ROBOTIC STEERING: MECHANISTIC FINETUNING OF VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Action Models (VLAs) promise to extend the remarkable success of foundation models in vision and language to robotics. Yet, unlike those models, usable VLAs for robotics require finetuning to contend with complex physical factors like robot embodiment, environment characteristics, and spatial relationships. Current fine-tuning methods adapt the same set of parameters regardless of the visual, linguistic, and physical characteristics of a particular task. Inspired by functional specificity in neuroscience, we hypothesize that it is *more effective to fine-tune components of model representations specific to a given task.* In this work, we introduce **Robotic Steering**, a novel mechanistic finetuning approach that identifies task-specific representations in the attention-head space to selectively adapt VLAs. In particular, we use few-shot examples to identify and selectively finetune only the VLA attention heads that align with the specific physical, visual, and linguistic requirements of a task. Through comprehensive on-robot evaluations using a Franka Emika robot arm, we demonstrate that Robotic Steering matches or outperforms full-head LoRA across all tested tasks. Crucially, Robotic Steering demonstrates superior robustness under environmental and task variations compared to standard LoRA finetuning, while enabling faster, more compute-efficient, and interpretable experimentation. Grounded in mechanistic interpretability, Robotic Steering offers a controllable, efficient, and generalizable framework for adapting VLAs to the diverse physical requirements of robot tasks.

1 INTRODUCTION

> *"It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail."*
>
> — Abrahan Harold Maslow, *The Psychology of Science* [46]

Vision-Language-Action (VLA) models represent an emerging paradigm that extends foundation models to robotics by applying next token prediction across vision, language, and physical action spaces [37, 49, 62, 70]. While large-scale robotic datasets [12, 14, 36] have enabled unprecedented training scales, VLAs have yet to achieve the impressive generalization of language and vision-language models. Unlike those models that demonstrate remarkable zero-shot adaptation, VLAs require targeted finetuning for each specific deployment environment, establishing a paradigm where practitioners must adapt models to match the exact specifications of their intended task.

This reality raises a philosophical question: what constitutes a "task" in robotics? A seemingly straightforward manipulation objective such as picking up a mug can have many physical instantiations when considering real-world perturbations [2, 22] such as a camera position, the color of the mug, the table height, or even variations of the robot initial position by a few centimeters. Unlike vision and language domains where tasks have clear boundaries, robotics operates in a continuous space of physical variations where the slightest environmental perturbation can fundamentally alter the required model behavior. We propose that few-shot expert demonstrations better specify what a robotic "task" is, as they contain the valuable physical information inextricably linked to the task definition. Unlike linguistic descriptions alone, these demonstrations encode the physical properties

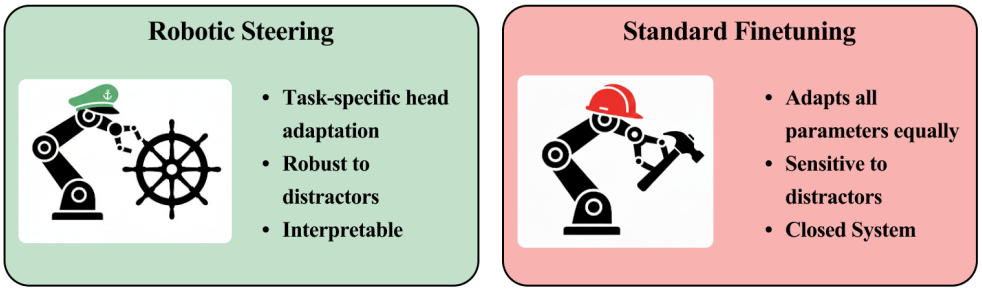


Figure 1: **Robotic Steering enables efficient task adaptation** by finetuning specific heads of a vision-language-action (VLA) model. Standard finetuning trains all parameters, and by intelligently selecting heads to finetune, we find Robotic Steering to be more interpretable and robust to distractors.

of the deployment scenario: the exact angle the robot grasps from, how cluttered the workspace is, what lighting conditions exist, and countless other factors that determine successful execution.

Given task specification through few-shot demonstrations, the key challenge becomes: how can we effectively make use of these demonstrations to learn an embodied task efficiently? Current finetuning methods like LoRA [31] adapt the *same set of parameters regardless of the specific requirements of each task*. In contrast, we take inspiration from functional specificity in neuroscience, which suggests that certain brain regions are specialized for particular tasks [19, 35], and from mechanistic interpretability in machine learning, which has shown that specific attention heads in transformers encode distinct capabilities [27, 50]. Building on these insights, we introduce a novel paradigm: using the few-shot demonstrations themselves to identify which attention heads encode the task-relevant representations, then selectively finetuning only those components. This approach recognizes that different tasks engage different model capabilities, for example grasping from above requires different visual and spatial reasoning than pushing sideways, and adapts the model accordingly.

We introduce Robotic Steering, the first approach to leverage mechanistic interpretability for finetuning task-specific representations of VLAs. Our method consists of three steps, each addressing a key challenge in VLA adaptation. First, we perform semantic attribution to identify task-relevant attention heads. Given a set of few-shot demonstrations of a task, we extract activations from each attention head as the base model performs a forward pass on the examples. We then select heads whose activations perform best on a lightweight k-NN regression task to predict the ground truth actions for the examples. By identifying these task-specific heads, we can achieve more precise adaptation than uniformly finetuning all parameters. Our second step is to freeze the visual encoder, action expert, and LLM backbone while applying targeted finetuning to only the queries and MLP parameters associated with selected heads using LoRA adaptors. Finally, the resulting model deploys as a standard checkpoint without additional overhead. Unlike other mechanistic approaches that require activation interventions during runtime, our finetuned weights integrate seamlessly into existing VLA deployment pipelines. An overview is shown in Figure 1 and Figure 2.

We summarize the main contributions of our work: (i) We introduce Robotic Steering, the first method combining mechanistic interpretability with robotic finetuning for controllable adaptation through semantic attribution of attention heads; (ii) Through comprehensive on-robot evaluations using a Franka Emika robot arm, we demonstrate that Robotic Steering matches or outperforms full-head LoRA across all tested tasks while requiring less runtime and fewer parameters; (iii) Our approach exhibits superior generalization under environmental distractors, including variations in lighting, object properties, and scene configurations, compared to standard finetuning methods; (iv) We provide a practical framework producing standard model checkpoints deployable without additional inference overhead, making mechanistic finetuning accessible for real-world robotic systems.

2 Related Work

Few-Shot Adaptation in Vision-Language-Action Models. Large Language Models (LLMs) [3, 34, 54, 64] and Large Multimodal Models (LMMs) [1, 4, 42, 43, 51, 60, 61] have demonstrated

remarkable capabilities through large-scale pretraining and causal token prediction. Vision-Language-Action models (VLAs) represent the current frontier of robot policy learning [15, 37, 55, 62, 70] and is enabled by large-scale datasets [12, 14, 36]. This scale of training has demonstrated generalization across embodiments and tasks. The state-of-the-art π-series models—π_0 [10] and $\pi_{0.5}$ [53]—use flow matching for continuous action generation along with large-scale data to achieve impressive zero-shot transfer. Despite these advances, VLAs struggle with few-shot adaptation to new environments.

Researchers have explored various few-shot techniques: in-context learning approaches [45, 58, 67] condition on demonstrations without weight updates but face context limitations; parameter-efficient methods [25, 31, 32, 39, 44] and specialized adaptations [38, 57] reduce trainable parameters; meta-learning [20, 23, 68] and behavior retrieval [17, 40, 66] enable rapid adaptation given access to prior data. However, these methods operate at the level of entire weight matrices without considering which components encode physical reasoning. Thus, they lack interpretability and fail to leverage VLAs' structured representations, motivating our mechanistic approach. We also note other work in steering in robotics focuses on guiding the action denoising process of diffusion policies [16, 48] or designing inference-time action sampling metrics [48]. Instead, our work takes a mechanistic approach that more selectively finetunes a VLA.

Mechanistic Interpretability. Recent advances in mechanistic interpretability have revealed how model behavior can be precisely manipulated through internal representations. Early research [8, 9, 69] established frameworks for understanding semantic encoding in neural networks, while activation steering methods [52, 59, 65] demonstrated parameter-free behavior modification. The discovery of specialized components like induction heads [50] and task-specific neurons [28] led to task vector abstractions [26, 63], with parallel work on sparse autoencoders [13] and superposition [18] providing tools for decomposing representations.

An emerging line of work leverages few-shot mechanistic interpretability for model adaptation through task vector methods [11, 29, 33, 47], which concentrate task-relevant information in specific attention heads or activation subspaces for efficient parameter-free adaptation. Research in multimodal representations has revealed how vision-language models structure cross-modal concepts through multimodal neurons [24], mechanistic understanding [56], text-based decomposition [5, 21], and knowledge localization [6, 7]. The comprehensive survey by Lin et al. [41] provides a broader overview of these approaches. While these methods have succeeded in language and vision domains, our work is the first to apply mechanistic interpretability to vision-language-action models, leveraging these insights to identify and adapt components responsible for physical reasoning in robotic control.

3 METHODS

In this section, we present Robotic Steering, our approach for enabling finetuning of task-specific components of Vision-Language-Action models through mechanistic interpretability. Our method identifies and selectively finetunes attention heads that encode task-relevant physical reasoning, allowing VLAs to learn new capabilities while preserving existing ones. We begin with preliminaries on VLA architectures, followed by our three-step approach: (1) identifying task-relevant attention heads through k-NN regression, (2) selective finetuning of identified components, and (3) standard inference with finetuned weights.

3.1 PRELIMINARIES

Vision-Language-Action Models. VLAs extend the transformer architecture to robotic control by processing visual observations and language instructions to predict continuous action vectors. Given an observation o_t consisting of image frames and optional language instruction, a VLA predicts an action vector $a_t \in \mathbb{R}^d$ containing control values (e.g., joint velocities, gripper commands). Modern VLAs like π_0 [10] formulate this as a conditional generation problem, where actions are produced through autoregressive token prediction or flow matching. The model processes inputs as a sequence of visual tokens, language tokens, and robot state information, combining multimodal information for action prediction.

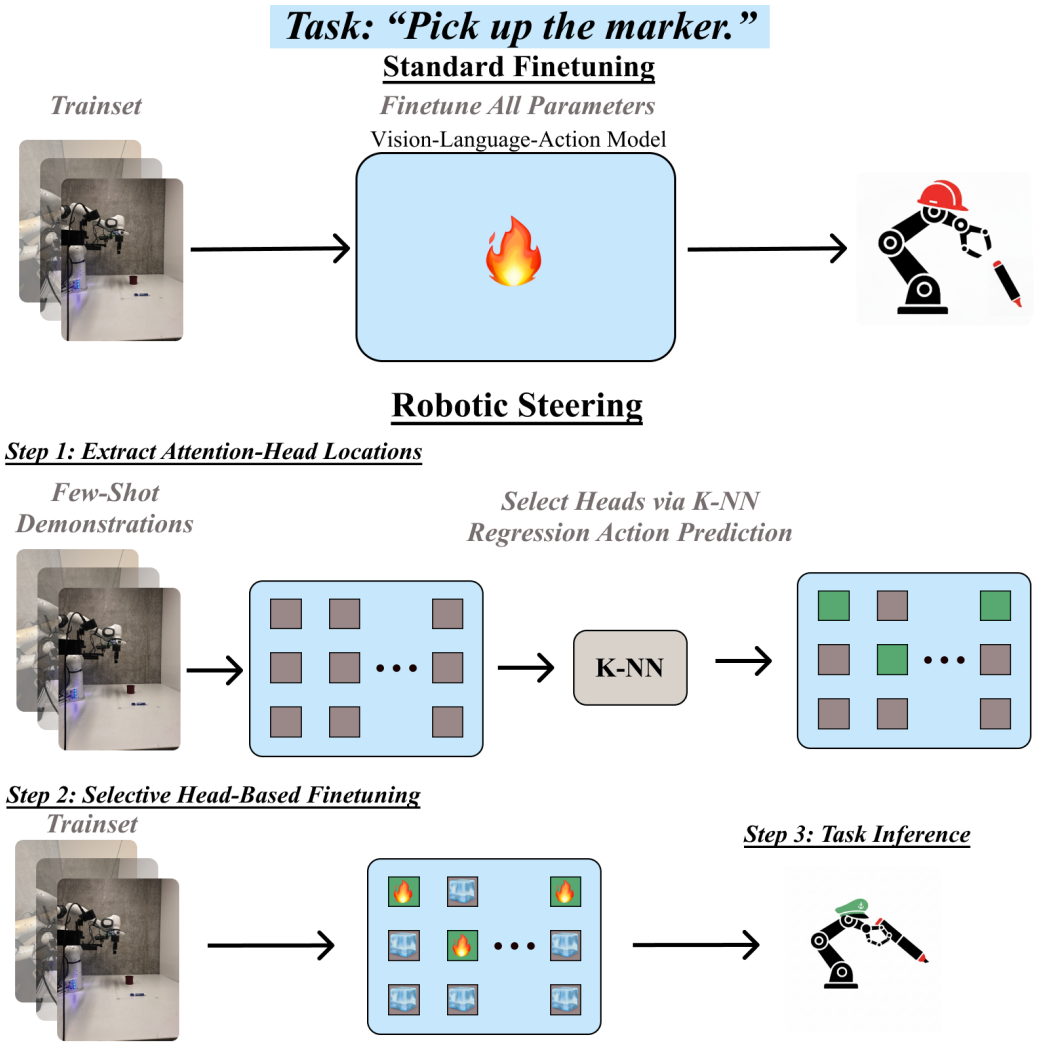


Figure 2: **Method.** Robotic Steering enables targeted adaptation of VLAs by detecting attention heads encoding task-relevant information, finetuning only these components, and reusing the updated model for standard inference.

Multi-Head Attention. For a transformer with L layers and H attention heads per layer, each head (l, h) computes:

$$\mathbf{h}_l^h(x_i) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V \tag{1}$$

where Q, K, V are query, key, and value projections. For action prediction in VLAs, we focus on activations at the final token position $\mathbf{h}_l^h(x_T)$, which aggregates information across the entire sequence.

3.2 Step 1: Identifying Task-Relevant Attention Heads

Our key insight is that within a VLA's attention mechanism, specific heads naturally specialize in encoding physical concepts relevant to particular manipulation tasks. We identify these heads through their ability to retrieve examples with similar action patterns.

Extracting Head Activations. Suppose we are given a frozen VLA and few-shot demonstrations $\mathcal{D} = \{(\tau_1, a_1), (\tau_2, a_2), \ldots, (\tau_N, a_N)\}$, where each trajectory τ_i consists of T timesteps. Each timestep t contains the VLA's input observation: visual tokens from camera images, language tokens

from task instructions, and robot state information (e.g., joint angles). The corresponding $a_i \in \mathbb{R}^{T \times d}$ are action vectors across all timesteps.

For each timestep t in trajectory τ_i, we extract the attention vector $\mathbf{h}_l^h(\tau_i^t)$ for every head (l, h). Importantly, we work at the timestep level rather than trajectory level—each timestep becomes an individual example in our retrieval set.

k-NN Regression for Head Evaluation. To evaluate each head's relevance, we assess its ability to retrieve timesteps with similar actions. The intuition is that if a head's representation groups together observations that require similar physical actions, then this head encodes task-relevant features worth finetuning. In order to make head selection more efficient, we employ the keyframe extraction approach suggested in [67]. Functionally, however, the approach is identical with or without this step. More details can be found in Section A.2 of the Supplementary material.

For a query observation q from trajectory τ_i at timestep t:

We first find the k nearest neighbor timesteps from all other trajectories based on cosine similarity in head (l, h)'s representation space:

$$\mathcal{N}_k^{l,h}(q) = \text{top-}k \left\{ \frac{\mathbf{h}_l^h(q) \cdot \mathbf{h}_l^h(\tau_j^s)}{\|\mathbf{h}_l^h(q)\| \|\mathbf{h}_l^h(\tau_j^s)\|} \right\}_{j \neq i, s} \tag{2}$$

Second, we predict the action by averaging the actions of retrieved neighbors:

$$\hat{a}_t^{l,h} = \frac{1}{k} \sum_{(\tau_j^s) \in \mathcal{N}_k^{l,h}(q)} a_j^s \tag{3}$$

Finally, we compute the head's score as the mean squared error across all queries:

$$\text{score}(l, h) = \frac{1}{|\mathcal{D}| \cdot T} \sum_{\tau_i \in \mathcal{D}} \sum_{t=1}^{T} \|\hat{a}_t^{l,h} - a_i^t\|^2 \tag{4}$$

We select the top-m heads with lowest scores:

$$\mathcal{H}_{\text{task}} = \{(l, h) \mid \text{score}(l, h) \text{ is among } m \text{ lowest scores}\} \tag{5}$$

These heads learn representations that effectively map task-specific observations to other observations requiring similar actions within the few-shot demonstration trajectories, making them ideal candidates for task-specific finetuning.

3.3 Step 2: Selective Finetuning with LoRA

Having identified task-relevant heads $\mathcal{H}_{\text{task}}$, we perform targeted finetuning while preserving the model's general capabilities.

Sparse Parameter Updates. We freeze all model components except the query projections of selected heads. For each head $(l, h) \in \mathcal{H}_{\text{task}}$, we apply Low-Rank Adaptation (LoRA) [31]:

$$W_Q'^{l,h} = W_Q^{l,h} + B^{l,h} A^{l,h} \tag{6}$$

where $B^{l,h} \in \mathbb{R}^{d \times r}$ and $A^{l,h} \in \mathbb{R}^{r \times d}$ are low-rank matrices with rank $r \ll d$. We also finetune the MLP layers associated with the selected attention blocks.

Training Objective. Our approach is flexible and compatible with any VLA training objective. We simply finetune the selected heads using the same loss function as the base model—whether that's flow matching loss for diffusion-based models like π_0 or cross-entropy for discretized action spaces. This selective updating acts as a targeted refinement that enhances task performance without broadly overwriting the model's parameters.

3.4 Step 3: Inference

After selective finetuning, inference proceeds through standard forward passes with the finetuned weights. Unlike many mechanistic interpretability methods that require computing and manipulating

Table 1: Performance comparison of Robotic Steering on new and in-domain tasks. Methods are finetuned on tasks with varying example sizes (20-200), then evaluated on both new and original in-domain tasks for 20 trials under the same task and environmental settings.

Method	Training Time	Trainable Params	New Tasks		In-Domain Tasks		
			Place Marker in Cup	Push Button Hard	Pick Cube	Place Cube in Bowl	Push Bowl to Cup
Zero-shot	-	-	10%	0%	0%	0%	0%
Full-head LoRA	239 min	1785.9M	75%	65%	75%	60%	60%
Robotic Steering (KNN)	189 min	78.8M	**80%**	**75%**	**90%**	**85%**	**65%**

activations at inference time, our approach produces a standard model checkpoint deployable without additional computational overhead or specialized procedures. The model simply uses the finetuned weights for the selected heads while maintaining frozen weights elsewhere, preserving both new task capabilities and existing skills through this selective modification.

4 Evaluation

In our work, we evaluate our method on a variety of real-world on-robot tasks using the strong π_0 VLA to demonstrate the effectiveness of our approach on realistic, physically-grounded usecases. We select tasks of diverse difficulties and skills and deeper experimentation and ablation that showcases the many unique qualities of our approach including its performance, robustness, and interpretability. We present more details as follows:

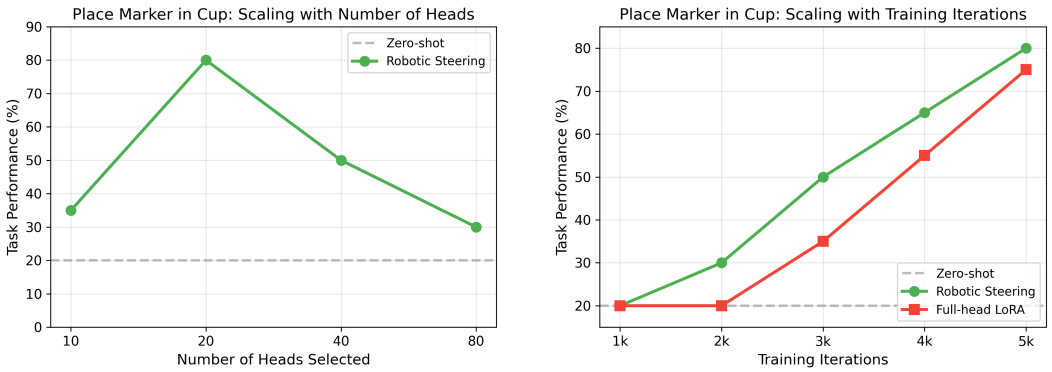
4.1 Implementation Details

While our method is model-agnostic, we use π_0 [10], a state-of-the-art VLA that uses flow matching for continuous action generation. Our entire implementation is in Jax [], which notably lacks convenient hooks to easily extract activations from the model. Thus, we highlight the development of such functionality for a Jax-based model as a core technical contribution of our work. We finetune the model using 2 NVIDIA RTX A6000 GPUs, emphasizing the lightweight nature of our approach. We extract attention activations from the model's PaliGemma [] LLM backbone with 18 layers with 8 heads each, selecting $m = 20$ heads for finetuning based on k-NN regression with $k = 5$ neighbors. The LoRA rank is set to $r = 8$, and we finetune for 5,000 steps for our main experiments using varying number of demonstrations depending on the difficulty of the task. More implementation details can be found in Supplementary Section B.

4.2 Robotic Setup Details

We follow the setup from DROID [36] exactly, using a 7-DoF Franka Emika Panda robot arm with a Robotiq gripper and a low-level Polymetis controller []. As suggested by DROID, we enable two of the three cameras for both finetuning and inference: the left arm camera and wrist camera. We record each example episode at 6 Hz. All data collection is performed on-robot using teleoperation, with each task controlling for the exact objects used to ensure fair evaluation across methods.

We evaluate a total of 5 primary tasks with the following language instructions: (1) "place marker in cup", (2) "push button hard", (3) "pick red cube", (4) "place green cube in red bowl", and (5) "push red bowl to red cup". Tasks (1) and (2) are considered new tasks requiring 200 training samples due to their difficulty based on action complexity, physical demands (e.g., manipulating small objects like markers), and unique task specifications such as pressing the button hard in a particular demonstrated manner. Tasks (3)-(5) are in-domain tasks that require only 20 training samples. Importantly, all experiments use only 20 few-shot expert demonstrations for head selection when applicable, regardless of the total training data available. All models are finetuned for 5,000 iterations as detailed in Section 4.1 (implementation details). More details about the robot and the task setup can be found in Section C of the Supplement.



(a) Number of heads

(b) Training iterations

Figure 3: Scaling experiments on *Place Marker in Cup* task. (a) Success rate versus number of selected attention heads. (b) Success rate versus training iterations for Robotic Steering and Full-head LoRA, both starting from the same zero-shot baseline

5 Results

Our main results are shown in Table 1. The crucial insight of Robotic Steering is that few-shot expert demonstrations can encode the physical nuances of robotic tasks and more importantly inform which task-specific components of a model to finetune.

Indeed, our results demonstrate that Robotic Steering matches or outperforms LoRA's success rate on all evaluated tasks. This is true for both simpler in-domain tasks which are similar to DROID [36] dataset tasks and excitingly more challenging, new tasks which we even provide 200 examples for finetuning. This demonstrates that Robotic Steering is a broadly effective finetuning approach, leveraging just 20 demonstrations to surpass the LoRA baseline. It is worth noting that none of these tasks are trivial given the physical context of on-robot evaluation as we see that zero-shot performance is near 0% success rate for all tasks. While π_0 is a SoTA VLA, it is remains brittle to generalization when faced with variations in environment conditions, robot embodiments, and language instructions. Beyond task performance, Table 1 demonstrates that Robotic Steering is significantly more computationally efficient than full-head LoRA, reducing finetuning time by 21% while using 96% fewer parameters. This efficiency is crucial for practical robotics, where rapid iteration and experimentation in new environments is essential. We present additional results and ablations in Section A of the Appendix.

5.1 Ablations

We perform a comprehensive ablation study of Robotic Steering on the *Place Marker in Cup* task to understand the impact of key design choices. For all ablations, we use the base π-0 model.

Varying number of attention heads. In Figure 3a (a), we examine the impact of varying the number of selected attention heads used in our method. We find that performance peaks at 20 heads (80% success rate) and decreases with both fewer and more heads. This suggests that an optimal subset of heads exists for task-specific adaptation, where too few heads lack sufficient representational capacity while too many heads introduce noise or conflicting signals.

Scaling with training iterations. We investigate how our method scales with the number of training iterations compared to Full-head LoRA. As shown in Figure 3b, Robotic Steering demonstrates faster initial learning and achieves higher final performance (80%) compared to Full-head LoRA (75%) after 5k iterations. This result suggests that our approach scales well, surpassing or at least matching LoRA's capabilities of scaling performance with further training.

Head selection approach. Our results in Table 3 show that K-NN regression, our approach for head selection, slightly outperforms Causal Mediation Analysis (CMA) [63] and REINFORCE [30, 33]. CMA, specifically causal ablation in our experiments, selects heads by adding noise to each head

Table 2: Generalization performance under environmental variations and transfer to related tasks after training on Place Marker in Cup. ↓ indicates performance drop from base condition.

Method	Base Performance	+ Lighting Distractor	+ Positional Variation	Unseen Task: Pick Mug	Unseen Task: Pl. Cube in Bowl
Zero-shot	10%	0%	0%	0%	0%
Full-head LoRA	75%	40% ↓47%	45% ↓X%	0%	40%
Robotic Steering	**80%**	**60%** ↓25%	**55%** ↓X%	**30%**	**55%**

Table 3: Ablation studies on head selection methods and training components. All methods use 20 few-shot examples and select top-20 heads for adaptation.

Method	Place Marker in Cup	Head Selection Time (min)	Fine-tuning Time (min)
Head Selection Methods			
CMA	15%	58 min	51 min
REINFORCE	80%	93 min	51 min
K-NN Regression (Ours)	**80%**	**17 min**	51 min
Training Components			
Queries only	10%	17 min	52 min
Queries + MLP (Ours)	**80%**	**17 min**	**51 min**

and measuring the resulting performance drop on the 20 few-shot demonstrations. REINFORCE optimizes head selection through gradient-based search to maximize task performance. While all three methods achieve comparable task success rates as shown in Table 3, K-NN regression offers a crucial advantage: significantly lower runtime. This is due to K-NN regression not requiring model inference and evaluation for head selection. Once the activations are computed, K-NN regression boils down to a simple and very efficient retrieval-based regression on the activations themselves.

Training Components. We also carefully ablate the recipe for which precise components of the model to finetune. Of course, when selecting heads, it is natural to finetune their queries, but we also question whether additionally finetuning their MLPs, yields any benefit. Our results in Table 3 suggest that indeed finetuning both the queries and MLPs associated with the selected task-specific heads yields improvements in success rate. This suggests that the feedforward projection following attention is important to adapt for VLA finetuning. We do not consider finetuning the parameters of the keys and values as they are shared per layer in π_0's base LLM [10].

5.2 Additional Experiments

In this subsection, we present experiments that demonstrate additional properties and capabilities of Robotic Steering, beyond its use for improving task-specific performance. Additional visualizations can be found in Supplementary Section A.2.2. For all experiments, we use the π_0 model with our steering method trained on *Place Marker in Cup*.

Robustness to environmental distractors. We evaluate the robustness and generalization of our method to common environmental variations that occur in real-world robotic deployment. We test our model trained on the base *Place Marker in Cup* task under two challenging conditions: lighting variations and positional distractors. As shown in Table 2, Robotic Steering shows significantly less degradation in the face of environmental variations. This demonstrates that our sparse head selection naturally filters out features sensitive to task-irrelevant variations while preserving task-critical attention heads.

Zero-shot transfer to related tasks. A key advantage of our approach is the ability to transfer learned steering vectors to related manipulation tasks without additional training. In Table 2, we evaluate the heads selected for *Place Marker in Cup* on two related tasks: *Pick Mug* and *Place Cube in Bowl*. Despite being trained only on the marker placement task, our method achieves 30% success

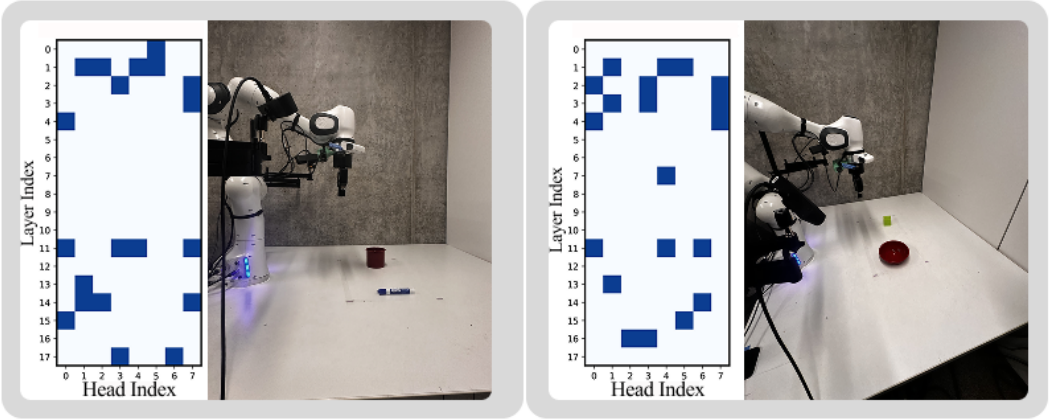


Figure 4: **Task and Head Selection Visual Left**: Attention heads selected by **Robotic Steering** and task visual for *Place Marker in Mug*; **Right:** Attention heads selected by **Robotic Steering** and task visual for *Place Cube in Bowl*

on *Pick Mug*, while Full-head LoRA completely fails (0%). This suggests that the sparse selected heads capture generalizable representations that transfer across tasks with similar action spaces and object interactions.

Visualizing selected attention heads. To understand what our method learns, we visualize the attention patterns of the top-selected heads for *Place Marker in Cup*, *Pick Cube*, and *Push Bowl to Cup* tasks in Figure 4. The visualizations reveal that different tasks activate distinct sets of heads. Intuitively, this aligns with the notion of functional specificity, except in models' attention heads. This interpretability is a key advantage of our approach: unlike black-box finetuning methods, we can directly inspect which attention mechanisms are being leveraged for each task.

6 Conclusion

In this work, we introduce Robotic Steering, which demonstrates that few-shot demonstrations can specify physically-grounded embodied tasks and help identify which specific attention heads in VLAs encode task-relevant physical reasoning. By selectively finetuning only these heads, we match or exceed full LoRA performance while using 96% fewer parameters and achieving superior robustness to environmental variations. Our visualizations reveal that different manipulation tasks activate distinct attention patterns, providing mechanistic insight into how VLAs encode physical tasks.

This work opens exciting research directions at the intersection of mechanistic interpretability and robotic learning. Future methods could explore alternative head selection approaches beyond K-NN regression, investigate finer-grained selection at the parameter or neuron level, or develop compositional schemes where multiple task-specific adaptations combine without interference. More fundamentally, our results suggest that the question of "what to finetune" deserves equal attention to "how to finetune", a shift that could transform how we adapt foundation models for robotics. As VLAs scale to billions of parameters, the ability to precisely identify and modify task-relevant components will become essential for practical deployment across the wide variety of physical contexts robots must master.

References

[1] The claude 3 model family: Opus, sonnet, haiku. URL `https://api.semanticscholar.org/CorpusID:268232499`.

[2] Abrar Anwar, Rohan Gupta, and Jesse Thomason. Contrast sets for evaluating language-guided robot policies. In *Conference on Robot Learning (CoRL)*, 2024.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966, 2023.

[5] Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. Decomposing and interpreting image representations via text in vits beyond clip. In *NeurIPS*, 2024.

[6] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. *arXiv preprint arXiv:2406.04236*, 2024. URL https://arxiv.org/abs/2406.04236.

[7] Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Ryan A. Rossi, Nanxuan Zhao, Vlad I. Morariu, Varun Manjunatha, and Soheil Feizi. On mechanistic knowledge localization in text-to-image generative models. *arXiv preprint arXiv:2405.01008*, 2024. URL https://arxiv.org/abs/2405.01008.

[8] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

[9] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Understanding the role of individual units in a deep neural network. In *Proceedings of the National Academy of Sciences*, volume 117, pp. 30071–30077, 2020.

[10] 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. pi_0: A vision-language-action flow model for general robot control. *CoRR*, abs/2410.24164, 2024.

[11] Tianning Chai, Chancharik Mitra, Brandon Huang, Gautam Rajendrakumar Gare, Zhiqiu Lin, Assaf Arbelle, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, Deva Ramanan, et al. Activation reward models for few-shot model alignment. *arXiv preprint arXiv:2507.01368*, 2025.

[12] Open-X Embodiment Collaboration. Open x-embodiment: Robotic learning datasets and rt-x models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

[13] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. URL https://arxiv.org/abs/2309.08600.

[14] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *CoRR*, abs/1910.11215, 2019.

[15] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[16] Maximilian Du and Shuran Song. Dynaguide: Steering diffusion polices with active dynamic guidance. *arXiv preprint arXiv:2506.13922*, 2025.

[17] Maximilian Du, Suraj Nair, Dorsa Sadigh, and Chelsea Finn. Behavior retrieval: Few-shot imitation learning by querying unlabeled datasets. *arXiv preprint arXiv:2304.08742*, 2023.

[18] Nicholas Elhage, Neel Nanda, Catherine Olsson, Mor Geva, Akbir Khan Reddy, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. URL https://arxiv.org/abs/2209.10652.

[19] Evelina Fedorenko, Michael K. Behr, and Nancy Kanwisher. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433, 2011. doi: 10.1073/pnas.1112937108.

[20] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In *Conference on Robot Learning (CoRL)*, 2017.

[21] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023. URL https://arxiv.org/abs/2310.05916.

[22] Jensen Gao, Suneel Belkhale, Sudeep Dasari, Ashwin Balakrishna, Dhruv Shah, and Dorsa Sadigh. A taxonomy for evaluating generalist robot policies. *arXiv preprint arXiv:2503.01238*, 2025.

[23] Ali Ghadirzadeh, Xi Chen, Petra Poklukar, Chelsea Finn, MÃĕrten BjÃűrkman, and Danica Kragic. Bayesian meta-learning for few-shot policy adaptation across robotic platforms. pp. 1274–1280, 09 2021. doi: 10.1109/IROS51168.2021.9636628.

[24] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030. URL https://distill.pub/2021/multimodal-neurons.

[25] Shresth Grover, Akshay Gopalkrishnan, Bo Ai, Henrik I. Christensen, Hao Su, and Xuanlin Li. Enhancing generalization in vision-language-action models by preserving pretrained representations. 2025. URL https://api.semanticscholar.org/CorpusID:281315107.

[26] Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, Singapore, December 2023. Association for Computational Linguistics.

[27] Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*, 2023.

[28] Evan Hernandez, Belinda Z. Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023. URL https://arxiv.org/abs/2304.00740.

[29] Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. Finding visual task vectors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[30] Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. Finding visual task vectors. In *European Conference on Computer Vision (ECCV)*, pp. 257–273. Springer, 2025.

[31] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[32] Shengchao Hu, Wanru Zhao, Weixiong Lin, Li Shen, Ya Zhang, and Dacheng Tao. Prompt tuning with diffusion for few-shot pre-trained policy generalization. *arXiv preprint arXiv:2411.01168*, 2024.

[33] Brandon Huang, Chancharik Mitra, Assaf Arbelle, Leonid Karlinsky, Trevor Darrell, and Roei Herzig. Multimodal task vectors enable many-shot multimodal in-context learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22124–22153, 2024.

[34] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

[35] Nancy Kanwisher. Domain specificity in face perception. *Nature neuroscience*, 3(8):759–763, 2000.

[36] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, ..., Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[37] 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. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[38] Puhao Li, Yingying Wu, Ziheng Xi, Wanlin Li, Yuzhe Huang, Zhiyuan Zhang, Yinghan Chen, Jianan Wang, Song-Chun Zhu, Tengyu Liu, and Siyuan Huang. Controlvla: Few-shot object-centric adaptation for pre-trained vision-language-action models. *arXiv preprint arXiv:2506.16211*, 2025. URL https://arxiv.org/abs/2506.16211.

[39] Anthony Liang, Ishika Singh, Karl Pertsch, and Jesse Thomason. Transformer adapters for robot learning. In *CoRL 2022 Workshop on Pre-training Robot Learning*, 2022.

[40] Li-Heng Lin, Yuchen Cui, Amber Xie, Tianyu Hua, and Dorsa Sadigh. Flowretrieval: Flow-guided data retrieval for few-shot imitation learning. *Conference on Robot Learning (CoRL)*, 2024.

[41] Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A. Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, et al. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*, 2025. URL https://arxiv.org/abs/2502.17516.

[42] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[44] Zuxin Liu, Jesse Zhang, Kavosh Asadi, Yao Liu, Ding Zhao, Shoham Sabach, and Rasool Fakoor. Tail: Task-specific adapters for imitation learning with large pretrained models. *arXiv preprint arXiv:2310.05905*, 2023.

[45] Yecheng Jason Ma, Joey Hejna, Ayzaan Wahid, Chuyuan Fu, Dhruv Shah, Jacky Liang, Zhuo Xu, Sean Kirmani, Peng Xu, Danny Driess, Ted Xiao, Jonathan Tompson, Osbert Bastani, Dinesh Jayaraman, Wenhao Yu, Tingnan Zhang, Dorsa Sadigh, and Fei Xia. Vision language models are in-context value learners. *arXiv preprint arXiv:2411.04549*, 2024. doi: 10.48550/arXiv.2411.04549. URL https://arxiv.org/abs/2411.04549.

[46] Abraham Harold Maslow. *The Psychology of Science*. Harper & Row, New York,, 1966.

[47] Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. Sparse attention vectors: Generative multimodal model features are discriminative vision-language classifiers. *arXiv preprint arXiv:2412.00142*, 2024.

[48] Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. *arXiv preprint arXiv:2410.13816*, 2024.

[49] Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning. *ArXiv*, abs/2406.11815, 2024. URL https://api.semanticscholar.org/CorpusID:270559839.

[50] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[51] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[52] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023. URL https://arxiv.org/abs/2312.06681.

[53] Physical Intelligence, Kevin Black, Noah Brown, Danny Driess, Chelsea Finn, Sergey Levine, et al.
$pi_{0.5}$: A vision-language-action model with open-world generalization. *CoRR*, abs/2504.16054, 2025.

[54] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

[55] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[56] Sarah Schwettmann, Oliver Watkins, Martin Wattenberg, and Shan Carter. Towards mechanistic interpretability of multimodal neurons. *arXiv preprint arXiv:2301.11796*, 2023. URL https://arxiv.org/abs/2301.11796.

[57] Mingchen Song, Xiang Deng, Guoqiang Zhong, Qi Lv, Jia Wan, Yinchuan Li, Jianye Hao, and Weili Guan. Few-shot vision-language action-incremental policy learning. *arXiv preprint arXiv:2504.15517*, 2025.

[58] Kaustubh Sridhar, Souradeep Dutta, Dinesh Jayaraman, and Insup Lee. Ricl: Adding in-context adaptability to pre-trained vision-language-action models. *arXiv preprint arXiv:2508.02062*, 2025. URL https://arxiv.org/abs/2508.02062.

[59] Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[60] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024.

[61] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[62] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[63] Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

[65] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024. URL https://arxiv.org/abs/2308.10248.

[66] Amber Xie, Rahul Chand, Dorsa Sadigh, and Joey Hejna. Data retrieval with importance weights for few-shot imitation learning. *Conference on Robot Learning (CoRL)*, 2025.

[67] Yida Yin, Zekai Wang, Yuvan Sharma, Dantong Niu, Trevor Darrell, and Roei Herzig. In-context learning enables robot action prediction in llms. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8972–8979. IEEE, 2025.

[68] Tianhe Yu, Chelsea Finn, Sudeep Dasari, Annie Xie, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. In *Robotics: Science and Systems (RSS)*, 2018.

[69] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 41, pp. 2131–2145, 2018.

[70] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, ..., Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proc. of the 7th Conference on Robot Learning (CoRL)*, pp. 2165–2183. PMLR, 2023.