

VLA's Are Confined yet Capable of Generalizing to Novel Tasks

Quanyi Li

Independent Researcher
quanyili0057@gmail.com

Abstract

Vision-language-action models (VLAs) often perform well on demonstrated tasks but struggle significantly when required to extrapolate or recombine skills used in different tasks in novel ways. For instance, VLAs might learn from the demonstration to put the cream cheese in the bowl and put the bowl on top of the cabinet, yet still fail to put the cream cheese on top of the cabinet. This motivates us to investigate whether VLAs merely overfit to demonstrated tasks or still hold the potential to extrapolate. Our study uses *text latent* as the ingredient; it is a task-specific vector derived from the models' hidden states. It thus encodes semantics necessary for completing a task and can be used to reconstruct the associated task behavior by writing it to the model's residual stream. Furthermore, we find that skills used in distinct tasks can be combined to produce novel behaviors by blending their respective *text latent*. Applying this to π_0 , we increase its success rate from 9% to 83% on the proposed *libero-ood* benchmark, which features 20 tasks extrapolated from standard LIBERO tasks. This reveals that the skill representations encoded in *text-latent* are individual yet composable, while π_0 fails to autonomously combine these representations for extrapolation. This also validates the design of *libero-ood*; it comprises tasks that the model fails, yet should be able to complete. We then tested other VLAs on *libero-ood*, and none of them achieved a success rate higher than 21%. Further analysis reveals VLAs share a common pattern to exhibit spatial overfitting, associating object names with where the object is spatially located in the demonstrated scene rather than achieving true object and goal understanding. Code is available at: <https://github.com/QuanyiLi/pi0-text-latent>.

1 Introduction

Building toward generalist, vision-language-action models (VLAs) trained with large-scale multi-modal datasets [O'Neill *et al.*, 2023] have shown remarkable visual and language generalizability, leading to strong performance on diverse

manipulation tasks [Brohan *et al.*, 2023; Kim *et al.*, 2024; O'Neill *et al.*, 2023; Li *et al.*, 2023b; Huang *et al.*, 2024; *et al.*, 2024; Zhen *et al.*, 2024; Black *et al.*, 2024; Team *et al.*, 2024; Hou *et al.*, 2025; Qu *et al.*, 2025; Zheng *et al.*, 2024]. Typically, for satisfactory deployment on new tasks, VLAs are fine-tuned using demonstrations of target tasks [Kim *et al.*, 2025]. Though this paradigm ensures good in-distribution generalization to light conditions [Li *et al.*, 2024b] and small perturbations of scene layout [Liu *et al.*, 2024], we empirically find that VLAs struggle with out-of-distribution (OOD) tasks, particularly those extrapolated from tasks that they can perform well individually. For instance, a VLA might successfully "put the cream cheese in the bowl" and "put the bowl on top of the cabinet", yet fail to perform the extrapolated task of "put the cream cheese on top of the cabinet", even though the motion primitives required by this task, "picking the cream cheese" and "reaching the top of the cabinet", have already been learned in different tasks. This raises a question: **Do VLAs merely overfit to demonstrated trajectories or do they learn composable representations that support broader generalization?** We study this on the SOTA VLA π_0 [Black *et al.*, 2024] with its *text latent*, which is a vector derived from the model's internal states.

To identify the *text latent* for a given task, we run π_0 on the corresponding task demonstrations and record the hidden states of text tokens for each transformer layer. After this, we average all collected layer-wise features and obtain *text latent*. By writing the *text latent* to the text tokens' residual streams [Elhage *et al.*, 2021] of π_0 , we can reconstruct the associated task trajectory without providing the task prompt. We also find that unembedding *text latent* [nostalgebraist, 2020] produces alternative task prompts, which can instruct the model to finish the corresponding tasks with about 70% success rate. However, these alternative prompts are encrypted and unreadable, enabling private instruction and backdoor attacks. Furthermore, we find that by using *Text Latent Interpolation* (TLI) to blend the representations learned for two tasks, the model can combine respective sub-behaviors or skills to complete novel tasks. Specifically, TLI injects the temporal interpolation of the two respective *text latents* to the residual stream, with the mixing ratio linearly adjusted at each timestep. As a result, the intervention guides the model towards Task 1's behavior at the beginning, and as the episode progresses, the influence smoothly shifts towards Task 2's behavior. As shown

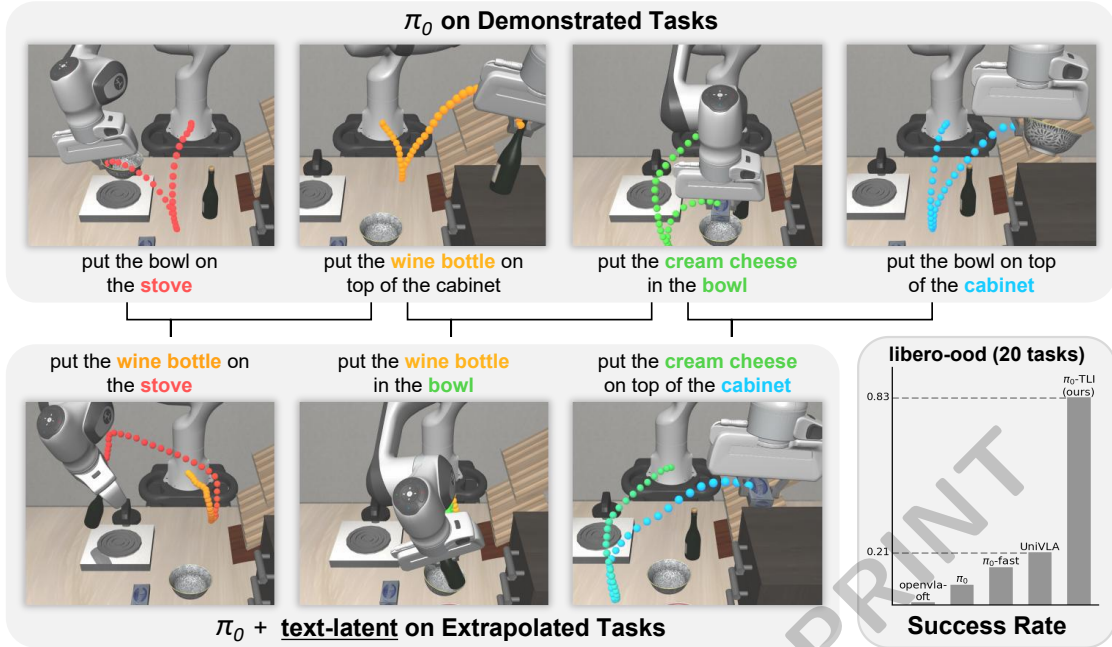


Figure 1: UniVLA, openvla-off, π_0 , and π_0 -fast achieve less than 21% success rate on the proposed *libero-ood* benchmark. It consists of tasks extrapolated from standard LIBERO [Liu *et al.*, 2024] tasks that those VLAs have been fine-tuned with. By applying the proposed *Text Latent Interpolation* (TLI), we improve the performance of π_0 up to 83%. Behaviors of π_0 -TLI on three exemplary extrapolated tasks are shown in the second row, while the first row shows the associated two base tasks.

in Fig. 1, π_0 with *text latent* can stitch together sub-trajectories used in different tasks to finish extrapolated tasks, even if the newly composed trajectories don’t exist in the training data.

To verify that learned representations can broadly support OOD generalization with TLI, we introduce the *libero-ood* task suite. It comprises 20 challenging extrapolated tasks derived from the three standard LIBERO suites: *libero-goal*, *libero-spatial*, and *libero-object*. Each task in *libero-ood* is designed such that while the individual movements required for grasping and placement are present in separate training tasks, the specific combination of these movements is novel. By applying TLI, the success rate of π_0 on *libero-ood* increases from 9% to 83%. And it manages to complete 19 out of 20 tasks at least once in 10 repeated runs. **This significant improvement confirms that the skill representations, encoded in text latents, generally exist and are composable, while the model cannot combine them autonomously and hence failed on novel extrapolated tasks.** On the other hand, this result validates the design of *libero-ood* as a suitable benchmark for OOD generalizability for all VLAs. Rather than presenting VLAs in completely impossible OOD tasks (e.g., holding a steering wheel to drive a car), *libero-ood* comprises OOD tasks that the model has the potential to solve by simply mixing the learned representation. Therefore, we tested other SOTA VLAs on *libero-ood*, including π_0 -fast [Pertsch *et al.*, 2025], openvla-off [Kim *et al.*, 2025], and UniVLA [Bu *et al.*, 2025]. Though they achieve about a 95% success rate on the three standard LIBERO task suites after fine-tuning, their success rate on *libero-ood* is less than 21%, highlighting their limitations in extrapolation. Further qualitative analysis reveals that VLAs commonly exhibit spatial overfitting, asso-

ciating object names with their locations in the demonstrated scene. When instructed to pick an object, they always move to where the object has been placed before, ignoring its current location. It uncovers that VLAs fail to learn true object or goal understanding even after fine-tuning.

2 Related Work

2.1 Vision-Language-Action Models

VLAs are usually initialized from vision-language models (VLMs) pretrained with large-scale cross-modality data. To adapt it for decision-making, we need to further train it with large-scale cross-embodiment datasets like Open X-Embodiment [O’Neill *et al.*, 2023], following a fine-tuning on specific tasks with fewer demonstrations [Brohan *et al.*, 2023; O’Neill *et al.*, 2023; Li *et al.*, 2023b; Kim *et al.*, 2024; Durante *et al.*, 2024; Huang *et al.*, 2024; et al., 2024; Zhen *et al.*, 2024; Black *et al.*, 2024; Team *et al.*, 2024; Hou *et al.*, 2025; Bu *et al.*, 2025; Qu *et al.*, 2025; Wang *et al.*, 2024; NVIDIA *et al.*, 2025; Team, 2025; Pertsch *et al.*, 2025; Zheng *et al.*, 2024]. Though VLAs are all built upon the transformer-based pre-trained VLMs [Vaswani *et al.*, 2023], it remains unclear what is the best way to decode actions. OpenVLA [Kim *et al.*, 2024], RT-2 [Brohan *et al.*, 2023], and π_0 -fast [Pertsch *et al.*, 2025] follow the next-token prediction manner used by VLMs and LLMs and decode discrete action tokens, which will be converted to continuous actions according to different tokenization schemes. Another widely adopted way to predict action is to use a regression head or diffusion model with extra parameters [NVIDIA *et al.*, 2025; Black *et al.*, 2024; Kim *et al.*, 2024; Wen *et al.*, 2024;

Lee *et al.*, 2024; Chi *et al.*, 2023; Kim *et al.*, 2025; Li *et al.*, 2024a]. As the VLM is the core component for VLAs’ ability, we choose to mine the meaningful internal representation of the transformer part of π_0 , which adopts an additional action expert to generate continuous actions with flow matching.

2.2 Mechanistic Interpretability (MI)

This is a subfield of interpretability, where researchers reverse engineer the model’s internal computations to understand how it works [Rai *et al.*, 2025; Elhage *et al.*, 2021; Vaswani *et al.*, 2023] and reconstruct or activate certain behaviors. Some early works analyze image classification models [Olah *et al.*, 2020; Bau *et al.*, 2017; Zhou *et al.*, 2015] and find that neurons play the role of feature detectors for patterns from simple curves to complex objects like cats [Cammarata *et al.*, 2020]. Recent studies of MI on LLM and VLM identifies important circuits that can perform specific tasks like induction head [Olsson *et al.*, 2022] to output repeating words, function vectors [Todd *et al.*, 2024; Luo *et al.*, 2024] to produce antonyms, attention head to detect number or shape [Gandelsman *et al.*, 2024], neurons contributing to recognize specific objects [Gandelsman *et al.*, 2024], and internal features making VLMs hallucinate [Jiang *et al.*, 2025]. However, there are limited works that study the neural policies or VLAs from the MI perspective. The most relevant works study adversarial attack with feature attribution [Wang *et al.*, 2025], symbolic representation uncovering [Lu *et al.*, 2025], and motion-relevant neuron identifying and characterization [Li *et al.*, 2023a; Häon *et al.*, 2025]. Our work finds causal effects between the internal representations and VLAs’ behaviors. As a result, we can produce obscure prompts with *logit lens* [nostalgebraist, 2020], enabling private instruction or backdoor attack. In addition, the identified functional component enables the π_0 for extrapolated tasks, where all SOTA VLAs struggle.

3 Method

We start by formulating how transformer-based VLAs work. Then, we illustrate how to identify *text latent* and use them to change the model’s internal features, steering its behavior.

3.1 Preliminary

Existing VLAs adopt VLMs as encoders to fuse both vision and language information. Concretely, a pre-trained vision encoder, e.g., CLIP [Radford *et al.*, 2021] and SigLIP [Zhai *et al.*, 2023], is used to generate d -dimensional sequential image embeddings from image patches, followed by d -dimensional text embeddings that are tokenized and projected from the task description. For some VLAs [Kim *et al.*, 2024], the task description is encapsulated with certain context, like “what actions the robot should take to {*task description*}”, which introduces extra tokens. In addition, the proprioceptive state will be projected [Kim *et al.*, 2024] or tokenized [Pertsch *et al.*, 2025] into the d -dimensional shared space to work with image-text embeddings. After tokenization, we assume there are totally m d -dimensional embeddings, $e = \{e^i : e^i \in \mathbb{R}^d, i = 1 \dots m\}$, obtained from image, text, and robot proprioception, which will go through L transformer layers.

Except for the last layer, each layer l outputs hidden states $h_l = \{h_l^i : h_l^i \in \mathbb{R}^d, i = 1 \dots m\}$. The action is then generated by $a = f(e, h)$, where $h = \{h_l : h_l \in \mathbb{R}^{m \times d}, l = 1 \dots L - 1\}$ is hidden states of all tokens across $L - 1$ layers. More precisely, the action generation uses the KV cache derived from e and h , where the image-text embeddings e are passed through the first transformer layer (layer 0) to compute key and value projections, and the hidden states h_l are passed through their subsequent layer $l + 1$ to compute the corresponding KV projections. To simplify the notation, we condition the action generation on e and h . This formulation applies to most existing VLAs as long as the VLM is reused to do action generation in an autoregressive way [Kim *et al.*, 2024; Qu *et al.*, 2025; Brohan *et al.*, 2023; Pertsch *et al.*, 2025] or the extra action prediction module is also transformer-based [NVIDIA *et al.*, 2025; Black *et al.*, 2024], which can take the KV-cache to do causal self-attention.

3.2 Text Latent

As VLAs’ behavior depends on the task description, we hypothesize that the skill representations are embedded in the internal representations of text (task description) tokens. To study this mechanism, we denote the text embeddings as $e^T = \{e^i : e^i \in \mathbb{R}^d, i \in T\}$ and their hidden state at each layer as $h_l^T = \{h_l^i : h_l^i \in \mathbb{R}^d, i \in T\}$, where T is the set of text tokens indices. Therefore, the hidden states of all text tokens across all $L - 1$ layers can be represented by a tensor as $h^T \in \mathbb{R}^{L-1 \times |T| \times d}$ after preserving orders and stacking them. By denoting the rest of the embeddings and their hidden states as $e^- = e \setminus e^T$ and $h^- = h \setminus h^T$, we can update the action generation function at timestep i as

$$a = f(e^T, h^T(i), e^-(i), h^-(i)) \quad (1)$$

Note that text embedding, e^T , doesn’t condition on i , since it is fixed throughout the whole episode, while the rest image and proprioceptive tokens are changed at different timesteps. Our goal is to manipulate the $h^T(i)$ at each decision-making step to activate behaviors with specific semantics.

Text latent is the ingredient for doing this. It has the same shape as h^T and thus can be written into the text tokens’ residual stream. We identify it by averaging a set of text hidden states $\{h^T(i) : h^T(i) \in \mathbb{R}^{L-1 \times n \times d}, i \in B\}$, where B is all timestep indices of a demonstrated episode for the target task, and $h^T(i)$ is thus obtained by forwarding the model with the observation at timestep i . As multiple demonstrated episodes exist for a single task in the training sets, the average is thus taken over K demonstrations. Therefore, the *text latent* can be calculated by element-wise average:

$$\mathcal{T} = \frac{1}{\sum_{k=1}^{K-1} |B_k|} \sum_{k=1}^{K-1} \sum_{i \in B_k} h^T(i) \quad (2)$$

In our reconstruction experiment, we demonstrate that $\mathcal{T}$ captures the most essential knowledge for finishing the corresponding task. And blending them enables skill combination for task extrapolation.

3.3 Text Latent Interpolation

A task derived from two base tasks can be interpreted as starting with Task 1 and gradually switching to Task 2. Though

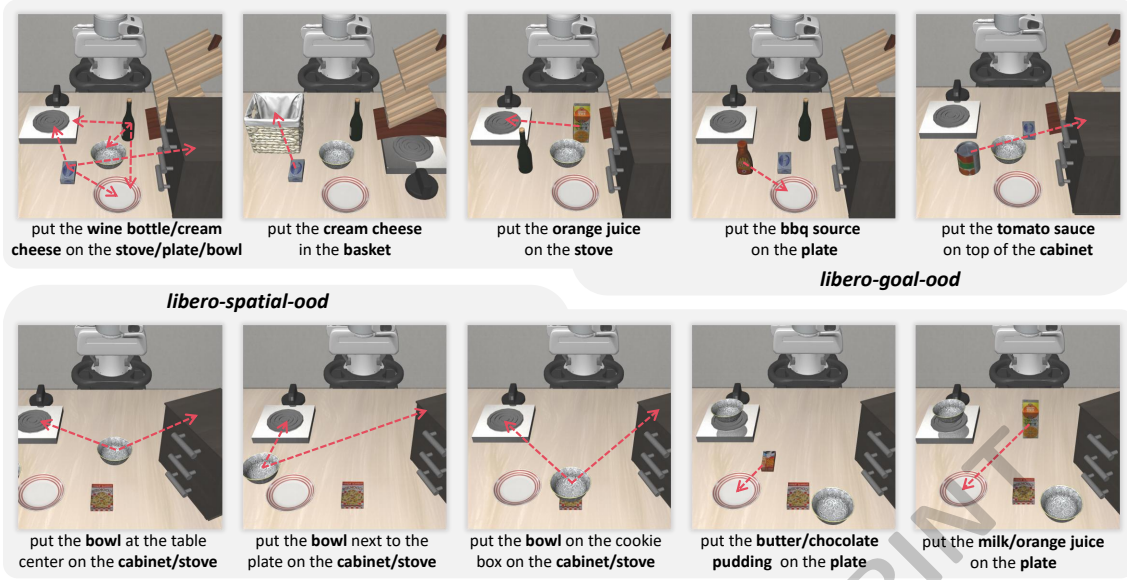


Figure 2: Visualization of scene layouts, object to pick and where to place (denoted by the red arrows), and prompts for tasks in *libero-ood*, which can be further split into *libero-goal-ood* and *libero-spatial-ood*. *libero-goal-ood* includes six extrapolated tasks that require combining sub-skills learned from *libero-goal* and operating in the same scene layout as *libero-goal* tasks (the top-left figure), while the remaining four tasks slightly change the layout and additionally demand transferring knowledge about objects. On the other hand, *libero-spatial-ood* provides six tasks that require transferring the skills learned in *libero-goal* and *libero-spatial* to place the object on the cabinet or stove. The remaining four tasks require placing unseen objects from *libero-object* at new locations on the plate.

text token ids are discrete and cannot be changed smoothly throughout the episode, we can approximate a continuous transition by linearly interpolating the text embeddings of the two base task prompts, e_1^T and e_2^T , at timestep i . This is called *Text Embedding Interpolation* (TEI).

$$e^T = e^T(i) = (1 - \alpha)e_1^T + \alpha e_2^T, \quad \alpha = i/\lambda, \quad 0 \leq i \leq \lambda,$$

where the hyperparameter λ controls the transition speed. The λ is set to 24, the average number of policy-execution steps, for π_0 , except for *put the wine bottle in the bowl*. As the wine bottle is near the bowl, we shorten λ to 14 for this task, so the second task behavior is activated sooner.

Intuitively, TEI rewrites the task prompt at every step with a weighted blend of the two source instructions. Additionally, we can leave the target task prompt (and thus its embedding) unchanged and operate on the model’s residual stream with the two respective *text latent* ($\mathcal{T}^1, \mathcal{T}^2$). We term this approach *Text Latent Interpolation* (TLI), which modifies the text hidden states $h^T(i)$ as follows:

$$h^T(i) = h^T(i) + [(1 - \alpha)\mathcal{T}^1 + \alpha\mathcal{T}^2] - [(1 - \alpha)\mathcal{T}^2 + \alpha\mathcal{T}^1]$$

At the beginning of the episode, Task 2’s context is suppressed and subtracted from the residual stream. As the episode progresses ($\alpha : 0 \rightarrow 1$), Task 2’s context is gradually injected into the residual stream while Task 1’s context fades out and is finally suppressed. TLI can also be applied to a specific layer l by replacing $h_l^T(i)$ with the interpolated pair $(\mathcal{T}_l^1, \mathcal{T}_l^2)$ in the same fashion. The full interpolation procedure is listed in Algorithm 1. Both TEI and TLI can be deployed independently or jointly. In either case, the ratio i/λ is clipped between 0 and 1. If the length of the text dimension of $\mathcal{T}^1$ or $\mathcal{T}^2$ differs

from the number of tokens of the target prompt, we truncate it or pad it with zeros to match the length of the target prompt, as the task descriptions are of a similar length.

Algorithm 1 TEI and TLI during Task Execution

Require: *text latent* $\mathcal{T}^1, \mathcal{T}^2$, Interpolation steps λ , Initial observations o , Max timestep J

for $i = 1$ to J **do**

$e^T, h^T(i), e^-(i), h^-(i) \leftarrow \text{encode}(o)$

if *Text Embedding Interpolation* (TEI)

$$e^T = (1 - i/\lambda)e_1^T + (i/\lambda)e_2^T$$

if *Text Latent Interpolation* (TLI)

$$h^T(i) += (1 - i/\lambda)(\mathcal{T}^1 - \mathcal{T}^2) + i/\lambda(\mathcal{T}^2 - \mathcal{T}^1)$$

$a = f(e^T, h^T(i), e^-(i), h^-(i))$ ▷ Decode action

$o \leftarrow \text{simulation}(a)$ ▷ Forward simulation

end for

4 Experiments

Benchmark. We conducted experiments with LIBERO [Liu *et al.*, 2024], a simulation environment widely employed for evaluating Vision-Language-Action (VLA) models [Black *et al.*, 2024; Pertsch *et al.*, 2025; Kim *et al.*, 2025; Qu *et al.*, 2025; Zheng *et al.*, 2024; Bu *et al.*, 2025]. We use three standard task suites from LIBERO: *libero-goal*, *libero-object*, and *libero-spatial*. Each suite contains 10 tasks, and most of

Original Prompt	Prompt unembedded from <i>text-latent</i> (tokens)
pick up the black bowl next to the plate and place it on the plate	'Δϱ' 'QSize', 'U000e0062', 'black', 'Δϱ' 'Δϱ' 'QSize', 'U000e0062', 'plate', 'QSize', 'QSize', 'QSize', 'QSize', 'U000e0062', 'QSize'
pick up the black bowl on the ramekin and place it on the plate	'Δϱ' 'Δϱ' 'U000e0062', 'black', 'Δϱ' 'U000e0062', 'U000e0062', 'Δϱ' 'kin', 'QSize', 'QSize', 'Δϱ' 'U000e0062', 'U000e0062', 'QSize'
pick up the black bowl on the stove and place it on the plate	'Δϱ' 'Δϱ' 'U000e0062', 'black', 'Δϱ' 'U000e0062', 'U000e0062', 'stove', 'QSize', 'QSize', 'Δϱ' 'U000e0062', 'U000e0062', 'QSize'
pick up the black bowl from table center and place it on the plate	'Δϱ' 'QSize', 'U000e0062', 'black', 'QSize', 'FROM', 'table', 'Δϱ' 'QSize', 'QSize', 'Δϱ' 'U000e0062', 'U000e0062', 'Δϱ'

Table 1: Four exemplary Prompts decoded from $\mathcal{T}_3$ for *libero-spatial* tasks. These human-unreadable prompts can instruct π_0 to finish tasks with around 70% success rate. This enables adversarial attacks and private instructions.

them require pick-and-place. We also introduce a novel task suite, *libero-ood*, comprising two sub-suites *libero-goal-ood* and *libero-spatial-ood*. Each also contains 10 extrapolated tasks. The key idea behind these 20 tasks is to ensure that both the grasping and placement locations have individually appeared in the training data, so the policy has already learned how to reach these locations. However, the specific trajectory connecting these two locations has not been demonstrated. Thus, solving these tasks requires the policy to stitch together sub-trajectories it has learned from demonstrated tasks.

Experiments. Our main experimental goal is to examine whether π_0 learns individual yet composable task representations, so it can complete these 20 extrapolated tasks by combining learned representations, encoded in *text latent*. For each task in the three standard suites, we identify the corresponding *text latent* using 20 demonstrations, calculated according to equation 2. We first demonstrate on π_0 that *text latent* encapsulates the essential knowledge required to complete standard LIBERO tasks by reconstructing the task behavior without clear task prompts. Second, we show that by interpolating between *text latent* or *text embeddings*, π_0 can complete challenging extrapolated tasks in *libero-ood* suite, where tasks pose difficulties for current state-of-the-art VLAs. Finally, our analysis across specific tasks reveals that VLAs’ behavior relies on the mechanistic memorization of demonstrated trajectories, rather than genuine object recognition.

Evaluation. For each task, we execute 10 independent runs using different random seeds. Therefore, the final success rate for each task suite is calculated as the proportion of successful episodes across the total 100 runs (10 tasks $\times$ 10 runs/task). All experiments are conducted with Nvidia RTX 4090.

4.1 Task Reconstruction

In this experiment, we use *text latent* to reconstruct the behavior or trajectory learned for the three standard task suites, demonstrating that individual task representation is encoded in *text latent*. We use two ways to examine the effectiveness of *text latent*. The first way is to mask all the text tokens or set each text token to the space character (" ") so the text prompt doesn’t provide any task information. We can then add back the *text latent* to each hidden layer representation when executing the policy. The second way is to unembed the early layer vectors of *text latent* into a set of token ids [nostalgebraist, 2020], producing alternative prompts. Thus, we can feed the alternative prompt to the model directly.

The results are shown in Table 2. The first line is the official performance of π_0 with the original task prompt, serving as the upper bound. The *Mask Prompt* experiment uses only

	libero-object	libero-goal	libero-spatial
<i>Original</i>	<i>0.98</i>	<i>0.95</i>	<i>0.97</i>
Mask Prompt	0.11	0.14	0.24
Blank Prompt	0.28	0.16	0.23
Blank Prompt+ $\mathcal{T}$	0.94	0.82	0.81
$\mathcal{T}_1$ -Prompt	<u>0.92</u>	<u>0.66</u>	0.98
$\mathcal{T}_2$ -Prompt	0.73	0.19	<u>0.85</u>
$\mathcal{T}_3$ -Prompt	0.56	-	0.68

Table 2: The success rate of different ways to reconstruct tasks.

image input for all task execution, while the *Blank Prompt* adds back text tokens but fills with space characters (" "). In both settings, the policy has no task instructions, indicating the lower bound of performance. The rest lines show the performance of different ways to reconstruct the task. For each task suite, the settings yielding the best performance are bolded, and the second-best performance is underlined. The experiment *Blank Prompt+ $\mathcal{T}$* shows that by writing the *text latent* to the model’s residual stream, $h^T(i) = h^T(i) + \mathcal{T}$, the tasks can be finished with a success rate higher than 80%, even if the prompt is blanked and provides none of the task information. **Thus, it confirms that *text latent* captures essential task knowledge, and can remind π_0 of a task by injecting it into the model’s internal states.**

As $\mathcal{T} \in \mathbb{R}^{L-1 \times n \times d}$, we can thus applying the language embedding matrix E to unembed $\mathcal{T}_l$ into tokens by calculating the cosine similarity between each column of E and $\mathcal{T}_l$, then selecting the indices with maximum similarity as new tokens, composing the alternative prompt. As shown in Table 2, we unembed $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ and find that for *libero-goal*, the prompt produced by unembedding $\mathcal{T}_1$ achieves a 66% success rate, while the next layer’s vector can not reconstruct the task well. However, for *libero-object* and *libero-spatial*, even the prompt unembedded by $\mathcal{T}_2$ and $\mathcal{T}_3$ can reconstruct tasks with an average success rate of 70%. **We find most prompts unembedded from *text latent* are encrypted and unreadable, even for people who are familiar with the original prompts.** This enables an application, obscure prompting, for private instruction or backdoor attacks. The unembedded prompts for four example tasks from *libero-spatial* are shown in table 1.

4.2 Task Extrapolation

As *text latent* encodes task context or skill representations, we then ask whether the behaviors learned from separate tasks can be recombined using the respective *text latent*. We refer to any task that demands such a behavior recombination as an

	UniVLA	openvla-oft	π_0 -fast	π_0	π_0 -TEI	π_0 -TLI	π_0 -TLI ⁺
libero-goal-ood	0.32	0.01	0.13	0.02	0.02	0.85	0.85
libero-spatial-ood	0.11	0	0.17	0.16	0.16	0.81	0.69
All	0.21	0.01	0.15	0.09	0.09	0.83	0.77

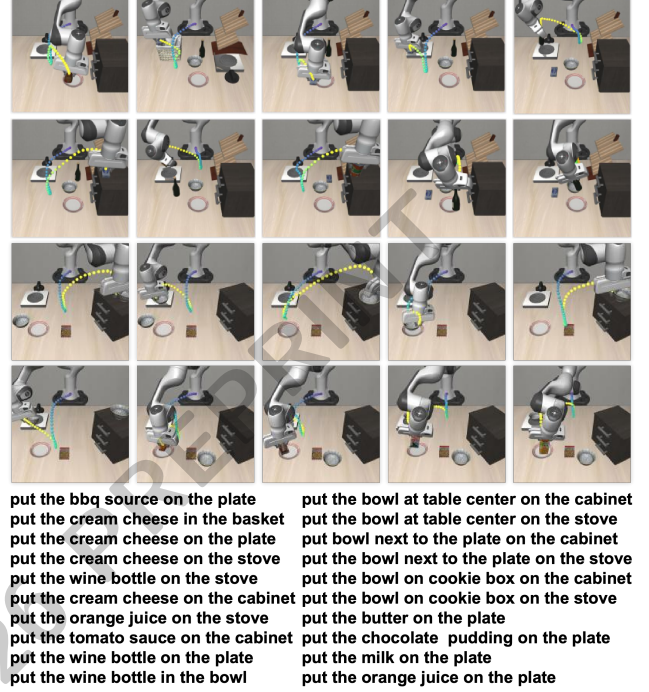
Table 3: The success rate of SOTA VLAs and our methods on the *libero-goal-ood* task suite.

extrapolated task. These new tasks keep objects’ locations or layouts the same as the scenes for collecting training demonstrations. Thus, for each *libero-ood* task, VLAs have learned to reach the grasping and placement locations, respectively. However, π_0 only shows a 9% success rate in *libero-ood*. After applying the proposed TLI for π_0 , we can largely improve its success rate to 83%. **This confirms that skill representations generally exist in π_0 that can be rearranged to solve novel tasks. It also validates that the tasks in *libero-ood* are within π_0 ’s capability, while it fails on it.** We further test other SOTA VLAs’ extrapolation ability using *libero-ood*. The results are shown in Table 3. None of the VLAs achieves a success rate higher than 21%. Among all VLAs, the best is UniVLA. It shows trajectory stitching behavior in some tasks, and better language following ability, which results from the disentangled training of the vision-language module and the action module [Huang *et al.*, 2025].

However, simply applying the TEI doesn’t produce good results. It suggests that latent stores more and critical information to retrieve specific skills. We also applied both TEI and TLI to the model inference together, yielding π_0 -TLI⁺. The results show that TEI doesn’t further improve the performance of π_0 -TLI for *libero-goal-ood*, while it even slightly harms the performance of *libero-spatial-ood*. We will use π_0 -TLI⁺ to reveal the spatial overfitting exhibited in π_0 in the next section.

4.3 Spatial Overfitting of π_0

As shown in Fig. 2, the rightmost 4 tasks in *libero-goal-ood* require the model to pick up objects that have never been shown in the scene of *libero-goal* before. In these 4 tasks, we swap the original objects with new ones from *libero-object*, while still keeping the displaced objects in the scene. Using both TEI and TLI, π_0 achieves an average success rate of 85% on these four tasks, which confirms the existence of *Spatial Overfitting* in π_0 . Specifically, consider the task *put the orange juice on the stove*. The orange juice is placed at the original position of the wine bottle, and the wine bottle is moved to the former position of the cream cheese. We can instruct π_0 to pick the orange juice, using the *text latents* and text embeddings of *put the wine bottle on top of the cabinet* with π_0 -TLI⁺. Both text embedding and hidden states request π_0 to pick the wine bottle, while π_0 ignores the current location of the wine bottle that is reachable at the location of the cream cheese, and it still mechanically moves to where the wine bottle is placed in the demonstrated scene. The trajectories produced by π_0 -TLI⁺ for all four tasks are shown in Fig. 4. **This behavior indicates that the π_0 does not understand object identity but instead maps object names to fixed locations, termed as *Spatial Overfitting*.** Thus, “cream cheese” actually means “the object at the location where the cream cheese appeared during training.” And π_0 can grasp anything placed at the location of

Figure 3: Trajectories resulted from π_0 -TLI on *libero-ood*. The only failure case is *put the bowl next to the plate on the stove*.

the cream cheese with this prompt.

4.4 Spatial Overfitting is Common in VLAs

	π_0	π_0 -fast	openvla-oft	UniVLA
Two-prompt	0.94	0.86	0.99	0.90
OOD-position	0	0	0	0

Table 4: The success rate on modified *libero-object* suite

The spatial overfitting commonly exists for VLAs besides π_0 , which can be confirmed by experiments on *libero-object* suite, where five tasks place the target object at the center of the scene, and the other five place it in the top-right corner. After fine-tuning with demonstrations, VLAs can always pick the central object with any of the five “center” prompts and the top-right object with any of the five “top-right” prompts. In the two-prompt experiments shown in Table 4, we always use the prompt, *pick up the cream cheese* to pick the object at the scene center, and the prompt, *pick up the alphabet soup* for the objects at the top-right corner. We then run the *libero-object* with the two prompts. All VLAs can maintain the performance on the *libero-object* with incorrect prompts, even though the

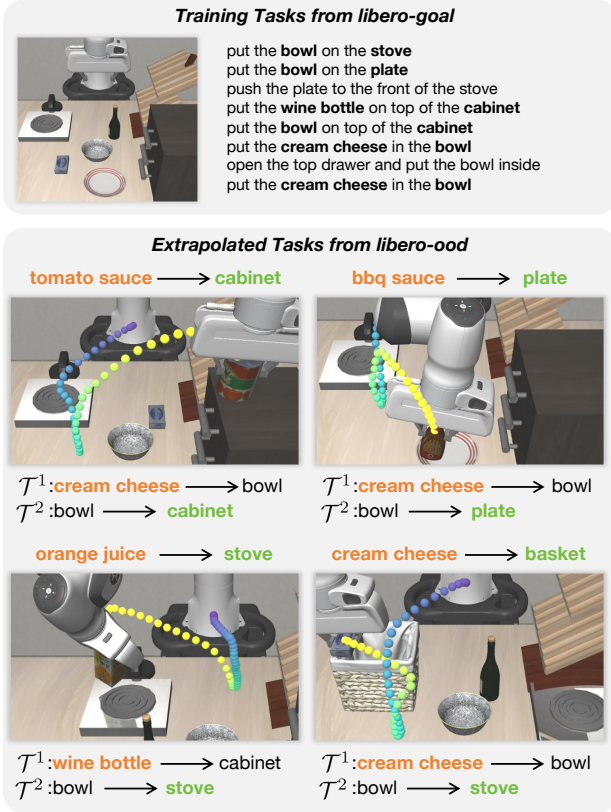


Figure 4: This figure shows the original tasks from *libero-ood* and four extrapolated tasks, requiring recombining the skills learned from original tasks. Trajectories of π_0 -TLI⁺ for the four tasks are shown. Object to grasp and place to drop are shown on top of the figure; The two *text latent* used to complete the task are listed underneath. These behaviors reveal spatial overfitting. For example, we can use the *text latent* and text embedding of "placing on the stove" to put any object into the basket that occupied the original location of the stove in the training data, ignoring the current location of the stove.

"cream cheese" and "alphabet soup" still appear in the scene.

In addition, when the target object is located anywhere other than the center or top-right, the policy fails. In the OOD-position experiment shown in table 4, we relocated the target objects to a place other than the centre or the top-right corner. As expected, all VLAs failed. The end effector still travelled to the original location of the target object, picked up whatever object happened to occupy that spot.

As *text latent* encodes the task context, we can analyze which part of the image observation contributes most to the formation of *text latent* for π_0 with logit lens [nostalgebraist, 2020; Jiang *et al.*, 2025]. Fig. 5 highlights the information extracted from the image observation to build *text-latent* in 4 *libero-object* tasks at different timesteps. We find that the π_0 always absorbs the information about the destination and the robot pose to build *text latent*, with less attention on the object to grasp or treat all objects equally. Thus, π_0 relies on the text instruction instead of image information to differentiate the two grasping locations. This suggests a poor vision-language alignment and shortcut learning.

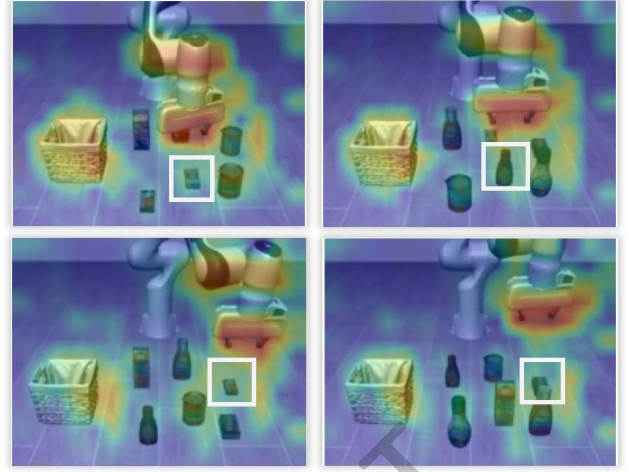


Figure 5: The object to grasp is framed in white. Even if the end effector is close to the object, the π_0 model is still not interested in the target object. Also, during the inference, π_0 always focuses on the posture of the end effector, suggesting that the image input is mainly for estimating the robot state.

5 Conclusion

Existing VLAs show promising in-distribution generalization after fine-tuning, but their out-of-distribution (OOD) robustness is understudied due to a lack of suitable benchmarks. This is primarily because it is difficult to determine how far new tasks should deviate from the training data distribution. For instance, it would be unreasonable to expect a robot arm trained solely on tabletop manipulation to hold a steering wheel and drive a car. An ideal OOD benchmark should therefore comprise tasks that VLAs have the potential to complete but currently cannot. In this work, we find that by using the learned internal representation, π_0 can be augmented to solve novel tasks that it can not complete on its own. Since the injected knowledge is derived from the model's own learned representations, we conclude that π_0 inherently possesses the capability to complete these new tasks, although this ability appears to be latent or "locked". Tasks constructed in this manner are thus well-suited for our purpose, leading us to introduce an OOD benchmark named *libero-ood*. Besides π_0 , we further evaluate other SOTA VLAs on LIBERO, including π_0 -fast, UniVLA, and openvla-oft. Unfortunately, none of them achieved a success rate higher than 21%. Their failure may stem from the same issue as π_0 , an inability to recombine learned representations into new skills, or skill representations don't emerge during the training. We thus make *libero-ood* public to encourage investigating their failure mode and developing more generalist models without using tricks like TLI.

Ethical Statement

This work aims to provide insights into building robust VLAs. However, the ingredient, *text latent*, can be used to compose obscure prompts, making the action goal untransparent.

Acknowledgments

We thank Bolei Zhou for donating the GPU used in this work.

References

- [Bau *et al.*, 2017] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations, 2017.
- [Black *et al.*, 2024] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- [Brohan *et al.*, 2023] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
- [Bu *et al.*, 2025] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions, 2025.
- [Cammarata *et al.*, 2020] Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020. <https://distill.pub/2020/circuits>.
- [Chi *et al.*, 2023] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [Durante *et al.*, 2024] Zane Durante, Bidipta Sarkar, Ran Gong, Rohan Taori, Yusuke Noda, Paul Tang, Ehsan Adeli, Shrinidhi Kowshika Lakshmikanth, Kevin Schulman, Arnold Milstein, et al. An interactive agent foundation model. *arXiv preprint arXiv:2402.05929*, 2024.
- [Elhage *et al.*, 2021] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>.
- [Et al., 2024] Andrew Sohn et al. Introducing rfm-1: Giving robots human-like reasoning capabilities, 2024.
- [Gandelsman *et al.*, 2024] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip’s image representation via text-based decomposition, 2024.
- [Hou *et al.*, 2025] Zhi Hou, Tianyi Zhang, Yuwen Xiong, Hengjun Pu, Chengyang Zhao, Ronglei Tong, Yu Qiao, Jifeng Dai, and Yuntao Chen. Diffusion transformer policy: Scaling diffusion transformer for generalist vision-language-action learning, 2025.
- [Huang *et al.*, 2024] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.
- [Huang *et al.*, 2025] Huang Huang, Fangchen Liu, Letian Fu, Tingfan Wu, Mustafa Mukadam, Jitendra Malik, Ken Goldberg, and Pieter Abbeel. Otter: A vision-language-action model with text-aware visual feature extraction, 2025.
- [Häon *et al.*, 2025] Bear Häon, Kaylene Stocking, Ian Chuang, and Claire Tomlin. Mechanistic interpretability for steering vision-language-action models, 2025.
- [Jiang *et al.*, 2025] Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations, 2025.
- [Kim *et al.*, 2024] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [Kim *et al.*, 2025] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success, 2025.
- [Lee *et al.*, 2024] Seungjae Lee, Yibin Wang, Haritheja Etukuru, H Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024.
- [Li *et al.*, 2023a] Quanyi Li, Zhenghao Peng, Haibin Wu, Lan Feng, and Bolei Zhou. Human-ai shared control via policy dissection, 2023.
- [Li *et al.*, 2023b] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023.
- [Li *et al.*, 2024a] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen, Yuanchun Shi, Jiaolong Yang, and Baining Guo. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation, 2024.
- [Li *et al.*, 2024b] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation, 2024.
- [Liu *et al.*, 2024] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

- [Lu *et al.*, 2025] Hong Lu, Hengxu Li, Prithviraj Singh Shahani, Stephanie Herbers, and Matthias Scheutz. Probing a vision-language-action model for symbolic states and integration into a cognitive architecture, 2025.
- [Luo *et al.*, 2024] Grace Luo, Trevor Darrell, and Amir Bar. Task vectors are cross-modal, 2024.
- [nostalgebraist, 2020] nostalgebraist. Interpreting GPT: The logit lens. LessWrong, Aug 2020.
- [NVIDIA *et al.*, 2025] NVIDIA, :, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llonet, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025.
- [Olah *et al.*, 2020] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. <https://distill.pub/2020/circuits/zoom-in>.
- [Olsson *et al.*, 2022] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- [O'Neill *et al.*, 2023] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.
- [Pertsch *et al.*, 2025] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models, 2025.
- [Qu *et al.*, 2025] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. Spatialvla: Exploring spatial representations for visual-language-action model, 2025.
- [Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [Rai *et al.*, 2025] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models, 2025.
- [Team *et al.*, 2024] 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.
- [Team, 2025] Gemini Robotics Team. Gemini robotics: Bringing ai into the physical world, 2025.
- [Todd *et al.*, 2024] Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models, 2024.
- [Vaswani *et al.*, 2023] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
- [Wang *et al.*, 2024] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers, 2024.
- [Wang *et al.*, 2025] Taowen Wang, Cheng Han, James Chenhao Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics, 2025.
- [Wen *et al.*, 2024] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*, 2024.
- [Zhai *et al.*, 2023] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.
- [Zhen *et al.*, 2024] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.
- [Zheng *et al.*, 2024] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies, 2024.
- [Zhou *et al.*, 2015] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns, 2015.