Mechanistic Finetuning of Vision-Language-Action Models via Few-Shot Demonstrations

Chancharik Mitra[1⋆], Yusen Luo[2*], Raj Saravanan[3*], Dantong Niu[3], Anirudh Pai[3], Jesse Thomason[2], Trevor Darrell[3], Abrar Anwar[2], Deva Ramanan[1], and Roei Herzig[3]

[1] Carnegie Mellon University[1]
[2] University of Southern California[2]
[3] University of California, Berkeley[3]
https://chancharikmitra.github.io/robosteering/

Abstract. Vision-Language Action (VLAs) models promise to extend the remarkable success of vision-language models (VLMs) to robotics. Yet, unlike VLMs in the vision-language domain, VLAs for robotics require finetuning to contend with varying physical factors like robot embodiment, environment characteristics, and spatial relationships of each task. We propose leveraging few-shot demonstrations of tasks to better capture this physical variability. Overfitting to the trained task is yet another problem of existing methods. Inspired by functional specificity, we hypothesize that finetuning only task-specific sparse model representations is both more effective and better retains pretrained model generality. In this work, we introduce **Robotic Steering**, a finetuning approach grounded in mechanistic interpretability that leverages few-shot demonstrations to identify and selectively finetune task-specific attention heads aligned with the physical, visual, and linguistic requirements of robotic tasks. Through comprehensive on-robot evaluations with a Franka Emika robot arm, we demonstrate that Robotic Steering outperforms LoRA while achieving superior robustness under task variation, reduced computational cost, and enhanced interpretability for adapting VLAs to diverse robotic tasks.

1 Introduction

Vision-Language-Action (VLA) models represent an emerging paradigm that extends foundation models to robotics by jointly modeling vision, language, and physical action spaces [42, 54, 69, 79]. While large-scale robotic datasets [15, 17, 40] have enabled significant scaling of model pretraining, VLAs have yet to achieve the impressive generalization of language and vision-language models. Unlike Vision-Language Models (VLMs) [5, 43, 56] 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.

⋆ Equal contribution.

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, 25], 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 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 [35] broadly adapt all model parameters regardless of task specifics, risking overfitting to the narrow training distribution and eroding the pretrained representations that enable generalization. In contrast, we take inspiration from functional specificity in neuroscience, which suggests that certain brain regions are specialized for particular tasks [21,39], and from mechanistic interpretability in machine learning, which has shown that specific attention heads in transformers encode distinct capabilities [29, 30, 55]. Building on these insights, we introduce a novel paradigm especially suited for robotics: using few-shot demonstrations, we first identify which attention heads encode task-relevant representations, and then specifically finetune those components. Because the base weights stay frozen and adaptation is low-rank and selective over task-relevant heads, the broad pretrained knowledge that makes VLAs capable zero-shot remains largely intact. Such task-relevant adaptation yields policies that are not only effective on the target task but also robust to environmental variation and transferable to related tasks.

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 of predicting 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 the query and output projections of the selected heads using LoRA adapters. This preserves the generality and robustness of the pretrained model. Finally, the resulting model deploys as a standard checkpoint without additional overhead. Unlike other

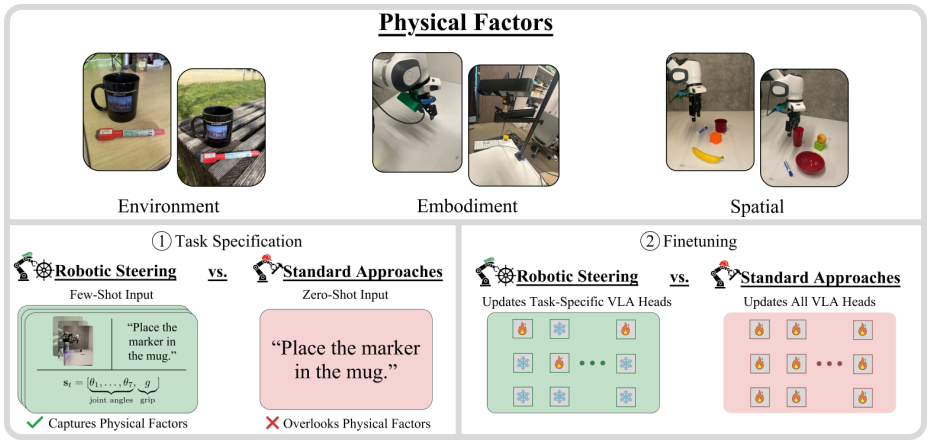


Fig. 1: Robotic Steering. Our method leverages few-shot examples that capture the inherent physical variability of robotic tasks to identify and selectively finetune only task-relevant attention heads. In contrast, standard approaches specify tasks only with a language expression and update all parameters. Because of this, our novel method is a more performant, efficient, and generalizable approach for few-shot VLA finetuning.

mechanistic approaches that require activation interventions during inference, our finetuned weights integrate seamlessly into existing VLA deployment pipelines. An overview is shown in Figure 1, and a detailed view is shown in 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 task generalization and environmental robustness, 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) [3,38,60,71] and Large Multimodal Models (LMMs) [1,4,49,50,57, 67,68] 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 [18,42,61,69,79] and is enabled by large-scale datasets [15,17,40]. This scale of training has demonstrated generalization across embodiments and tasks. The state-of-the-art π-series models—π_0 [12]

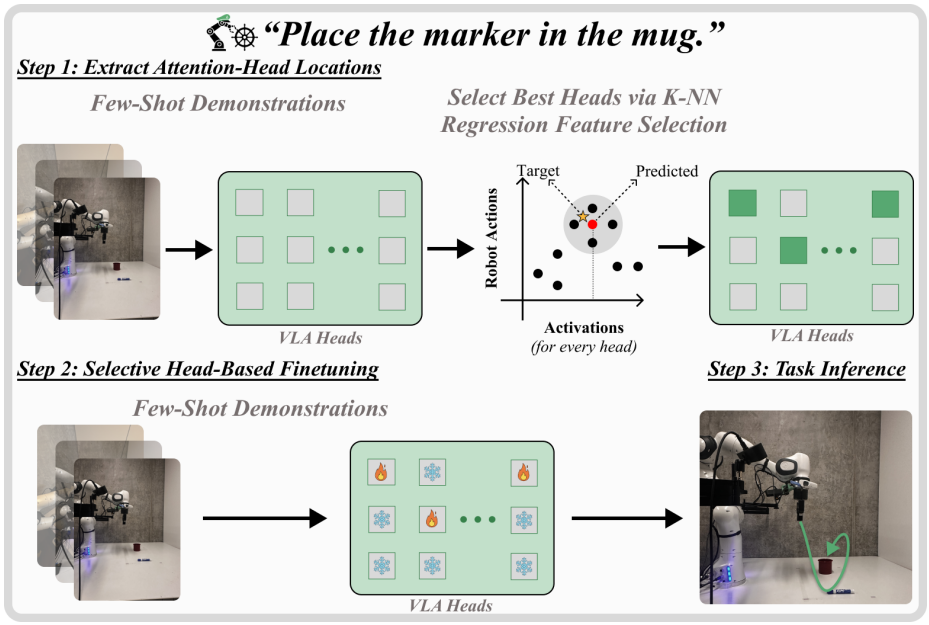


Fig. 2: Robotic Steering Approach. Robotic Steering enables targeted adaptation of VLAs by (1) using few-shot demonstrations to extract task-relevant attention heads, (2) finetuning only these components, (3) and using the sparsely finetuned weights for task inference.

and $\pi_{0.5}$ [59]—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 [52,64,76] condition on demonstrations without weight updates but face context limitations; parameter-efficient finetuning reduces trainable parameters through general-purpose adapters [28,35,36,45,51] and VLA-specific finetuning recipes and adapters [41,44,63,73,74]; meta-learning [23,26,77] and behavior retrieval [19,46,75] enable rapid adaptation given access to prior data. However, these methods update parameters without considering which components encode task-specific physical reasoning, lacking interpretability and failing to leverage VLAs' structured representations. This motivates our mechanistic approach that stands in contrast to previous work by identifying and selectively finetuning only task-relevant attention heads.

Mechanistic Interpretability. Recent advances in mechanistic interpretability have revealed how model behavior can be precisely manipulated through internal representations. Early research [9,10,78] established frameworks for understanding semantic encoding in neural networks, while activation steering methods [58,66,72] demonstrated parameter-free behavior modification. The discovery of specialized components like induction heads [55] and task-specific neurons [32] led to task

vector abstractions [31, 70], with parallel work on sparse autoencoders [16] and superposition [20] providing tools for decomposing representations.

An emerging line of work leverages few-shot mechanistic interpretability for model adaptation via task vector methods [14, 33, 37, 53], which concentrate task-relevant information in specific attention heads or activation subspaces. Research in multimodal representations has revealed how vision-language models structure cross-modal concepts through multimodal neurons [27], mechanistic understanding [62], text-based decomposition [6,24], and knowledge localization [7, 8]. We are also excited by concurrent work on finetuning-free activation steering in VLAs that looks at controlling VLAs' behavior on in-domain tasks (e.g. controlling robot height and speed for a task) [29]. The comprehensive survey by Lin et al. [48] provides a broader overview of such approaches. While these methods have succeeded in language and vision domains, our work is the first to apply mechanistic interpretability for *finetuning* VLAs to learn new tasks in an efficient, interpretable, and robust manner.

3 Methods

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

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 and $\pi_{0.5}$ [12,59] 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, leveraging this multimodal context for action prediction.
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 from a few-shot set of examples.

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 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 [76]. Functionally, however, the approach is identical with or without this step. More details can be found in Section **??** of the Supplementary. 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 a distance-weighted average of the retrieved neighbors' actions:

$$\hat{a}_t^{l,h} = \sum_{(\tau_j^s)\in\mathcal{N}_k^{l,h}(q)} w_j^s \, a_j^s, \quad w_j^s \propto \frac{1}{d_j^s + \epsilon} \tag{3}$$

where d_j^s is the cosine distance to neighbor (τ_j^s) and the weights are normalized over $\mathcal{N}_k^{l,h}(q)$. Finally, we compute the head's score as the mean squared error across all queries:

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

We fix the number of selected heads m and choose the neighborhood size k by grid search under the criterion above. Since $\text{score}_k(l, h)$ depends on k through the retrieved neighbors, we recompute the head ranking for each k in a small

grid and retain the top m. We then choose the k whose top-m set attains the lowest error, and take that set as $\mathcal{H}_{\text{task}}$:

$$\mathcal{H}_{\text{task}} = \{(l, h) \mid \text{score}_k(l, h) \text{ is among the } m \text{ lowest at the selected } k\}. \quad (5)$$

This selection is a direct scan over the grid, not an elbow or thresholding heuristic. We ablate the value of m in Section A.2 of the Supplementary. These heads learn representations that effectively map task-specific observations to other observations in a few-shot demonstration set that require similar actions, 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 on the parameters of those particular heads.

Head-Selective LoRA. Robotic Steering is a head-selective form of LoRA: where standard LoRA adapts every attention head, we adapt only the task-relevant subset $\mathcal{H}_{\text{task}}$. We train LoRA adapters on the query and output projections of the selected heads and freeze the attention adapters of all others:

$$W_Q'^{l,h} = W_Q^{l,h} + B_Q^{l,h} A_Q^{l,h}, \qquad W_O'^{l,h} = W_O^{l,h} + B_O^{l,h} A_O^{l,h}, \qquad (l, h) \in \mathcal{H}_{\text{task}}. \quad (6)$$

We leave the key and value projections frozen, since $\pi_{0.5}$'s Gemma backbone shares one key/value head per layer (multi-query attention), so they cannot be adapted head-by-head. The remaining trainable components, the LLM's FFN LoRA adapters and the flow-matching action head, are common to a standard LoRA finetuning paradigm and held identical across every method we compare. Head selection is thus the only difference between Robotic Steering and the LoRA baseline, so any performance gap reflects the impact of feature selection, not the usage of a larger finetuning budget.

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 $\pi_{0.5}$ or cross-entropy for discretized action spaces. This selective updating acts as a targeted refinement that enhances task performance without broadly overwriting all of 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 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.

4 Evaluation

In our work, we evaluate our method on a variety of real-world on-robot tasks using the strong π_0 and $\pi_{0.5}$ VLAs 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 [12] and $\pi_{0.5}$ [59], two state-of-the-art VLAs that use flow matching for continuous action generation. Our entire implementation is in Jax [13], 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 [11, 65] LLM backbone with 18 layers with 8 heads each, selecting $m = 20$ heads for finetuning based on k-NN regression with $k \in \{10, 20, 30, 40\}$ neighbors and choosing the k that yields the lowest mean squared error. The LoRA rank is set to $r = 8$, and we finetune for 5,000 steps for our main experiments using only 20 demonstrations of the target task. More implementation details are in Supp. Section B.

4.2 Robotic Setup Details

We follow the setup from DROID [40] exactly, using a 7-DoF Franka Emika Panda robot arm with a Robotiq gripper and a low-level Polymetis controller [47]. 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 objects used for fairness across methods.

We evaluate a total of 5 primary tasks with the following language instructions: (1) "place marker in mug", (2) "press red button hard", (3) "pick up red cube", (4) "place green cube in red bowl", and (5) "push red cup to red bowl". We collect 20 teleoperated demonstrations for all tasks for both head selection and finetuning, representing a sample-efficient few-shot paradigm. All models are finetuned for 5,000 iterations as detailed in Section 4.1 (implementation details). More details about the setup can be found in Section C of the Supplement.

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

Table 1: Results. We report the performance and computational cost of **Robotic Steering** applied to real, on-robot tasks. Finetuned methods are trained on 20 demonstrations. All approaches are evaluated using 40 trials under the same task settings.

Method	Computational Cost		Tasks				
	Training Time	Trainable Params	Place Marker in Mug	Press Button Hard	Pick Cube	Place Cube in Bowl	Push Cup to Bowl
Zero-shot Methods							
π_0-DROID	–	–	15%	0%	35%	5%	0%
$\pi_{0.5}$-DROID	–	–	10%	0%	20%	0%	0%
Finetuned Methods							
π_0 Full-head LoRA	239 min	29M	75%	45%	75%	60%	10%
π_0 **Robotic Steering (KNN)**	189 min	15.8M	**80%**	**75%**	**90%**	**85%**	**17.5%**
$\pi_{0.5}$ Full-head LoRA	214 min	29M	62.5%	**90%**	70%	65%	20%
$\pi_{0.5}$ **Robotic Steering (KNN)**	185 min	15.8M	**72.5%**	85%	**77.5%**	**80%**	**27.5%**

tasks and more importantly inform which task-specific components of a model to finetune for model adaptation.

Indeed, our results demonstrate that Robotic Steering matches or outperforms LoRA's success rate on all evaluated tasks across both π_0 and $\pi_{0.5}$ VLA models. This holds true for both simpler in-domain tasks which are similar to DROID [40] dataset tasks and more challenging new tasks. This demonstrates that Robotic Steering is a broadly effective finetuning approach, leveraging only 20 demonstrations for both head selection and finetuning 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 evidenced by near 0% zero-shot success rates for most tasks that were evaluated.

Interestingly, π_0 and $\pi_{0.5}$ exhibit similar performance on these short-horizon manipulation tasks. We hypothesize this similarity stems from two factors: first, π_0 already achieves strong performance on this task distribution, leaving limited room for improvement; and second, $\pi_{0.5}$'s methodological improvements over $\pi0$ are significantly centered around long-horizon reasoning tasks, wheras our tasks are shor-horizon for simpler and more consistent comparison.

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 trainable parameters. This efficiency is crucial for practical robotics, where rapid iteration and experimentation in new environments is essential.

Additional results and ablations are in Section A in Supp.

5.1 Ablations

We perform a comprehensive ablation study of Robotic Steering on the *Place Marker in Mug* task to understand the impact of key design choices. For all ablations, we use $\pi_{0.5}$.

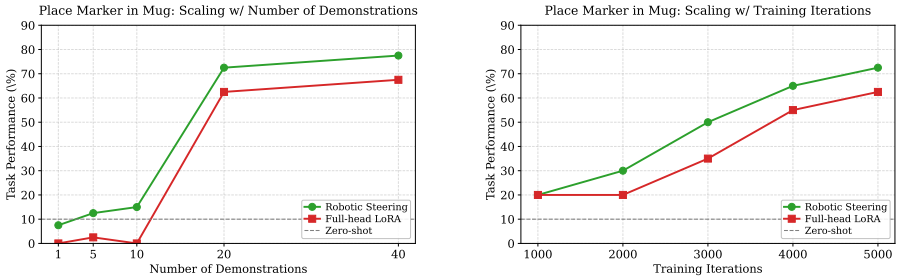
Table 2: Ablations. We compare the performance and cost of different design choices for head-selection and training configuration. All methods use 20 few-shot demos and select top-20 heads for adaptation.

Method	Place Marker in Mug	Head Selection Time (min)	Fine-tuning Time (min)	Activation Cache (M)
Head Selection Methods				
CMA	15%	58 min	188 min	92.83M
REINFORCE	80%	93 min	186 min	92.83M
K-NN Regression (Ours)	**80%**	**0.2 min**	185 min	92.83M
Training Components				
Queries only	10%	0.2 min	186 min	92.83M
Queries + MLP (Ours)	**80%**	**0.2 min**	**185 min**	92.83M

Head selection approach. Our results in Table 2 show that K-NN regression, our approach for head selection, slightly outperforms Causal Mediation Analysis (CMA) [70] and REINFORCE [34, 37]. CMA, more specifically implemented as causal ablation in our experiments, selects heads by adding noise to each head 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 2, 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 becomes a simple and efficient regression approach on the activations.

Training Components. We ablate which components to finetune alongside the selected heads. When adapting the selected task-specific heads, we finetune their query and output projections, and we ask whether additionally adapting the feedforward (FFN) path yields benefit. Table 2 shows that training the FFN LoRA on top of the selected heads' attention adapters improves success rate, indicating that the feedforward projection following attention is important to adapt for VLA finetuning. We do not finetune the key and value projections, since $\pi_{0.5}$'s Gemma backbone shares a single KV head across all query heads in a layer (multi-query attention).

Scaling number of demonstrations. In Figure 3a, we analyze the effect of varying the number of demonstrations used for fine-tuning. Robotic Steering consistently outperforms Full-head LoRA across all data scales, achieving higher success rates especially in low-data regimes. Performance improves sharply up to 20 demonstrations, reaching around 72.5% success rate for Robotic Steering compared to 62.5% for Full-head LoRA, and then saturates. This demonstrates that our method can make more efficient use of limited demonstrations while maintaining superior scalability.



(a) Number of Demonstrations **(b)** Training iterations

Fig. 3: Scaling Experiments. For the *Place Marker in Mug* task, we show the (a) success rate versus number of demonstrations and (b) success rate versus number of training iterations for Robotic Steering and full-head LoRA.

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 (72.5%) compared to Full-head LoRA (62.5%) after 5k iterations. This result suggests that our approach surpasses or at least matching LoRA's capabilities of scaling performance with further training.

5.2 Additional Experiments

A key strength of Robotic Steering is its flexibility and generalization across different training scenarios. In this section, we demonstrate that our approach not only improves task-specific performance but also exhibits robust generalization capabilities in three important dimensions: transfer to unseen tasks, multi-task learning, and robustness to environmental variations. Additional results can be found in Section A.2 of the Supplementary.

Generalization to unseen tasks. While trained exclusively on Place Marker in Mug and Place Cube in Bowl, our method generalizes successfully to an unseen task (Pick Mug). Excitingly, Robotic Steering leads to noticeable gains over LoRA on unseen tasks in both single and multi-task finetuning settings. This result shows us that while Robotic Steering leads to superior performance on the specific tasks in the train set, it also offers broader generalization to similar tasks, a capability that LoRA appears to lack. These results are shown in the left-hand columns of Table 3.

Multi-task training strategies. Because Robotic Steering selects task-specific heads, our approach naturally accommodates multiple strategies for multi-task learning. We explore two complementary approaches. *Non-overlapping head selection* trains each task on a distinct set of heads, where k-NN regression independently selects the 20 most task-relevant heads for each individual task. This allows tasks to specialize to their own optimal head subset. *Joint head selection* uses a unified set of examples and activations from both tasks to

Table 3: Generalization and Robustness. We compare **Robotic Steering** to full-head LoRA in both single and multi-task settings. We also compare both methods' generalization to unseen tasks and robustness to environmental variability. All methods are finetuned on 20 demonstrations *per task* and evaluated on 40 trials. Results are shown for $\pi_{0.5}$ VLA model.

		Task Generalization			Task Robustness		
Method	**Training Strategy**	**Place Marker in Mug**	**Place Cube in Bowl**	**Pick Mug (Unseen)**	**Lighting Variation**	**Form Variation**	**Distractor Object Variation**
Single-Task Training (Place Marker in Mug only)							
LoRA	Single	62.5%	20%	42.5%	25% ↓37.5%	22.5% ↓40%	30% ↓32.5%
Robotic Steering	Single	**72.5%**	**25%**	**65%**	**47.5%** ↓25%	**37.5%** ↓35%	**40%** ↓32.5%
Multi-Task Training (Place Marker + Place Cube)							
LoRA	Joint	37.5%	37.5%	0%	15%	12.5%	15%
Robotic Steering	Non-overlapping	**45%**	**65%**	12.5%	**30%**	17.5%	25%
Robotic Steering	Joint Selection	40%	47.5%	**20%**	22.5%	**20%**	**27.5%**

select a single set of 20 heads optimized for joint performance, enabling shared representations across the task pair. Across both multi-task approaches, Robotic Steering outperforms full-head LoRA consistently across both trained tasks suggesting that Robotic Steering works well as a multi-task learning framework. Please refer to the bottom left of Table 3 for these results.

Robustness to environmental variations. A critical concern in real-world robotics is robustness to environmental changes that occur during deployment. We evaluate both Robotic Steering and LoRA under three types of perturbations to the Place Marker in Mug task: lighting variation, form variation (e.g. different marker shapes), and the introduction of distractor objects. Importantly, Robotic Steering demonstrates significantly greater robustness across all three conditions compared to full-head LoRA. This suggests that task-specific sparse head selection naturally filters out task-irrelevant features while preserving robust, task-critical representations. This result is particularly exciting as it challenges a reasonable assumption that sparse, task-specific parameter selection might be brittle; instead, our approach demonstrates that mechanistic steering produces generalizable and robust finetuning even under distributional shift. These findings are shown in the right-hand columns of Table 3.

Generalization in Simulation. The real-robot experiments above demonstrate Robotic Steering's robustness to environmental variation and its ability to generalize across tasks. To provide a reproducible simulation anchor for these findings, we evaluate on the LIBERO-Plus benchmark [22], which augments the standard LIBERO suite with object-layout and lighting perturbations at test time (Table 4).

Table 4: Simulation Results on LIBERO-Long. We compare **Robotic Steering** (RS) to full-head LoRA finetuning using the Long suite in the LIBERO-Plus benchmark, which introduces environmental variation at evaluation time. Both finetuned methods train on π_0 using 20 demonstrations from *Task 0 only* . We evaluate within-task robustness (Task 0 under varied conditions) and cross-task generalization (Tasks 1–9, unseen during finetuning). Robotic Steering retains substantially more of π_0's zero-shot capability than LoRA across both axes of evaluation.

	Trained	**Unseen Tasks**									**Average**	
Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Unseen**	**All**
Object Layout Variation												
π_0 ZS	94.1	94.1	46.2	91.3	34.3	81.3	68.6	76.9	28.0	66.7	65.8	68.9
π_0 + LoRA	41.2	91.2	**65.4**	78.3	8.6	66.7	65.7	53.8	8.0	23.1	51.1	50.0
π_0 + RS	**76.5**	**94.1**	53.8	**95.7**	**31.4**	**72.9**	**74.3**	**69.2**	0.0	**51.3**	**60.8**	**62.5**
Lighting Variation												
π_0 ZS	92.3	100.0	81.0	87.8	78.8	100.0	90.3	100.0	8.3	84.2	80.1	80.7
π_0 + LoRA	53.8	100.0	**90.5**	59.2	3.0	86.4	51.6	100.0	0.0	21.1	51.0	51.1
π_0 + RS	**92.3**	100.0	73.8	**89.8**	**78.8**	**90.9**	**77.4**	100.0	0.0	**55.3**	**72.0**	**73.0**

Both methods finetune π_0 on Task 0 under a single fixed setting; LIBERO-Plus then evaluates under environmental conditions unseen during training, mirroring the distributional shift we study on the real robot. The simulation results reinforce the same pattern observed in hardware: LoRA overfits to the narrow training distribution and degrades well below the original zero-shot performance, both on the trained task and across all nine unseen tasks. Robotic Steering, by contrast, retains far more of the pretrained model's capability. Under lighting variation, for instance, Robotic Steering preserves the full zero-shot accuracy on the trained task while LoRA drops by nearly 40 percentage points; on unseen tasks, the gap is equally pronounced. These results corroborate our real-robot findings at scale: mechanistic steering adapts the policy without the catastrophic forgetting that conventional finetuning induces, yielding robust transfer across both tasks and environments. Refer to Table 4 for results.

Interpreting task-specific attention heads. To understand what our method learns, we visualize the attention patterns of the top-selected heads for Place Marker in Mug and Place Cube in Bowl in Figure 4. The visualizations confirm that different tasks activate distinct sets of heads, consistent with the principle of functional specificity in attention mechanisms. This interpretability can be a fundamental advantage of Robotic Steering: unlike general adaptation methods, we can directly inspect and understand which attention mechanisms are being leveraged for each task. We encourage future works to explore the interpretability of these task representations further. Visualizations of more tasks can be found in Section A.3 of the Supplement.

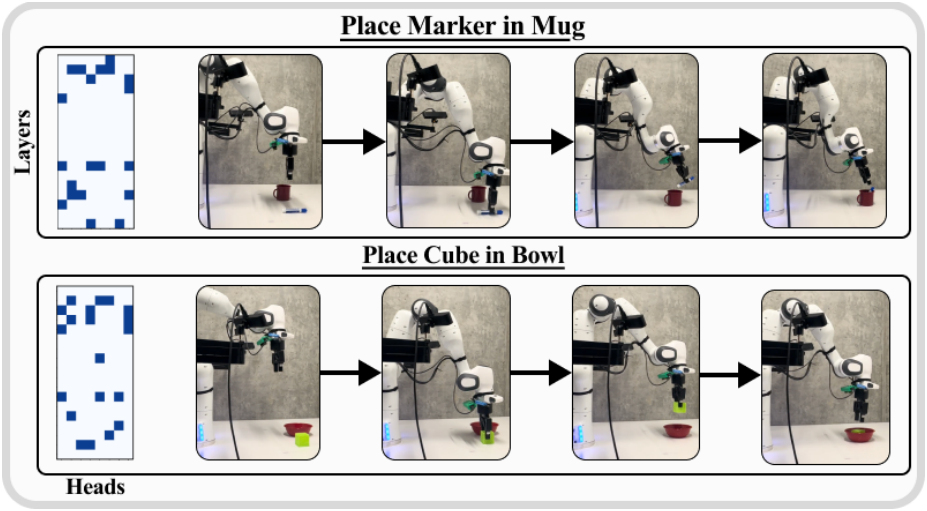


Fig. 4: Task and Attention Head Selection Visual. We show the selected heads and task visuals for Robotic Steering on the *Place Marker in Mug* task (top) and the *Place Cube in Bowl* task (bottom).

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 attention heads in VLAs encode task-relevant physical reasoning. By selectively finetuning these heads, we match or exceed full LoRA performance while using 96% fewer parameters, generalizing to unseen tasks, and being robust 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 parameter-level feature-selection approaches, and consider applications to long-horizon tasks. More fundamentally, our results suggest that the question of "what to finetune" deserves equal attention as "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 learn to master.

We consider potential negative ethical impacts of our work. We note that the ability to surgically modify specific model behaviors while leaving the rest of the pretrained model intact could, in principle, be exploited to inject targeted behavioral changes that are difficult to detect precisely because the model's overall

behavior remains largely unchanged. Nevertheless, we believe this same capability is what makes mechanistic approaches uniquely promising: the ability to isolate and inspect task-specific components opens the door to more interpretable and controllable robotic foundation models, a direction we are excited to see pursued in future work done by the community.

7 Limitations

We outline a few limitations of our work. Firstly, Robotic Steering requires open-source access to the weights which may not be possible on closed-source models. As such, it is impossible to directly apply our method for selecting task-relevant attention heads. Secondly, although our approach is completely embodiment agnostic, we evaluate on a single Franka Emika Panda robot arm due to resource constraints. We encourage the application of our work to a wide variety of embodiments, environments, and task types. Lastly, our interpretable feature extraction focuses on attention heads of the LLM, when there can be informative features in the other model components as well (e.g. action expert features, visual encoder features, etc.). We are enthusiastic about future research that further explores intersection of mechanistic interpretability and robotics.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No(s) (NSF grant number: DGE2140739). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. The claude 3 model family: Opus, sonnet, haiku. https://api.semanticscholar.org/CorpusID:268232499
2. Anwar, A., Gupta, R., Thomason, J.: Contrast sets for evaluating language-guided robot policies. In: Conference on Robot Learning (CoRL) (2024)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. ArXiv **abs/2308.12966** (2023)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. ArXiv **abs/2502.13923** (2025), https://api.semanticscholar.org/CorpusID:276449796
6. Balasubramanian, S., Basu, S., Feizi, S.: Decomposing and interpreting image representations via text in vits beyond clip. In: NeurIPS (2024)

7. Basu, S., Grayson, M., Morrison, C., Nushi, B., Feizi, S., Massiceti, D.: Understanding information storage and transfer in multi-modal large language models. arXiv preprint arXiv:2406.04236 (2024), https://arxiv.org/abs/2406.04236
8. Basu, S., Rezaei, K., Kattakinda, P., Rossi, R.A., Zhao, N., Morariu, V.I., Manjunatha, V., Feizi, S.: On mechanistic knowledge localization in text-to-image generative models. arXiv preprint arXiv:2405.01008 (2024), https://arxiv.org/abs/2405.01008
9. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6541–6549 (2017)
10. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Understanding the role of individual units in a deep neural network. In: Proceedings of the National Academy of Sciences. vol. 117, pp. 30071–30077 (2020)
11. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
12. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L.X., Tanner, J., Vuong, Q., Walling, A., Wang, H., Zhilinsky, U.:
pi_0: A vision-language-action flow model for general robot control. CoRR **abs/2410.24164** (2024)
13. Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: composable transformations of Python+NumPy programs (2018), http://github.com/jax-ml/jax
14. Chai, T., Mitra, C., Huang, B., Gare, G.R., Lin, Z., Arbelle, A., Karlinsky, L., Feris, R., Darrell, T., Ramanan, D., et al.: Activation reward models for few-shot model alignment. arXiv preprint arXiv:2507.01368 (2025)
15. Collaboration, O.X.E.: 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). https://doi.org/10.1109/ICRA57147.2024.10611477
16. Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600 (2023), https://arxiv.org/abs/2309.08600
17. Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., Schmeckpeper, K., Singh, S., Levine, S., Finn, C.: Robonet: Large-scale multi-robot learning. CoRR **abs/1910.11215** (2019)
18. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
19. Du, M., Nair, S., Sadigh, D., Finn, C.: Behavior retrieval: Few-shot imitation learning by querying unlabeled datasets. arXiv preprint arXiv:2304.08742 (2023)
20. Elhage, N., Nanda, N., Olsson, C., Geva, M., Reddy, A.K., et al.: Toy models of superposition. arXiv preprint arXiv:2209.10652 (2022), https://arxiv.org/abs/2209.10652

21. Fedorenko, E., Behr, M.K., Kanwisher, N.: Functional specificity for high-level linguistic processing in the human brain. Proceedings of the National Academy of Sciences **108**(39), 16428–16433 (2011). https://doi.org/10.1073/pnas.1112937108
22. Fei, S., Wang, S., Shi, J., Dai, Z., Cai, J., Qian, P., Ji, L., He, X., Zhang, S., Fei, Z., Fu, J., Gong, J., Qiu, X.: Libero-plus: In-depth robustness analysis of vision-language-action models. arXiv preprint arXiv:2510.13626 (2025)
23. Finn, C., Yu, T., Zhang, T., Abbeel, P., Levine, S.: One-shot visual imitation learning via meta-learning. In: Conference on Robot Learning (CoRL) (2017)
24. Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting clip's image representation via text-based decomposition. arXiv preprint arXiv:2310.05916 (2023), https://arxiv.org/abs/2310.05916
25. Gao, J., Belkhale, S., Dasari, S., Balakrishna, A., Shah, D., Sadigh, D.: A taxonomy for evaluating generalist robot policies. arXiv preprint arXiv:2503.01238 (2025)
26. Ghadirzadeh, A., Chen, X., Poklukar, P., Finn, C., BjÃűrkman, M., Kragic, D.: Bayesian meta-learning for few-shot policy adaptation across robotic platforms. pp. 1274–1280 (09 2021). https://doi.org/10.1109/IROS51168.2021.9636628
27. Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., Olah, C.: Multimodal neurons in artificial neural networks. Distill (2021). https://doi.org/10.23915/distill.00030, https://distill.pub/2021/multimodal-neurons
28. Grover, S., Gopalkrishnan, A., Ai, B., Christensen, H.I., Su, H., Li, X.: Enhancing generalization in vision-language-action models by preserving pretrained representations (2025), https://api.semanticscholar.org/CorpusID:281315107
29. Häon, B., Stocking, K.C., Chuang, I., Tomlin, C.: Mechanistic interpretability for steering vision-language-action models. In: Lim, J., Song, S., Park, H.W. (eds.) Proceedings of The 9th Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 305, pp. 2743–2762. PMLR (27–30 Sep 2025), https://proceedings.mlr.press/v305/haon25a.html
30. Hendel, R., Geva, M., Globerson, A.: In-context learning creates task vectors. arXiv preprint arXiv:2310.15916 (2023)
31. Hendel, R., Geva, M., Globerson, A.: In-context learning creates task vectors. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 9318–9333. Association for Computational Linguistics, Singapore (Dec 2023)
32. Hernandez, E., Li, B.Z., Andreas, J.: Inspecting and editing knowledge representations in language models. arXiv preprint arXiv:2304.00740 (2023), https://arxiv.org/abs/2304.00740
33. Hojel, A., Bai, Y., Darrell, T., Globerson, A., Bar, A.: Finding visual task vectors. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
34. Hojel, A., Bai, Y., Darrell, T., Globerson, A., Bar, A.: Finding visual task vectors. In: European Conference on Computer Vision (ECCV). pp. 257–273. Springer (2025)
35. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
36. Hu, S., Zhao, W., Lin, W., Shen, L., Zhang, Y., Tao, D.: Prompt tuning with diffusion for few-shot pre-trained policy generalization. arXiv preprint arXiv:2411.01168 (2024)
37. Huang, B., Mitra, C., Arbelle, A., Karlinsky, L., Darrell, T., Herzig, R.: Multimodal task vectors enable many-shot multimodal in-context learning. In: Advances in Neural Information Processing Systems. vol. 37, pp. 22124–22153 (2024)

38. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b. ArXiv **abs/2310.06825** (2023)
39. Kanwisher, N.: Domain specificity in face perception. Nature neuroscience **3**(8), 759–763 (2000)
40. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., ..., Levine, S., Finn, C.: Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint arXiv:2403.12945 (2024)
41. Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint arXiv:2502.19645 (2025)
42. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
43. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
44. Li, P., Wu, Y., Xi, Z., Li, W., Huang, Y., Zhang, Z., Chen, Y., Wang, J., Zhu, S.C., Liu, T., Huang, S.: Controlvla: Few-shot object-centric adaptation for pre-trained vision-language-action models. arXiv preprint arXiv:2506.16211 (2025), https://arxiv.org/abs/2506.16211
45. Liang, A., Singh, I., Pertsch, K., Thomason, J.: Transformer adapters for robot learning. In: CoRL 2022 Workshop on Pre-training Robot Learning (2022)
46. Lin, L.H., Cui, Y., Xie, A., Hua, T., Sadigh, D.: Flowretrieval: Flow-guided data retrieval for few-shot imitation learning. Conference on Robot Learning (CoRL) (2024)
47. Lin, Y., Wang, A.S., Sutanto, G., Rai, A., Meier, F.: Polymetis. https://facebookresearch.github.io/fairo/polymetis/ (2021)
48. Lin, Z., Basu, S., Beigi, M., Manjunatha, V., Rossi, R.A., Wang, Z., Zhou, Y., Balasubramanian, S., Zarei, A., Rezaei, K., et al.: A survey on mechanistic interpretability for multi-modal foundation models. arXiv preprint arXiv:2502.17516 (2025), https://arxiv.org/abs/2502.17516
49. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
50. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
51. Liu, Z., Zhang, J., Asadi, K., Liu, Y., Zhao, D., Sabach, S., Fakoor, R.: Tail: Task-specific adapters for imitation learning with large pretrained models. arXiv preprint arXiv:2310.05905 (2023)
52. Ma, Y.J., Hejna, J., Wahid, A., Fu, C., Shah, D., Liang, J., Xu, Z., Kirmani, S., Xu, P., Driess, D., Xiao, T., Tompson, J., Bastani, O., Jayaraman, D., Yu, W., Zhang, T., Sadigh, D., Xia, F.: Vision language models are in-context value learners. arXiv preprint arXiv:2411.04549 (2024). https://doi.org/10.48550/arXiv.2411.04549, https://arxiv.org/abs/2411.04549
53. Mitra, C., Huang, B., Chai, T., Lin, Z., Arbelle, A., Feris, R., Karlinsky, L., Darrell, T., Ramanan, D., Herzig, R.: Sparse attention vectors: Generative multimodal model features are discriminative vision-language classifiers. arXiv preprint arXiv:2412.00142 (2024)

54. Niu, D., Sharma, Y., Biamby, G., Quenum, J., Bai, Y., Shi, B., Darrell, T., Herzig, R.: Llarva: Vision-action instruction tuning enhances robot learning. ArXiv **abs/2406.11815** (2024), https://api.semanticscholar.org/CorpusID:270559839
55. Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al.: In-context learning and induction heads. arXiv preprint arXiv:2209.11895 (2022)
56. OpenAI, :, Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., MÄĚdry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A.T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C.J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D.P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F.P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H.P., Ren, H., Chang, H., Chung, H.W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J.G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J.Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J.W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M.J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M.O.T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N.,

Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R.G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., Malkov, Y.: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276

57. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023)
58. Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., Turner, A.M.: Steering llama 2 via contrastive activation addition. arXiv preprint arXiv:2312.06681 (2023), https://arxiv.org/abs/2312.06681
59. Physical Intelligence, Black, K., Brown, N., Driess, D., Finn, C., Levine, S., et al.: $pi_{0.5}$: A vision-language-action model with open-world generalization. CoRR **abs/2504.16054** (2025)
60. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018), https://api.semanticscholar.org/CorpusID:49313245
61. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T., et al.: A generalist agent. arXiv preprint arXiv:2205.06175 (2022)
62. Schwettmann, S., Watkins, O., Wattenberg, M., Carter, S.: Towards mechanistic interpretability of multimodal neurons. arXiv preprint arXiv:2301.11796 (2023), https://arxiv.org/abs/2301.11796
63. Song, M., Deng, X., Zhong, G., Lv, Q., Wan, J., Li, Y., Hao, J., Guan, W.: Few-shot vision-language action-incremental policy learning. arXiv preprint arXiv:2504.15517 (2025)
64. Sridhar, K., Dutta, S., Jayaraman, D., Lee, I.: Ricl: Adding in-context adaptability to pre-trained vision-language-action models. arXiv preprint arXiv:2508.02062 (2025), https://arxiv.org/abs/2508.02062
65. Steiner, A., Pinto, A.S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., et al.: Paligemma 2: A family of versatile vlms for transfer. arXiv preprint arXiv:2412.03555 (2024)
66. Subramani, N., Suresh, N., Peters, M.: Extracting latent steering vectors from pretrained language models. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 566–581. Association for Computational Linguistics, Dublin, Ireland (May 2022)
67. Team, G.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. ArXiv **abs/2403.05530** (2024)

68. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
69. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)
70. Todd, E., Li, M.L., Sharma, A.S., Mueller, A., Wallace, B.C., Bau, D.: Function vectors in large language models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
71. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv **abs/2302.13971** (2023)
72. Turner, A.M., Thiergart, L., Leech, G., Udell, D., Vazquez, J.J., Mini, U., MacDiarmid, M.: Steering language models with activation engineering. arXiv preprint arXiv:2308.10248 (2024), https://arxiv.org/abs/2308.10248
73. Wang, Y., Ding, P., Li, L., Cui, C., Ge, Z., Tong, X., Song, W., Zhao, H., Zhao, W., Hou, P., Huang, S., Tang, Y., Wang, W., Zhang, R., Liu, J., Wang, D.: VLA-Adapter: An effective paradigm for tiny-scale vision-language-action model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 18638–18646 (2026). https://doi.org/10.1609/aaai.v40i22.38931
74. Wen, J., Zhu, Y., Li, J., Zhu, M., Wu, K., Xu, Z., Liu, N., Cheng, R., Shen, C., Peng, Y., Feng, F., Tang, J.: TinyVLA: Towards fast, data-efficient vision-language-action models for robotic manipulation. IEEE Robotics and Automation Letters pp. 3988–3995 (2025)
75. Xie, A., Chand, R., Sadigh, D., Hejna, J.: Data retrieval with importance weights for few-shot imitation learning. Conference on Robot Learning (CoRL) (2025)
76. Yin, Y., Wang, Z., Sharma, Y., Niu, D., Darrell, T., Herzig, R.: In-context learning enables robot action prediction in llms. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8972–8979. IEEE (2025)
77. Yu, T., Finn, C., Dasari, S., Xie, A., Zhang, T., Abbeel, P., Levine, S.: One-shot imitation from observing humans via domain-adaptive meta-learning. In: Robotics: Science and Systems (RSS) (2018)
78. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 41, pp. 2131–2145 (2018)
79. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., ..., Arenas, M.G., Han, K.: 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)