulder, CO, USA, 2024. IEEE. ISBN 979-8-3315-1698-7.

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *CoRR*, abs/2410.24164, 2024. doi: 10.48550/ARXIV.2410.24164.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong T. Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: robotics transformer for real-world control at scale. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu, editors, *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi: 10.15607/RSS.2023.XIX.025.

[5] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Anal.*, 71:102062, 2021. doi: 10.1016/J.MEDIA.2021.102062.

[6] Alejandra Ciria, Guido Schillaci, Giovanni Pezzulo, Verena V. Hafner, and Bruno Lara. Predictive processing in cognitive robotics: A review. *Neural Computation*, 33:1402–1432, April 2021. ISSN 0899-7667. doi: 10.1162/neco_a_01383.

[7] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller,

Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[8] Anca D. Dragan and Siddhartha S. Srinivasa. A policy-blending formalism for shared control. *Int. J. Robotics Res.*, 32(7):790–805, 2013. doi: 10.1177/0278364913490324.

[9] Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9367–9385. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.558.

[10] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630, 2024. doi: 10.1038/S41586-024-07421-0.

[11] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, Brian Ichter, Danny Driess, Jiajun Wu, Cewu Lu, and Mac Schwager. Foundation models in robotics: Applications, challenges, and the future. *CoRR*, abs/2312.07843, 2023. doi: 10.48550/ARXIV.2312.07843.

[12] Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7764–7771. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6280.

[13] Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In Dana Kulic, Gentiane Venture, Kostas E. Bekris, and Enrique Coronado, editors, *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. doi: 10.15607/RSS.2024.XX.090.

[14] Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Varma Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Shibo Zhao, Yu Quan Chong, Chen Wang, Katia P. Sycara, Matthew Johnson-Roberson, Dhruv Batra, Xiaolong Wang, Sebastian A. Scherer, Zsolt Kira, Fei Xia, and Yonatan Bisk. Toward general-purpose robots via foundation models: A survey and meta-analysis. *CoRR*, abs/2312.08782, 2023. doi: 10.48550/ARXIV.2312.08782.

[15] Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In Karen Liu, Dana Kulic, and Jeffrey Ichnowski, editors, *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR, 2022. URL https://proceedings.mlr.press/v205/ichter23a.html.

[16] Shervin Javdani, Siddhartha S. Srinivasa, and J. Andrew Bagnell. Shared autonomy via hindsight optimization. In Lydia E. Kavraki, David Hsu, and Jonas Buchli, editors, *Robotics: Science and Systems XI, Sapienza*

University of Rome, Rome, Italy, July 13-17, 2015, 2015. doi: 10.15607/RSS.2015.XI.032. URL http://www.roboticsproceedings.org/rss11/p32.html.

[17] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Paul Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR, 2024. URL https://proceedings.mlr.press/v270/kim25c.html.

[18] Kaiqu Liang, Zixu Zhang, and Jaime F. Fisac. Introspective planning: Aligning robots' uncertainty with inherent task ambiguity. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8451a20c5a7e0ee5671dda28f7daf7f3-Abstract-Conference.html.

[19] Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. Uncertainty quantification for in-context learning of large language models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3357–3370. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.184.

[20] Tingkai Liu, Ari S. Benjamin, and Anthony M. Zador. Token-level uncertainty-aware objective for language model post-training. *CoRR*, abs/2503.16511, 2025. doi: 10.48550/ARXIV.2503.16511.

[21] Dylan P. Losey, Craig G. McDonald, Edoardo Battaglia, and Marcia K. O'Malley. A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction. *Applied Mechanics Reviews*, 70(1), January 2018. ISSN 2379-0407. doi: 10.1115/1.4039145.

[22] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: efficient action tokenization for vision-language-action models. *CoRR*, abs/2501.09747, 2025. doi: 10.48550/ARXIV.2501.09747.

[23] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models, 2025. URL https://arxiv.org/abs/2501.09747.

[24] Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=kJUS5nD0vPB.

[25] Arindam Sharma and Cristina David. Assessing correctness in llm-based code generation via uncertainty estimation. *CoRR*, abs/2502.11620, 2025. doi: 10.48550/ARXIV.2502.11620.

[26] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *CoRR*, abs/2412.05563, 2024. doi: 10.48550/ARXIV.2412.05563.

[27] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale, 2024. URL https://arxiv.org/abs/2308.12952.

[28] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators, 2023.

[29] Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. *CoRR*, abs/2503.00172, 2025. doi: 10.48550/ARXIV.2503.00172.