THE LINEAR REPRESENTATION HYPOTHESIS FOR VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The *linear representation hypothesis* (LRH) has become a standard lens for measuring and intervening on semantic information through the internal representations of large language models (LLMs). A growing body of work has begun extending this perspective to vision-language-action (VLA) models, but the dynamical nature of embodied interaction introduces an additional challenge. Unlike semantic attributes commonly studied in LLMs, such as gender or language, a physical quantity of interest (QoI) in a VLA evolves jointly with the system dynamics: the representation influences the actions selected by the policy, which alter the physical state and, in turn, the next representation.

In this paper, we develop a theoretical, *signature*-based formulation of the LRH for VLA that unifies representations and policies. On the representation side, we establish the existence of representations from which the future evolution of a QoI under a candidate action trajectory can be recovered via *linear probing*. On the policy side, we introduce a *signature generalized linear model* for stochastic action chunks. This structure yields a monotonic change in the expected future QoI along linear paths in natural parameter space, enabling *linear steering*. We construct an explicit oracle representation in a planar control-affine navigation experiment and verify the predicted linear probing and steering mechanisms.

1 INTRODUCTION

The linear representation hypothesis (LRH) in large language models (LLMs) states that high-level concepts, such as gender and language, have representations that occupy particular linear directions, or low-dimensional subspaces, in their representation space (Mikolov et al., 2013; Elhage et al., 2022; Park et al., 2024). Its appeal is that this geometry provides a simple mathematical interface to otherwise opaque models. After identifying a linear representation corresponding to a concept, it can be used in two ways: *probing* measures the concept by projecting representations onto that direction (Alain & Bengio, 2016; Belinkov, 2022), while *steering* intervenes on the concept by moving representations along it (Li et al., 2023b; Turner et al., 2023; Zou et al., 2023). These two operations have become a standard interface to understand the internal structures of LLMs, supported by substantial empirical evidence (Li et al., 2023a; Marks & Tegmark, 2024; Templeton et al., 2024) and, more recently, by systematic theoretical analysis (Park et al., 2024; Jiang et al., 2024; Park et al., 2025; Garg et al., 2026).

It is natural to seek the same interface for robot control with vision- language-action (VLA) models, where the motivation is even stronger. VLA models (Brohan et al., 2023; Kim et al., 2024; Ghosh et al., 2024; Black et al., 2025) are large, expensive to fine-tune, and trained on demonstration data that is costly to collect (Open X-Embodiment Collaboration, 2024; Khazatsky et al., 2024); retraining is rarely an option, e.g., when a policy must be made more cautious or when one needs to understand what the policy has inferred from the scene. The representation space offers a cheaper point of access. If task-relevant physical information is encoded in a representation space and can be measured, then we can use it without additionally instrumenting the environment or modifying the model weights. If behavior is steerable, it can be adjusted at inference time without collecting new robot demonstrations. Recent work has increasingly built on this premise (Lu et al., 2025; Häon et al., 2025; Zhang et al., 2026; Miao et al., 2026), bringing much of the LLM-side LRH framework into VLAs: physical concepts, their representations, and the probing and steering.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

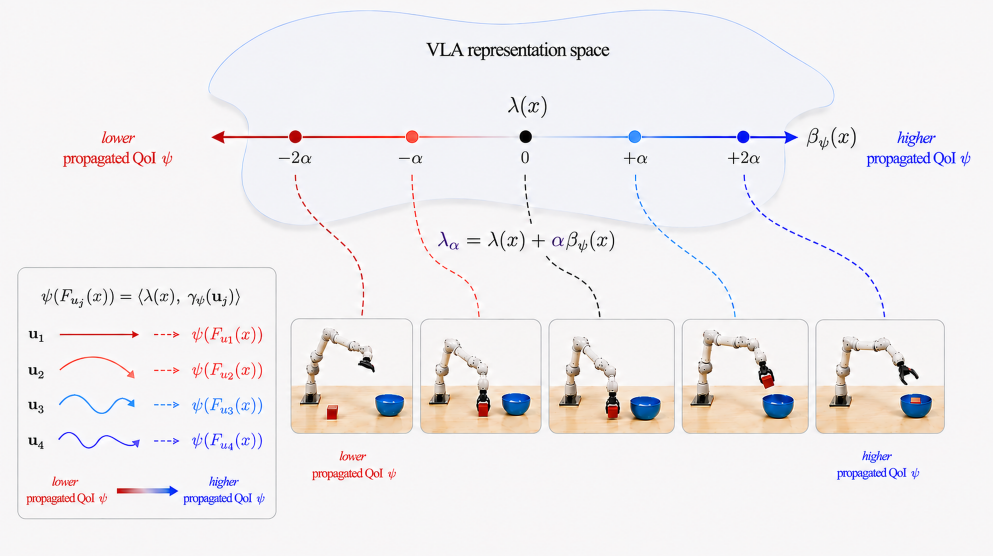

Figure 1: **Linear probing and steering in VLA representations.** Given a current state x with representation $\lambda(x)$, candidate action trajectories $\mathbf{u}_j$ define action-conditioned linear probes that predict the propagated quantity of interest (QoI), $\psi(F_{\mathbf{u}_j}(x)) = \langle \lambda(x), \gamma_\psi(\mathbf{u}_j) \rangle$. Steering the representation along $\lambda_\psi(x)$, via $\lambda(x) \mapsto \lambda(x) + \alpha\lambda_\psi(x)$, monotonically shifts the propagated QoI, allowing the policy to favor lower or higher physical outcomes.

But the objects of interest are not quite the same. In an LLM, commonly probed semantic directions, such as gender $\texttt{Rep}(\text{“Female”}) - \texttt{Rep}(\text{“Male”})$ or language $\texttt{Rep}(\text{“English”}) - \texttt{Rep}(\text{“French”})$, are relatively static. In a VLA, by contrast, many quantities of interest pertain to physical systems and, crucially, their evolutions: the pose of a manipulated object, or the force at a contact. A fixed representation direction may suffice to read out such a quantity at the current state, but decision-making often depends on how that quantity changes under action. For example, what matters is whether the object will reach a desired height or whether the contact force will not become excessive throughout task execution. Such future quantities depend not only on the current representation, but also on the action trajectory executed from that state and the dynamics through which those actions unfold. Although recent empirical studies on VLA models have shown promising results (Lu et al., 2025; Buurmeijer et al., 2026), a rigorous theoretical framework for characterizing this dynamical coupling remains largely undeveloped relative to the LRH literature for LLMs.

We therefore develop a mathematical abstraction of VLA models centered on the representation–action–dynamics interface. Our contributions are as follows.

1. **A control-theoretic formulation of the LRH for VLA models.** We connect VLA models to the underlying controlled dynamical system, leading to VLA-specific notions of *probing* and *steering*. Building on these notions, we formulate the LRH for VLA models and identify the key challenges arising from their architectural differences from LLMs.
2. **Shared linear representations for linear probing.** We show constructively that an ideal VLA representation can simultaneously encode the information of multiple physical quantities of interest, with each propagated value under a candidate action trajectory admitting a *linear probe* in representation space.
3. **Generalized linear policies for linear steering.** We introduce a path signature-based exponential family formulation of VLA policies, whose geometry provides a dually flat statistical manifold. Building on this geometry, we show that *linear probing* gives rise to *linear steering* in the representation space.
4. **Explicit numerical verification.** To provide a concrete picture, we construct an oracle representation that realizes our theoretical framework. We then numerically illustrate the resulting linear probing and steering mechanisms.

Basic background on VLA models. We consider a minimal abstraction of a VLA model. Let x denote the underlying physical state, which is not directly accessible to the model. Instead, the VLA receives an observation $o = h(x)$, which may include visual inputs, multiple camera views, or other sensor modalities. Given a language instruction ℓ, the VLA backbone maps o to an internal representation $\bar{\lambda}(o, \ell)$. Throughout the paper, we fix ℓ and suppress its dependence. Motivated by recent work on physically meaningful visual representations (Bardes et al., 2024; Assran et al., 2025), we conjecture that such information is accessible from the VLA representation. We therefore define $\lambda(x) := \bar{\lambda}(h(x), \ell)$. The representation $\lambda(x)$ is then passed to a stochastic action policy π, such as a diffusion-based action-chunk policy. For our analysis, we model the resulting action chunk as a continuous-time open-loop trajectory over a horizon H, $\mathbf{u} := \{\mathbf{u}(t)\}_{t\in[0,H]} \sim \pi(\cdot \mid \lambda(x))$. The resulting VLA pipeline is summarized as

$$\boxed{x \xrightarrow{h} o \xrightarrow{\bar{\lambda}(\cdot,\ell)} \lambda(x) \xrightarrow{\pi} \mathbf{u} \sim \pi(\cdot \mid \lambda(x)).}$$

We ask whether the LRH extends to this largely unexplored architectural setting.

2 A Control-Theoretic Formulation of VLA Models

2.1 VLA Models as Controlled Dynamical Systems

We now connect the VLA model to the underlying physical dynamics by interpreting the action trajectory $\mathbf{u}$ as a control input to a nonlinear control-affine system:

$$\dot{x}(t) = f_0(x(t)) + \sum_{i=1}^{m} f_i(x(t))u^i(t) \tag{1a}$$

$$o(t) = h(x(t)) \tag{1b}$$

Here, $x(t) \in \mathcal{X} \subseteq \mathbb{R}^n$ denotes the unobserved physical state, $\mathbf{u}(t) = (u^1(t), \ldots, u^m(t)) \in \mathcal{U} \subset \mathbb{R}^m$ denotes the control input, f_0 is the drift vector field, f_i is the control vector field associated with the i-th control input, and $h : \mathcal{X} \to \mathcal{O} \subseteq \mathbb{R}^o$ is the observation map. Let $\mathcal{U}_H$ denote the set of admissible open-loop control paths $\mathbf{u} : [0, H] \to \mathcal{U}$. For each $\mathbf{u} \in \mathcal{U}_H$, we denote by $F_\mathbf{u} : \mathcal{X} \to \mathcal{X}$ the corresponding state-transition map over the horizon H, defined by $x(H) = F_\mathbf{u}(x(0))$.

We next define a hidden scalar quantity $\psi : \mathcal{X} \to \mathbb{R}$ on the physical state space analogous to a high-level concept in the LLM setting.

Definition 2.1 (Quantity of Interest (QoI))**.** Let $\psi : \mathcal{X} \to \mathbb{R}$ be a real-valued function of the physical state x. For example, $\psi(x) = x_i$ may extract the i-th component of the state, such as a coordinate of a robot's position. More generally, ψ may represent any physically meaningful scalar quantity derived from x, such as a control Lyapunov function or a value function.[1]

2.2 Probing and Steering in VLA Models

We now adapt the notions of *probing* and *steering* to the VLA setting. Unlike LLM representations, VLA representations are coupled to actions that induce physical state transitions through interaction with the environment. We must therefore reformulate probing and steering to reflect this interaction.

For *probing*, we consider a physically meaningful QoI $\psi(x)$ and ask whether its future value under a candidate action trajectory $\mathbf{u}$ can be inferred from the VLA representation.

Probing Problem

Given a QoI $\psi : \mathcal{X} \to \mathbb{R}$ and a candidate action trajectory $\mathbf{u} \in \mathcal{U}_H$, can the propagated QoI

$$\psi(F_\mathbf{u}(x))$$

be predicted using only the VLA representation $\lambda(x)$ and $\mathbf{u}$, without explicit access to the underlying physical state or dynamics?

[1] This general notion of a QoI may also support the study of safety and planning in representation space.

A representation is informative for probing if, together with $\mathbf{u}$, it is sufficient to predict the action-conditioned propagated QoI $\psi(F_{\mathbf{u}}(x))$.

Existing VLA probing methods focus on static QoI, asking whether $\psi(x)$ can be decoded directly from the internal representation (Lu et al., 2025; Bhardwaj et al., 2026), with promising empirical results. However, static probing does not capture the action-conditioned evolution of the physical quantity. This distinction is particularly important for VLA models, whose latent representations both generate state-transitioning actions and provide a space for steering them (Häon et al., 2025).[2]

For *steering*, we ask whether an intervention in the representation space can alter the action trajectory generated by the policy so as to induce a desired change in the propagated physical state.

Steering Problem

Given a QoI $\psi : \mathcal{X} \to \mathbb{R}$, can we modify the VLA representation $\lambda(x)$ so that the resulting policy generates an action trajectory $\widetilde{\mathbf{u}} \sim \pi(\cdot \mid \texttt{Steer}(\lambda(x)))$ that drives the propagated QoI

$$\psi\big(F_{\widetilde{\mathbf{u}}}(x)\big)$$

toward a desired value?

In general, a steering map $\lambda(x) \mapsto \texttt{Steer}(\lambda(x))$ need not admit a simple algebraic form. Recent work on VLA steering has therefore explored more structured interventions on internal representations, including conceptor-based subspace operators and learned distribution-matching maps that go beyond a single additive steering direction (Miao et al., 2026; Khayatan et al., 2026).

Together, these two problems characterize the role of the representation λ in a VLA model: *probing* asks whether the representation contains sufficient information to predict the effect of a candidate action, whereas *steering* asks whether modifying the representation can systematically alter that effect through the policy.

3 Linear Probing and Steering in VLA Models

The linear representation hypothesis posits that both problems above admit a common resolution: linear structure in the VLA representation space. We extend the LRH for LLMs (Park et al., 2024) to VLA models by characterizing two complementary notions: *linear probing* and *linear steering*.

Definition 3.1 (Linear Probing in VLA)**.** Let $\psi : \mathcal{X} \to \mathbb{R}$ be a QoI. We say that ψ admits linear probing in the representation $\lambda : \mathcal{X} \to \mathbb{R}^d$ if there exists a map $\gamma_\psi : \mathcal{U}_H \to \mathbb{R}^d$ such that

$$\psi(F_{\mathbf{u}}(x)) = \langle \lambda(x), \gamma_\psi(\mathbf{u}) \rangle_{\mathbb{R}^d}, \qquad \forall\, x \in \mathcal{X},\ \mathbf{u} \in \mathcal{U}_H.$$

Note that this structure differs from the LLM probing setting. There, the probe is typically given by a single global direction $\gamma_\psi \in \mathbb{R}^d$, and the QoI is estimated from that direction alone; in our setting, the corresponding object depends on the action trajectory $\mathbf{u}$. Such dependence again reflects the interaction structure discussed above. The usual static probing setting is recovered by restricting to a no-transition case, in which action dependence is no longer required.

Challenges in linear probing. In the LLM setting, the relevant linear measurement direction γ_ψ is naturally tied to the rows of the output projection (or unembedding) matrix preceding the softmax. In contrast, VLA models do not provide an analogous, predefined measurement structure for propagated physical QoIs. A key challenge is therefore to characterize an action-conditioned map $\gamma : \mathcal{U}_H \to \mathbb{R}^d$ that associates action trajectory with the representation space.

Definition 3.2 (Linear Steering in VLA)**.** Let $\psi : \mathcal{X} \to \mathbb{R}$ be a QoI. We say that ψ admits linear steering at $x \in \mathcal{X}$ if there exists a nonzero vector $\lambda_\psi(x) \in \mathbb{R}^d$ such that

$$\alpha \longmapsto \mathbb{E}_{\mathbf{u} \sim \pi(\cdot \mid \lambda(x) + \alpha \lambda_\psi(x))} \left[\psi(F_{\mathbf{u}}(x))\right] \quad \text{is nondecreasing.}$$

[2] This is closely related to LeCun's view that an agentic system should be able to predict the consequences of contemplated actions (LeCun, 2022). A similar perspective is embodied in explicit world-model formulations, such as (Ha & Schmidhuber, 2018). Importantly, however, such predictive capability need not rely on an explicit world model and may also emerge implicitly within VLA models.

In our formulation, unlike standard LLM steering approaches that rely on a fixed global direction $\lambda_\psi \in \mathbb{R}^d$, the steering direction $\lambda_\psi(x)$ is generally state-conditioned. This is a natural extension in the VLA setting, since the propagated QoI $\psi(F_{\mathbf{u}}(x))$ depends on both the current state and the action trajectory through the system dynamics; consequently, the same change in action can have different effects on the propagated QoI at different states.

Challenges in linear steering. To control a physical QoI through a linear translation $\lambda \mapsto \lambda + \alpha\lambda_\psi$, we need to understand how such an intervention changes the downstream policy distribution and, consequently, the resulting physical behavior. To characterize this relationship, we adopt the information-geometric perspective of the *Softmax* (Park et al., 2026), viewing $\lambda \mapsto \pi(\cdot \mid \lambda)$ as a parameterization of a family of VLA policies. This motivates us to seek a generalized linear model (GLM) parameterization of the VLA policy under which the geometry induced by variations in λ admits an explicit characterization. However, owing to their structural differences from LLMs, modern generative VLA policies do not naturally admit such a convenient parameterization.

4 Existence of Representation and Measurement Maps

In this section, we take a closer look at *linear probing* in VLA models and address two fundamental theoretical questions as follows:

1. Can the VLA representation λ capture the physical information necessary to characterize multiple quantities of interest, $\Psi = \{\psi_j\}_{j=1}^p$?
2. For each QoI ψ_j, can we construct an action-conditioned measurement map γ_{ψ_j} such that the propagated QoI admits a linear probe from the VLA representation λ?

Before stating our main result, we introduce a key object used in its construction. Analogous to the role of unembedding vectors in the softmax-based measurement of LLMs, the action-conditioned measurement map γ can be chosen to factor through a *signature* of the action trajectory.

Definition 4.1 (Signature of an Action Trajectory)**.** Let $\mathcal{A} := \{0, 1, \dots, m\}$ denote the alphabet of action indices, with $u^0(t) \equiv 1$, and let $\mathcal{A}^\star$ denote the set of all finite words (or multi-indices) over $\mathcal{A}$, including the empty word $\emptyset$. For a word $I = (i_1, \dots, i_k) =: i_1 \cdots i_k \in \mathcal{A}^\star$, define

$$S(\mathbf{u})^I_{0,H} := \int_{0 \le t_1 \le \cdots \le t_k \le H} u^{i_1}(t_1) \cdots u^{i_k}(t_k)\, \mathrm{d}t_1 \cdots \mathrm{d}t_k,$$

with $S(\mathbf{u})^\emptyset_{0,H} = 1$. The collection $S(\mathbf{u}) := \big(S(\mathbf{u})^I_{0,H}\big)_{I \in \mathcal{A}^\star} \in \mathcal{S}$ is called the *signature* of $\mathbf{u}$.

The signature provides a principled representation of a continuous-time path, encoding its geometric and sequential structure through iterated integrals. It has been widely studied in mathematics, machine learning, and control (Lyons, 1998; Kidger et al., 2019; 2020; Bank et al., 2025). We refer the reader to (Chevyrev & Kormilitzin, 2026) for further details.

To address the challenges in linear probing described above, we interpret the signature $S : \mathcal{U}_H \to \mathcal{S}$ as an infinite-dimensional realization of the action-conditioned measurement map γ. For each QoI ψ_j, we consider a finite-dimensional measurement map that can be obtained through a linear map

$$L_{\psi_j} : \mathcal{S} \to \mathbb{R}^d, \qquad \gamma_{\psi_j}(\mathbf{u}) := (L_{\psi_j} \circ S)(\mathbf{u}).$$

The range of L_{ψ_j} therefore specifies a QoI-specific linear subspace of the representation space, while the signature parameterizes the measurement direction within that subspace as a function of the action trajectory. Our main result below establishes the existence of a shared VLA representation λ where these QoI-specific measurement structures can coexist.

Theorem 4.2 (Existence of Representation and Measurement Maps)**.** *Consider the control-affine system in Eq.* (1a)*, a finite collection of QoIs $\Psi = \{\psi_j\}_{j=1}^p$, and a compact set $K \subset \mathcal{X}$. Suppose the uniform Chen–Fliess conditions (Appendix A) hold on the horizon-H reachable set from K. Then for every $\varepsilon > 0$, there exist $d \in \mathbb{N}$, a shared representation map $\lambda : K \to \mathbb{R}^d$, and linear maps $L_{\psi_j} : \mathcal{S} \to \mathbb{R}^d$, $j = 1, \dots, p$, such that, with $\gamma_{\psi_j} := L_{\psi_j} \circ S : \mathcal{U}_H \to \mathbb{R}^d$, we have*

$$\max_{1 \le j \le p} \sup_{(x,\mathbf{u}) \in K \times \mathcal{U}_H} \big|\psi_j(F_{\mathbf{u}}(x)) - \big\langle \lambda(x), \gamma_{\psi_j}(\mathbf{u}) \big\rangle_{\mathbb{R}^d}\big| < \varepsilon.$$

The result is established constructively via the classical Chen–Fliess expansion (Chen, 1977; Fliess, 1981), as detailed in Appendix B.1. Importantly, we do not assume that the learned VLA representation $\lambda(x) = \bar{\lambda}(h(x), \ell)$ coincides exactly with the Lie-derivative-based construction. Rather, the latter establishes the existence of physically meaningful structure required by our framework. Theorem 4.2 provides a theoretical framework consistent with empirical linear probing results observed in existing VLA representations (Lu et al., 2025; Buurmeijer et al., 2026), while also providing an oracle against which arbitrary learned representations can be quantitatively assessed

Theorem 4.2 answers Questions 1 and 2 affirmatively, in an existential sense. A desirable VLA model can admit a shared representation λ that supports multiple QoIs, with each QoI equipped with an action-conditioned linear measurement of the form $\gamma_{\psi_j} = L_{\psi_j} \circ S$. Thus, the action dependence is captured by the signature, with L_{ψ_j} specifying the QoI-specific measurement structure.

5 Geometry of the Representation Space

Section 4 addressed the representation-side question of whether propagated QoIs can admit action-conditioned linear probes. Linear probing alone, however, does not determine how an intervention on the representation changes the downstream action distribution. Thus, we next introduce a geometry on the representation space for formulating *linear steering*. To this end, we view stochastic action-chunk policies, such as autoregressive and diffusion policies, as defining a family of conditional distributions over action trajectories $\mathbf{u} \in \mathcal{U}_H$

$$\mathcal{P} := \{\pi(\cdot \mid \lambda) : \mathcal{U}_H \to \mathbb{R}_{\geq 0}\},$$

and regard this family as parameterized by λ. To emphasize that the policy is parameterized by the representation λ, we henceforth write the policy as π_λ. Motivated by the geometric structure induced by the softmax-based distribution family in (Park et al., 2026), we would like to introduce a corresponding notion of geometry in representation space through $\mathcal{P}$.

For the policy-side analysis, we adopt the following identification. In general, the distribution family induced by a modern generative policy may possess a highly complex parametric structure. To obtain a tractable yet expressive representation of path-dependent statistics, we lift action trajectories $\mathbf{u} \mapsto S(\mathbf{u})$ into the signature space $\mathcal{S}$. This construction leverages the universality of linear functionals on signatures for approximating continuous path functionals (Lyons, 2014).

We then place this construction in a Hilbert-space setting by replacing the finite-dimensional representation space $\mathbb{R}^d$ with the Hilbert signature space $\mathcal{H}_S$, obtained by completing $\mathcal{S}$ under a suitable inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_S}$. This formulation naturally bridges the representation and policy modules. Specifically, we regard $\lambda(x) \in \Theta \subseteq \mathcal{H}_S$ simultaneously as the VLA representation and as the natural parameter of the signature exponential family. This identification allows us to draw on the well-developed body of work on linear representations in LLMs, including their geometry, and steering (Park et al., 2024; Marks & Tegmark, 2024; Park et al., 2026).

Definition 5.1 (Signature Exponential Family)**.** Given a base measure μ on the action-chunk space $\mathcal{U}_H$, we define the log-partition function $A(\lambda) := \log \int_{\mathcal{U}_H} \exp(\langle \lambda, S(\mathbf{u}) \rangle_{\mathcal{H}_S}) \mathrm{d}\mu(\mathbf{u})$, whenever the integral is finite. The *signature exponential family* generated by μ is

$$\mathcal{P}_\mu := \left\{ \frac{\mathrm{d}\pi_\lambda}{\mathrm{d}\mu}(\mathbf{u}) = \exp(\langle \lambda, S(\mathbf{u}) \rangle_{\mathcal{H}_S} - A(\lambda)) : \lambda \in \Theta \right\}$$

where $\Theta := \{\lambda \in \mathcal{H}_S : A(\lambda) < \infty\}$ is the natural parameter space.

The resulting exponential-family structure further admits a natural GLM perspective (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989), in which the conditional distribution of $\mathbf{u} \,|\, x$ belongs to an exponential family with natural parameter determined by x.

Definition 5.2 (Signature Generalized Linear Model)**.** Suppose that $\mathbf{u} \,|\, x$ follows a signature exponential family with base measure μ on $\mathcal{U}_H$. A *signature generalized linear model* specifies a state-conditioned natural parameter map $\lambda : \mathcal{X} \to \Theta$, which induces the conditional distribution

$$\frac{\mathrm{d}\pi_\lambda}{\mathrm{d}\mu}(\mathbf{u} \mid x) = \exp\big(\langle \lambda(x), S(\mathbf{u}) \rangle_{\mathcal{H}_S} - A(\lambda(x))\big).$$

We introduce the signature GLM as an architecture-agnostic statistical abstraction linking representation changes to downstream action distributions, rather than as a characterization of specific VLA architectures. By abstracting away the internal generative mechanism, it provides a mathematically interpretable geometry for analyzing representation–policy interactions. Specifically, this exponential family structure endows the statistical manifold with a *dually flat geometry* (Amari & Nagaoka, 2000), with dual coordinate $\varphi(\lambda) := \nabla A(\lambda) = \mathbb{E}_{\mathbf{u}\sim\pi_\lambda}[S(\mathbf{u})]$ and dual space $\Phi := \varphi(\Theta)$. For a fixed state $x \in \mathcal{X}$, define the state-conditioned steering direction $\lambda_\psi(x) := L_\psi^* \lambda(x) \in \mathcal{H}_S$ as in Definition 3.2, so that the expected propagated QoI is linear in the dual coordinate:

$$\mathbb{E}_{\mathbf{u}\sim\pi_\lambda}\big[\psi(F_\mathbf{u}(x))\big] = \langle \lambda(x), L_\psi\, \mathbb{E}_{\mathbf{u}\sim\pi_\lambda}[S(\mathbf{u})]\rangle_{\mathcal{H}_S} = \big\langle L_\psi^*\lambda(x), \varphi(\lambda)\big\rangle_{\mathcal{H}_S} = \langle \lambda_\psi(x), \varphi(\lambda)\rangle_{\mathcal{H}_S}\,.$$

A desired QoI level $c \in \mathbb{R}$ therefore corresponds to the dual affine hyperplane $\Phi_{\psi,x}(c) := \{\varphi \in \Phi : \langle \lambda_\psi(x), \varphi\rangle_{\mathcal{H}_S} = c\}$, and we formulate steering as the reverse-KL projection onto it,

$$\widehat{\lambda} \in \operatorname*{arg\,min}_{\varphi(\lambda)\in\Phi_{\psi,x}(c)} D_{\mathrm{KL}}\big(\pi_\lambda \,\|\, \pi_{\lambda(x)}\big).$$

By dual Bregman geometry the solution is $\widehat{\lambda} = \lambda(x) + \alpha\lambda_\psi(x)$ for some $\alpha \in \mathbb{R}$: projecting onto a constant expected-QoI hyperplane induces the linear path

$$\lambda_\alpha := \lambda(x) + \alpha\, \lambda_\psi(x) \tag{2}$$

in the natural parameter space. For readers seeking further details, we provide a complete derivation in Appendix C.1. We now show that this path satisfies the desired notion of linear steering.

Theorem 5.3 (Reverse-KL Linear Steering for VLA)**.** *Fix $x \in \mathcal{X}$, and let λ_α denote the steering path in Eq.* (2)*. Define $J_{\psi,x}(\alpha) := \mathbb{E}_{\mathbf{u}\sim\pi_{\lambda_\alpha}}\big[\psi(F_\mathbf{u}(x))\big]$. Suppose QoI ψ admits an exact linear probe at x of the form $\psi\big(F_\mathbf{u}(x)\big) = \big\langle \lambda(x),\, L_\psi S(\mathbf{u})\big\rangle_{\mathcal{H}_S}$, $\forall \mathbf{u} \in \mathcal{U}_H$, for a bounded linear operator $L_\psi : \mathcal{S} \to \mathcal{H}_S$. Then, for every α such that $\lambda_\alpha \in \Theta$, $J_{\psi,x}(\alpha)$ is nondecreasing in α.*

See Appendix B.2 for the proof. Although we assume an exact linear probe, Theorem 4.2 shows that this regime can be approximated arbitrarily well with a sufficiently expressive representation as d becomes large. We therefore interpret the exact-probe assumption as an idealized limit in which the propagated QoI can be represented without approximation error. In this limit, the exact linear-steering guarantee is recovered. Establishing quantitative steering guarantees under finite probing error is a separate robustness question that we leave for future work.

6 Numerical Verification

We begin with a controlled numerical verification of the probing and steering theory developed in Sections 4 and 5. To isolate the predicted steering mechanism from representation-learning error, we instantiate the Chen–Fliess construction of Theorem 4.2 as an *oracle representation* for the controlled system. This yields an exact action-conditioned linear probe together with the corresponding signature exponential family policy. Due to space limitations, several intermediate steps are omitted here. Please see Appendix C.2 for the full derivation.

Setup. Given the language instruction ℓ = "*Navigate to the goal.*", consider a minimal planar control-affine system with state $x = (p_x, p_y, \theta) \in \mathbb{R}^3$ and input $\mathbf{u} = (1, v, \omega) \in \mathcal{U}_H$:

$$\dot{p}_x = v, \qquad \dot{p}_y = \theta v, \qquad \dot{\theta} = \omega,$$

over horizon $H = 1$ from $x_0 = (0, 0, 0)$, with nominal goal $p_{\mathrm{goal}} = (1, 0)$. We denote by $\mathbf{u}$ an entire action chunk over the horizon $[0, H]$. Each chunk is piecewise constant over $K = 8$ equal-length intervals of duration $\Delta t = H/K$ and is parameterized by a forward speed s and a turning rate r.

$$v(t) = s, \qquad \omega(t) = \begin{cases} r, & t < H/2, \\ -r, & t \geq H/2, \end{cases}$$

so that each chunk turns and then counter-turns, producing lateral motion while approximately restoring the terminal heading. Sweeping s over 11 uniform points in $[0.75, 1.25]$ and r over 61 uniform points in $[-2.2, 2.2]$ yields 671 candidates.

For visual reference, we place a circular obstacle centered at $p_{\mathrm{obs}} = (0.5, 0)$ with radius $r_{\mathrm{obs}} = 0.08$. The obstacle is not included in C_{task}; it serves to visualize the physical effect of lateral steering.

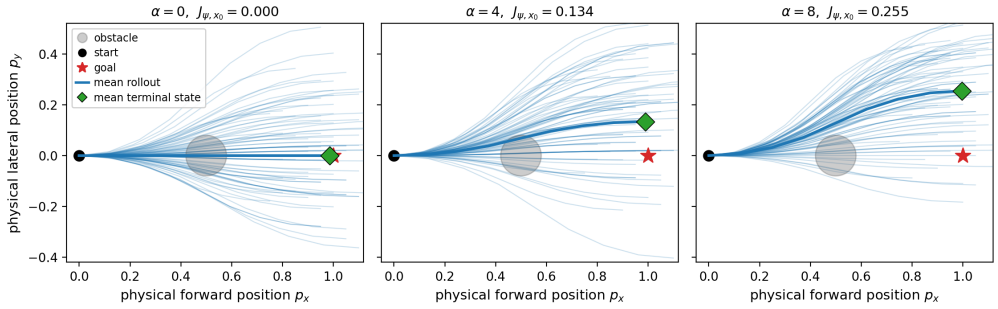

Figure 2: **Exact oracle steering in the physical workspace.** For QoI $\psi(x) = p_y$, increasing α along $\lambda_\alpha = \lambda(x_0) + \alpha\lambda_\psi(x_0)$ shifts the policy toward trajectories with increased propagated QoI p_y. Each panel shows 80 rollouts sampled i.i.d. from π_{λ_α}, the policy-averaged rollout (thick), and the policy-weighted mean terminal state (diamond).

Quantity of interest and exact oracle probe. We take $\psi(x) = p_y$, so $\psi(F_{\mathbf{u}}(x)) = p_y(H)$. Only three signature coordinates are required because they are the only ones with nonzero Chen–Fliess coefficients for $\psi(x) = p_y$: $\{\emptyset, 1, 21\}$. All other signature terms have zero coefficients. Thus, the Chen–Fliess expansion terminates exactly, hence the term *oracle representation*:

$$\psi(F_{\mathbf{u}}(x)) = \langle \lambda_\psi(x), S(\mathbf{u})\rangle_{\mathcal{H}_S} = p_y + \theta S(\mathbf{u})^1_{0,H} + S(\mathbf{u})^{21}_{0,H}, \qquad \lambda_\psi(x) = \begin{bmatrix} p_y, & \theta, & 1 \end{bmatrix}^\top.$$

To obtain a representation λ that is not aligned with the QoI, we represent it by an invertible linear change of coordinates of the form $\lambda(x) = M\lambda_\psi(x)$ and set $L_\psi = M^{-\top}$, $\gamma_\psi(\mathbf{u}) = L_\psi S(\mathbf{u})$, with

$$M = \begin{bmatrix} 0.10 & -0.04 & 0.18 \\ 0.14 & 0.03 & 0 \\ 0.02 & 0.16 & 0 \end{bmatrix}, \qquad \lambda(x_0) = M\lambda_\psi(x_0) = (0.18, 0, 0)^\top.$$

Since $M^\top M^{-\top} = I$, the bilinear relation $\psi(F_{\mathbf{u}}(x)) = \langle \lambda(x), \gamma_\psi(\mathbf{u})\rangle_{\mathcal{H}_S}$ holds exactly for every candidate chunk. The oracle VLA representation λ then plays two complementary roles: paired with the probe feature $\gamma_\psi(\mathbf{u})$, it decodes the propagated QoI; paired with the sufficient statistic $S(\mathbf{u})$, it serves as the natural parameter of the policy. These two pairings are connected through L_ψ.

Goal-directed VLA policy. We define a discrete base measure μ over candidate actions with action preference encoded by the goal-reaching cost

$$C_{\text{task}}(\mathbf{u}; x_0) = 8\big(p_x(H) - 1\big)^2 + 2p_y(H)^2 + 4\theta(H)^2 + 0.05\int_0^H \big(v^2 + \omega^2\big)\mathrm{d}t.$$

We set $\mu(\mathbf{u}) \propto \exp(-C_{\text{task}}(\mathbf{u}; x_0)/\tau)$ with $\tau = 0.20$, and instantiate the signature exponential family policy as $\pi_\lambda(\mathbf{u} \mid x_0) \propto \mu(\mathbf{u}) \exp(\langle \lambda, S(\mathbf{u})\rangle_{\mathcal{H}_S})$.

Since $\lambda_\psi(x_0) = (0, 0, 1)^\top$, the parameter $\lambda(x_0)$ contributes only the constant $\langle \lambda(x_0), S(\mathbf{u})\rangle_{\mathcal{H}_S} \equiv 0.18$ to the policy log-density. Hence, after normalization, the unsteered policy $\pi_{\lambda(x_0)}$ coincides with the base measure μ.

Linear steering. Following Theorem 5.3, we define $\lambda_\psi(x) = L_\psi^\top \lambda(x)$ and traverse the linear path $\lambda_\alpha = \lambda(x_0) + \alpha\lambda_\psi(x_0)$. Here $L_\psi^\top \lambda(x) = M^{-1}M\lambda_\psi(x) = \lambda_\psi(x)$, so $\lambda_\psi(x_0) = (0, 0, 1)^\top$ and

$$\langle \lambda_\psi(x_0), S(\mathbf{u})\rangle_{\mathcal{H}_S} = S^{21}_{0,H}(\mathbf{u}) = \psi(F_{\mathbf{u}}(x_0)),$$

the last equality because $p_y = \theta = 0$ at x_0. Since the log-density is affine in λ, the shift by $\alpha\lambda_\psi(x_0)$ factors out,

$$\pi_{\lambda_\alpha}(\mathbf{u} \mid x_0) \propto \mu(\mathbf{u})\, e^{\langle \lambda(x_0), S(\mathbf{u})\rangle_{\mathcal{H}_S}}\, e^{\alpha\langle \lambda_\psi(x_0), S(\mathbf{u})\rangle_{\mathcal{H}_S}} \propto \pi_{\lambda(x_0)}(\mathbf{u} \mid x_0)\, \exp\big(\alpha\, \psi(F_{\mathbf{u}}(x_0))\big),$$

so the steered policy is the nominal one tilted by the propagated QoI itself, and we report $J_{\psi,x_0}(\alpha)$.

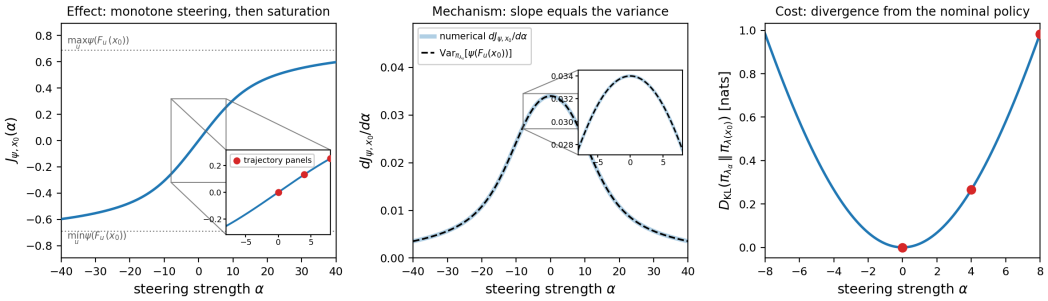

Figure 3: **Effect, mechanism, and cost of linear steering.** *Left:* $J_{\psi,x_0}(\alpha)$; dotted lines mark the extreme values of $\psi(F_{\mathbf{u}}(x_0))$ over the candidate set. *Middle:* numerical $\mathrm{d}J_{\psi,x_0}/\mathrm{d}\alpha$ against $\mathrm{Var}_{\pi_{\lambda_\alpha}}[\psi(F_{\mathbf{u}}(x_0))]$. *Right:* $D_{\mathrm{KL}}(\pi_{\lambda_\alpha}\|\pi_{\lambda(x_0)})$. Insets and the right panel show $\alpha \in [-8, 8]$; red markers are the $\alpha \in \{0, 4, 8\}$ panels of Fig. 2.

Results. Steering the representation λ_α moves the physical quantity it decodes. As α increases over $\{0, 4, 8\}$, the sampled rollouts shift laterally and the expected propagated QoI rises monotonically from 0.000 to $+0.134$ to $+0.255$ (Fig. 2), with the probe exact to 3×10^{-16} on all 671 candidates. Sweeping α continuously (Fig. 3) shows that this shift is predicted rather than merely observed: the numerical slope matches $\mathrm{Var}_{\pi_{\lambda_\alpha}}[\psi(F_{\mathbf{u}}(x_0))]$ to 10^{-11}, so the response to steering is determined by a quantity measurable from the nominal policy alone. It is also cheap, costing 0, 0.27, and 0.98 nats of KL at the three levels above. The steerable range is nonetheless bounded. Since ψ is bounded on the finite candidate set, large α concentrates π_{λ_α} on the extremal chunks, collapsing the same variance that sets the slope by an order of magnitude; J_{ψ,x_0} is therefore monotone but never affine, approaching 0.688 rather than growing.

7 Conclusion

We developed a formal mathematical framework for studying vision-language-action (VLA) models through the lens of the linear representation hypothesis. Unlike the standard large language model setting, physically meaningful quantities in a VLA must be understood through their evolution under candidate actions and the underlying controlled dynamics. Motivated by this distinction, we introduced action-conditioned linear probing of propagated quantities of interest (QoIs) and showed that signature-based representations provide a constructive route to shared linear representations for multiple physical QoIs. On the policy side, we introduced a signature generalized linear model and showed that an exact linear probe induces a linear steering direction along which the expected propagated QoI changes monotonically. These predictions were verified in a controlled oracle setting.

Scope and future directions. Our analysis focuses on an idealized regime characterized by two structural assumptions. *First*, the propagated QoI admits an exact action-conditioned linear probe. While Theorem 4.2 shows that the probing error can be made arbitrarily small with a sufficiently expressive representation, learned VLA representations need not satisfy this structure exactly. Nevertheless, existing empirical studies report encouraging probing performance on pretrained VLA models, and these results can be viewed as special cases of our broader action-conditioned formulation; we discuss this connection in detail in Appendix D. Developing refined probes that exploit this structure is left for future work. *Second*, we use the signature exponential family (SEF) as a tractable statistical geometry relating VLA representations to downstream stochastic policies. Modern VLA policies may rely on generative mechanisms such as diffusion or flow matching. Understanding when their induced action distributions are compatible with the SEF geometry, and how such compatibility can be assessed, provides a natural route toward extending the present framework.

Within this scope, our numerical study in Section 6 provides a controlled realization of the proposed probing and steering mechanisms, separating their theoretical structure from the additional complexities of learned representations and modern generative architectures. The framework thereby provides a concrete basis for investigating how such geometry emerges in pretrained VLA models.

POLICY ON LLM USE FOR RESEARCH AND WRITING PAPERS

We used generative AI tools to assist with language polishing and to discuss and refine aspects of the theoretical formulation. All mathematical statements and final formulations were independently verified by the authors.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society and Oxford University Press, 2000. doi: 10.1090/mmono/191.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Peter Bank, Christian Bayer, Paul P. Hager, Sebastian Riedel, and Tobias Nauen. Stochastic control with signatures. *SIAM Journal on Control and Optimization*, 63(5):3189–3218, 2025. doi: 10.1137/24M1667671.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Atiksh Bhardwaj, Edward Weiyi Duan, Prithwish Dan, Wei-Chiu Ma, and Preston Culbertson. Decoding task progress from VLA representations. *arXiv preprint arXiv:2608.13474*, 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems*, Los Angeles, CA, USA, 2025. doi: 10.15607/RSS.2025.XXI.010.

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

Hugo Buurmeijer, Carmen Amo Alonso, Aiden Swann, and Marco Pavone. Observing and controlling features in vision-language-action models, 2026. URL https://arxiv.org/abs/2603.05487.

Kuo-Tsai Chen. Iterated path integrals. *Bulletin of the American Mathematical Society*, 83(5): 831–879, 1977. doi: 10.1090/S0002-9904-1977-14320-6.

Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. In Christian Bayer, Gonçalo dos Reis, Blanka Horvath, and Harald Oberhauser (eds.), *Signature Methods in Finance*, pp. 3–64. Springer, Cham, 2026. doi: 10.1007/978-3-031-97239-3_1.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Michel Fliess. Fonctionnelles causales non linéaires et indéterminées non commutatives. *Bulletin de la Société Mathématique de France*, 109:3–40, 1981. doi: 10.24033/bsmf.1931.

Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis? In *Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5358–5376. PMLR, 2026.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems*, Delft, Netherlands, 2024. doi: 10.15607/RSS.2024.XX.090.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Bear Häon, Kaylene Caswell Stocking, Ian Chuang, and Claire Tomlin. Mechanistic interpretability for steering vision-language-action models. In *Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2743–2762. PMLR, 2025.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21879–21911. PMLR, 2024.

Pegah Khayatan, Sara Meziane, Jayneel Parekh, and Matthieu Cord. Dimas: Distribution matching for steering vision-language-action models, 2026. URL `https://arxiv.org/abs/2607.14280`.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*, Delft, Netherlands, 2024. doi: 10.15607/RSS.2024.XX.120.

Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signature transforms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2024.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023b.

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. In *IEEE International Conference on AI and Data Analytics*, 2025. doi: 10.1109/ICAD65464.2025.11114062.

Terry Lyons. Rough paths, signatures and the modelling of functions on streams. In Sun Young Jang, Young Rock Kim, Dae-Woong Lee, and Ikkwon Yie (eds.), *Proceedings of the International Congress of Mathematicians Seoul 2014*. Kyung Moon SA, 2014.

Terry J. Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998. doi: 10.4171/RMI/240.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 2 edition, 1989. ISBN 978-0-412-31760-6.

Miranda Muqing Miao, Subin Kim, Brandon Yang, and Lyle Ungar. Contrastive conceptor activation steering (COAST): Unlocking vision-language-action models through hidden states. *arXiv preprint arXiv:2605.17144*, 2026.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, 2013. Association for Computational Linguistics.

John A. Nelder and Robert W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972. doi: 10.2307/2344614.

Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation*, pp. 6892–6903. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, 2025.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. In *International Conference on Machine Learning*, 2026.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023. Version 1.

Jiachen Zhang, Junnan Nie, Junyi Lao, Wei Cheng, Chenghao Liu, Jiaxin Jiang, and Songfang Huang. What frozen VLAs already know about success: A probing study of value-like structure in foundation robot policies. *arXiv preprint arXiv:2605.28527*, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Regularity Conditions for the Chen–Fliess Expansion

We collect the notation and sufficient conditions used for the Chen–Fliess expansion in Theorem 4.2. Since the approximation result is local in the state variable, we formulate the conditions uniformly over an arbitrary compact set $K \subset \mathcal{X}$ and the admissible action class $\mathcal{U}_H$. The regularity and convergence constants below may depend on K, but are uniform over $K \times \mathcal{U}_H$.

Controlled system. Consider the control-affine system

$$\dot{x}(t) = f_0(x(t)) + \sum_{i=1}^{m} f_i(x(t))u^i(t), \tag{3}$$

and define $u^0(t) \equiv 1$. Then

$$\dot{x}(t) = \sum_{i=0}^{m} f_i(x(t))u^i(t).$$

Let $\mathcal{A} = \{0, 1, \ldots, m\}$ denote the input alphabet, and let $\mathcal{A}^\star$ denote the set of all finite words over $\mathcal{A}$, including the empty word $\emptyset$.

For a word $I = (i_1, \ldots, i_k) \in \mathcal{A}^\star$, we use the same chronological convention as in Definition 4.1. Namely,

$$S(\mathbf{u})^I_{0,H} := \int_{0 \le t_1 \le \cdots \le t_k \le H} \prod_{r=1}^{k} u^{i_r}(t_r)\, \mathrm{d}t_1 \cdots \mathrm{d}t_k, \qquad S(\mathbf{u})^\emptyset_{0,H} = 1. \tag{4}$$

With this convention, define the corresponding iterated Lie derivative by

$$\mathcal{L}_I \psi = \mathcal{L}_{f_{i_1}} \cdots \mathcal{L}_{f_{i_k}} \psi, \qquad \mathcal{L}_\emptyset \psi = \psi, \tag{5}$$

where $\mathcal{L}_{f_i}\phi(x) = D\phi(x) f_i(x)$. The ordering in Eqs. (4) and (5) is chosen so that the Chen–Fliess expansion takes the form

$$\psi(F_{\mathbf{u}}(x)) = \sum_{I \in \mathcal{A}^\star} \mathcal{L}_I \psi(x) S(\mathbf{u})^I_{0,H}.$$

A.1 Sufficient conditions for locally uniform convergence

Fix an arbitrary compact set $K \subset \mathcal{X}$. We impose the following sufficient conditions.

Assumption A.1 (Common reachable domain)**.** There exist an open set $D \subset \mathbb{R}^n$ and a compact set $K_H \subset D$ such that, for every $x \in K$ and every $\mathbf{u} \in \mathcal{U}_H$, the controlled system admits a unique solution on $[0, H]$, and

$$x_{\mathbf{u}}(t; x) \in K_H, \qquad \forall t \in [0, H].$$

Equivalently, the horizon-H reachable tube from K,

$$\mathcal{R}_H(K) := \{x_{\mathbf{u}}(t; x) : x \in K,\, \mathbf{u} \in \mathcal{U}_H,\, t \in [0, H]\},$$

satisfies $\mathcal{R}_H(K) \subset K_H \subset D$.

Assumption A.2 (Analytic regularity)**.** The vector fields $f_0, \ldots, f_m : D \to \mathbb{R}^n$ and each QoI $\psi_j : D \to \mathbb{R}$, $j = 1, \ldots, p$, are real analytic on D.

Assumption A.3 (Uniformly bounded controls)**.** Every admissible control satisfies

$$u^i \in L^\infty([0, H]), \qquad i = 1, \ldots, m,$$

and the admissible control class is uniformly bounded:

$$\bar{u} := \sup_{\mathbf{u} \in \mathcal{U}_H} \max\left\{1, \|u^1\|_{L^\infty}, \ldots, \|u^m\|_{L^\infty}\right\} < \infty. \tag{6}$$

Assumption A.4 (Uniform Lie-derivative growth)**.** There exist constants $C_K, M_K > 0$, possibly depending on K but independent of j, x, and I, such that

$$|\mathcal{L}_I \psi_j(x)| \le C_K M_K^{|I|} |I|!, \qquad \forall j \in \{1, \ldots, p\}, \quad \forall x \in K_H, \quad \forall I \in \mathcal{A}^\star. \tag{7}$$

Assumption A.5 (Local convergence radius)**.** The horizon and admissible control magnitude satisfy

$$q_K := (m+1)M_K\bar{u}H < 1. \tag{8}$$

Lemma A.6 (Uniform bound on signature coordinates)**.** *Under Assumption A.3, for every* $\mathbf{u} \in \mathcal{U}_H$ *and every* $I \in \mathcal{A}^\star$*,*

$$|S(\mathbf{u})^I_{0,H}| \le \frac{(\bar{u}H)^{|I|}}{|I|!}. \tag{9}$$

Proof. Let $I = (i_1, \dots, i_k)$, so that $|I| = k$. By Eq. (6),

$$|u^{i_r}(t)| \le \bar{u}, \qquad r = 1, \dots, k,$$

where the same bound holds for $i_r = 0$ because $u^0 \equiv 1$ and $\bar{u} \ge 1$. Hence

$$\begin{aligned} |S(\mathbf{u})^I_{0,H}| &\le \int_{0 \le t_1 \le \cdots \le t_k \le H} \prod_{r=1}^{k} |u^{i_r}(t_r)| \, \mathrm{d}t_1 \cdots \mathrm{d}t_k \\ &\le \bar{u}^k \operatorname{Vol}\{0 \le t_1 \le \cdots \le t_k \le H\}. \end{aligned}$$

Since the ordered k-simplex has volume $H^k/k!$,

$$|S(\mathbf{u})^I_{0,H}| \le \bar{u}^k \frac{H^k}{k!} = \frac{(\bar{u}H)^k}{k!}.$$

The empty-word case follows from $S(\mathbf{u})^\emptyset_{0,H} = 1$. □

Proposition A.7 (Locally uniform Chen–Fliess expansion)**.** *Let* $K \subset \mathcal{X}$ *be compact. Under Assumptions A.1 to A.5, for every* $j = 1, \dots, p$*,*

$$\psi_j(F_{\mathbf{u}}(x)) = \sum_{I \in \mathcal{A}^\star} \mathcal{L}_I \psi_j(x) S(\mathbf{u})^I_{0,H}, \tag{10}$$

and the series converges absolutely and uniformly over

$$j \in \{1, \dots, p\}, \qquad x \in K, \qquad \mathbf{u} \in \mathcal{U}_H.$$

Moreover, if

$$R_{j,N}(x, \mathbf{u}) := \sum_{|I| > N} \mathcal{L}_I \psi_j(x) S(\mathbf{u})^I_{0,H},$$

then

$$\max_{1 \le j \le p} \sup_{(x,\mathbf{u}) \in K \times \mathcal{U}_H} |R_{j,N}(x, \mathbf{u})| \le \frac{C_K q_K^{N+1}}{1 - q_K}, \qquad q_K = (m+1)M_K\bar{u}H < 1. \tag{11}$$

Proof. Fix $j \in \{1, \dots, p\}$, $x \in K$, and $\mathbf{u} \in \mathcal{U}_H$. For any sufficiently smooth scalar function $g : D \to \mathbb{R}$, the fundamental theorem of calculus along the controlled trajectory gives

$$\begin{aligned} g(x_{\mathbf{u}}(t; x)) - g(x) &= \int_0^t Dg(x_{\mathbf{u}}(s; x)) \dot{x}_{\mathbf{u}}(s; x) \, \mathrm{d}s \\ &= \sum_{i=0}^{m} \int_0^t \mathcal{L}_{f_i} g\big(x_{\mathbf{u}}(s; x)\big) u^i(s) \, \mathrm{d}s. \end{aligned}$$

Repeated application of this identity yields, for every $N \ge 0$,

$$\psi_j(F_{\mathbf{u}}(x)) = \sum_{|I| \le N} \mathcal{L}_I \psi_j(x) S(\mathbf{u})^I_{0,H} + \widetilde{R}_{j,N}(x, \mathbf{u}), \tag{12}$$

where $\widetilde{R}_{j,N}$ is a sum of iterated integrals of order $N+1$. Each such term contains an iterated Lie derivative of order $N+1$, evaluated at an intermediate point of the controlled trajectory.

By Assumption A.1, every such intermediate point belongs to K_H. Hence Assumption A.4 gives

$$\big|\mathcal{L}_I \psi_j\big(x_{\mathbf{u}}(t; x)\big)\big| \le C_K M_K^{N+1} (N+1)!$$

for every word w of length $N+1$. Using the uniform control bound and the volume of the ordered simplex, each remainder term is bounded by

$$C_K M_K^{N+1}(N+1)!\frac{(\bar{u}H)^{N+1}}{(N+1)!} = C_K(M_K\bar{u}H)^{N+1}.$$

Since there are $(m+1)^{N+1}$ words of length $N+1$,

$$|\widetilde{R}_{j,N}(x,\mathbf{u})| \le C_K q_K^{N+1}.$$

Since $q_K < 1$, this converges to zero uniformly over

$$j \in \{1,\dots,p\}, \qquad x \in K, \qquad \mathbf{u} \in \mathcal{U}_H.$$

Letting $N \to \infty$ in Eq. (12) therefore yields Eq. (10) uniformly on $K \times \mathcal{U}_H$.

For absolute convergence, let $|I| = k$. By Assumption A.4 and Lemma A.6,

$$\begin{aligned}\left|\mathcal{L}_I\psi_j(x)S(\mathbf{u})_{0,H}^I\right| &\le C_K M_K^k k!\frac{(\bar{u}H)^k}{k!}\\ &= C_K(M_K\bar{u}H)^k.\end{aligned}$$

There are $(m+1)^k$ words of length k. Therefore

$$\sum_{|I|=k}\left|\mathcal{L}_I\psi_j(x)S(\mathbf{u})_{0,H}^I\right| \le C_K q_K^k.$$

Since $q_K < 1$,

$$\sum_{k=0}^{\infty} C_K q_K^k = \frac{C_K}{1-q_K} < \infty.$$

The bound is independent of j, $x \in K$, and $\mathbf{u} \in \mathcal{U}_H$. Hence the Weierstrass M-test gives absolute and uniform convergence.

Finally,

$$\begin{aligned}|R_{j,N}(x,\mathbf{u})| &\le \sum_{k=N+1}^{\infty}\sum_{|I|=k}\left|\mathcal{L}_I\psi_j(x)S(\mathbf{u})_{0,H}^I\right|\\ &\le C_K\sum_{k=N+1}^{\infty} q_K^k\\ &= \frac{C_K q_K^{N+1}}{1-q_K}.\end{aligned}$$

Taking the maximum over j and the supremum over $K \times \mathcal{U}_H$ proves Eq. (11). □

B PROOFS

B.1 PROOF OF THEOREM 4.2

Proof. Fix an arbitrary compact set $K \subset \mathcal{X}$ and $\varepsilon > 0$. Let $\mathcal{A} = \{0, 1, \dots, m\}$, where $u^0(t) \equiv 1$, and let $\mathcal{A}^\star$ denote the set of all finite words over $\mathcal{A}$, including the empty word $\emptyset$.

By the locally uniform Chen–Fliess expansion established in Appendix A, for every $j \in \{1, \dots, p\}$,

$$\psi_j\big(F_{\mathbf{u}}(x)\big) = \sum_{I \in \mathcal{A}^\star} \mathcal{L}_I \psi_j(x) S(\mathbf{u})^I_{0,H}, \tag{13}$$

where the series converges absolutely and uniformly over

$$(x, \mathbf{u}) \in K \times \mathcal{U}_H.$$

More precisely, under the conditions of Appendix A, there are exactly $(m+1)^k$ words of length k,

$$\begin{aligned}\sum_{\substack{I \in \mathcal{A}^\star \\ |I|=k}} \big|\mathcal{L}_I \psi_j(x) S(\mathbf{u})^I_{0,H}\big| &\le (m+1)^k C_K M_K^k k! \frac{(\bar{u}H)^k}{k!} \\ &= C_K q_K^k.\end{aligned}$$

Consequently, if

$$R_{j,N}(x, \mathbf{u}) := \sum_{\substack{I \in \mathcal{A}^\star \\ |I|>N}} \mathcal{L}_I \psi_j(x) S(\mathbf{u})^I_{0,H}$$

denotes the truncation remainder, then

$$\max_{1 \le j \le p} \sup_{(x,\mathbf{u}) \in K \times \mathcal{U}_H} |R_{j,N}(x, \mathbf{u})| \le \frac{C_K q_K^{N+1}}{1 - q_K}. \tag{14}$$

Since $q_K < 1$, the right-hand side of Eq. (14) converges to zero as $N \to \infty$. Therefore, there exists $N \in \mathbb{N}$ such that

$$\frac{C_K q_K^{N+1}}{1 - q_K} < \varepsilon. \tag{15}$$

Define the finite word set

$$\mathcal{W}_N := \mathcal{A}^\star_{\le N} = \{I \in \mathcal{A}^\star : |I| \le N\},$$

and let

$$d_N := |\mathcal{W}_N| = \sum_{k=0}^{N} (m+1)^k.$$

Fix an ordering

$$\mathcal{W}_N = \{I_1, \dots, I_{d_N}\}.$$

For each QoI ψ_j, define the truncated Chen–Fliess coefficient vector

$$c_{j,N}(x) := \begin{bmatrix} \mathcal{L}_{I_1} \psi_j(x) \\ \vdots \\ \mathcal{L}_{I_{d_N}} \psi_j(x) \end{bmatrix} \in \mathbb{R}^{d_N}.$$

Construct the shared representation

$$\lambda_N(x) := \begin{bmatrix} c_{1,N}(x) \\ \vdots \\ c_{p,N}(x) \end{bmatrix} \in \mathbb{R}^{p d_N}. \tag{16}$$

Define the coordinate projection

$$P_N : \mathcal{S} \to \mathbb{R}^{d_N}$$

by

$$P_N s := \begin{bmatrix} s^{I_1} \\ \vdots \\ s^{I_{d_N}} \end{bmatrix}.$$

In particular,

$$P_N S(\mathbf{u}) = \begin{bmatrix} S(\mathbf{u})^{I_1}_{0,H} \\ \vdots \\ S(\mathbf{u})^{I_{d_N}}_{0,H} \end{bmatrix}.$$

For each $j \in \{1, \dots, p\}$, let

$$E_j : \mathbb{R}^{d_N} \to \mathbb{R}^{pd_N}$$

denote the canonical embedding into the j-th block:

$$E_j z = \begin{bmatrix} 0 \\ \vdots \\ z \\ \vdots \\ 0 \end{bmatrix}.$$

Define

$$L_{\psi_j,N} := E_j \circ P_N : \mathcal{S} \to \mathbb{R}^{pd_N}. \tag{17}$$

Since both P_N and E_j are linear, $L_{\psi_j,N}$ is linear.

The corresponding action-conditioned measurement map is

$$\gamma_{\psi_j,N}(\mathbf{u}) := L_{\psi_j,N} S(\mathbf{u}) = E_j P_N S(\mathbf{u}).$$

Using the standard Euclidean inner products on the corresponding spaces,

$$\begin{aligned} \left\langle \lambda_N(x), \gamma_{\psi_j,N}(\mathbf{u}) \right\rangle_{\mathbb{R}^{pd_N}} &= \langle c_{j,N}(x), P_N S(\mathbf{u}) \rangle_{\mathbb{R}^{d_N}} \\ &= \sum_{r=1}^{d_N} \mathcal{L}_{I_r} \psi_j(x)\, S(\mathbf{u})^{I_r}_{0,H} \\ &= \sum_{\substack{I \in \mathcal{A}^\star \\ |I| \le N}} \mathcal{L}_I \psi_j(x)\, S(\mathbf{u})^{I}_{0,H}. \end{aligned}$$

Combining this identity with Eq. (13) gives

$$\psi_j(F_{\mathbf{u}}(x)) - \left\langle \lambda_N(x), \gamma_{\psi_j,N}(\mathbf{u}) \right\rangle_{\mathbb{R}^{pd_N}} = R_{j,N}(x, \mathbf{u}).$$

Hence, by Eq. (15),

$$\begin{aligned} \max_{1 \le j \le p} \sup_{(x,\mathbf{u}) \in K \times \mathcal{U}_H} \left| \psi_j(F_{\mathbf{u}}(x)) - \left\langle \lambda_N(x), \gamma_{\psi_j,N}(\mathbf{u}) \right\rangle_{\mathbb{R}^{pd_N}} \right| &= \max_{1 \le j \le p} \sup_{(x,\mathbf{u}) \in K \times \mathcal{U}_H} |R_{j,N}(x, \mathbf{u})| \\ &\le \frac{C_K q_K^{N+1}}{1 - q_K} \\ &< \varepsilon. \end{aligned}$$

Finally, choose

$$d := pd_N, \qquad \lambda := \lambda_N, \qquad L_{\psi_j} := L_{\psi_j,N}.$$

Then

$$\gamma_{\psi_j} = L_{\psi_j} \circ S : \mathcal{U}_H \to \mathbb{R}^d$$

is linear in the signature coordinates and satisfies

$$\max_{1\le j\le p}\sup_{(x,\mathbf{u})\in K\times\mathcal{U}_H}\left|\psi_j(F_{\mathbf{u}}(x))-\left\langle\lambda(x),\gamma_{\psi_j}(\mathbf{u})\right\rangle_{\mathbb{R}^d}\right|<\varepsilon.$$

Since $K\subset\mathcal{X}$ was arbitrary, the approximation holds locally uniformly in the state variable. □

B.2 Proof of Theorem 5.3

Proof. Fix $x\in\mathcal{X}$, and write

$$\lambda_0=\lambda(x),\qquad \lambda_\psi=\lambda_\psi(x)=L_\psi^*\lambda_0.$$

We consider values of α for which

$$\lambda_\alpha=\lambda_0+\alpha\lambda_\psi\in\Theta,$$

so that $\pi_{\lambda_\alpha}(\cdot\mid x)$ is well defined.

By definition of the adjoint L_ψ^*,

$$\langle\lambda_0,L_\psi S(\mathbf{u})\rangle_{\mathcal{H}_S}=\left\langle L_\psi^*\lambda_0,S(\mathbf{u})\right\rangle_{\mathcal{H}_S}=\langle\lambda_\psi,S(\mathbf{u})\rangle_{\mathcal{H}_S}.$$

Since ψ admits an exact linear probe at x,

$$\psi\big(F_{\mathbf{u}}(x)\big)=\langle\lambda_0,L_\psi S(\mathbf{u})\rangle_{\mathcal{H}_S},\qquad\forall\,\mathbf{u}\in\mathcal{U}_H,$$

and therefore

$$\psi\big(F_{\mathbf{u}}(x)\big)=\langle\lambda_\psi,S(\mathbf{u})\rangle_{\mathcal{H}_S}. \tag{18}$$

Thus the propagated QoI coincides exactly with the sufficient-statistic direction induced by λ_ψ.

From Eq. (18),

$$\begin{aligned}J_{\psi,x}(\alpha)&=\mathbb{E}_{\mathbf{u}\sim\pi_{\lambda_\alpha}(\cdot|x)}\left[\langle\lambda_\psi,S(\mathbf{u})\rangle_{\mathcal{H}_S}\right]\\&=\left\langle\lambda_\psi,\mathbb{E}_{\mathbf{u}\sim\pi_{\lambda_\alpha}(\cdot|x)}\left[S(\mathbf{u})\right]\right\rangle_{\mathcal{H}_S}.\end{aligned}$$

For the signature exponential family, the dual coordinate is

$$\varphi(\lambda)=\nabla A(\lambda)=\mathbb{E}_{\mathbf{u}\sim\pi_\lambda(\cdot|x)}\left[S(\mathbf{u})\right].$$

Hence

$$J_{\psi,x}(\alpha)=\langle\lambda_\psi,\varphi(\lambda_\alpha)\rangle_{\mathcal{H}_S}. \tag{19}$$

Since the steering path is affine, i.e., $\lambda_\alpha=\lambda_0+\alpha\lambda_\psi$, its derivative with respect to α is

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\lambda_\alpha=\lambda_\psi.$$

Differentiating Eq. (19) along this path and applying the chain rule gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}J_{\psi,x}(\alpha)&=\left\langle\lambda_\psi,D\varphi(\lambda_\alpha)\left[\frac{\mathrm{d}\lambda_\alpha}{\mathrm{d}\alpha}\right]\right\rangle_{\mathcal{H}_S}\\&=\langle\lambda_\psi,D\varphi(\lambda_\alpha)\left[\lambda_\psi\right]\rangle_{\mathcal{H}_S}.\end{aligned}$$

Because $\varphi(\lambda)=\nabla A(\lambda)$, we have $D\varphi(\lambda)=\nabla^2A(\lambda)$, and therefore

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}J_{\psi,x}(\alpha)=\left\langle\lambda_\psi,\nabla^2A(\lambda_\alpha)[\lambda_\psi]\right\rangle_{\mathcal{H}_S}. \tag{20}$$

It remains to identify the quadratic form in Eq. (20). For the exponential family, the Hessian of the log-partition function is the covariance operator of the sufficient statistic. More precisely, for any $\eta,\zeta\in\mathcal{H}_S$,

$$\left\langle\eta,\nabla^2A(\lambda)[\zeta]\right\rangle_{\mathcal{H}_S}=\mathrm{Cov}_{\mathbf{u}\sim\pi_\lambda(\cdot|x)}\left(\langle\eta,S(\mathbf{u})\rangle,\langle\zeta,S(\mathbf{u})\rangle_{\mathcal{H}_S}\right).$$

Taking

$$\eta=\zeta=\lambda_\psi$$

yields

$$\left\langle \lambda_\psi, \nabla^2 A(\lambda_\alpha)[\lambda_\psi] \right\rangle_{\mathcal{H}_S} = \mathrm{Var}_{\mathbf{u}\sim\pi_{\lambda_\alpha}(\cdot|x)} \left[\langle \lambda_\psi, S(\mathbf{u}) \rangle_{\mathcal{H}_S} \right].$$

Combining this with Eq. (18) gives

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} J_{\psi,x}(\alpha) = \mathrm{Var}_{\mathbf{u}\sim\pi_{\lambda_\alpha}(\cdot|x)} \left[\psi\big(F_{\mathbf{u}}(x)\big) \right].$$

Since variance is nonnegative,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} J_{\psi,x}(\alpha) \geq 0.$$

In particular, for any $\alpha_1 \leq \alpha_2$ such that $\lambda_\alpha \in \Theta$ for all $\alpha \in [\alpha_1, \alpha_2]$,

$$\begin{aligned} J_{\psi,x}(\alpha_2) - J_{\psi,x}(\alpha_1) &= \int_{\alpha_1}^{\alpha_2} \frac{\mathrm{d}}{\mathrm{d}\alpha} J_{\psi,x}(\alpha)\, \mathrm{d}\alpha \\ &= \int_{\alpha_1}^{\alpha_2} \mathrm{Var}_{\mathbf{u}\sim\pi_{\lambda_\alpha}(\cdot|x)} \left[\psi\big(F_{\mathbf{u}}(x)\big) \right] \mathrm{d}\alpha \\ &\geq 0. \end{aligned}$$

Therefore $J_{\psi,x}(\alpha)$ is nondecreasing along the steering path $\lambda_\alpha = \lambda(x) + \alpha\lambda_\psi(x)$. □

C Detailed Derivations

C.1 KL Projection and the Induced Steering Path

We provide a derivation of the reverse-KL projection used in Eq. (2). Consider the exponential family induced by the base measure μ

$$\frac{\mathrm{d}\pi_\lambda}{\mathrm{d}\mu}(\mathbf{u}) = \exp(\langle \lambda, S(\mathbf{u})\rangle_{\mathcal{H}_S} - A(\lambda)), \qquad \lambda \in \Theta \subseteq \mathcal{H}_S,$$

where $A(\lambda) = \log \mathbb{E}_{\mathbf{u}\sim\pi_0}\left[e^{\langle \lambda, S(\mathbf{u})\rangle_{\mathcal{H}_S}}\right]$ is the log-partition function. We assume that the exponential family is regular and minimal, so that A is strictly convex on Θ. Its expectation coordinate is

$$\varphi(\lambda) = \nabla A(\lambda) = \mathbb{E}_{\mathbf{u}\sim\pi_\lambda}[S(\mathbf{u})], \qquad \Phi = \varphi(\Theta).$$

For a fixed state $x \in \mathcal{X}$, recall the state-conditioned steering direction

$$\lambda_\psi(x) = L_\psi^* \lambda(x) \in \mathcal{H}_S.$$

By construction, the expected propagated QoI is affine in the dual coordinate:

$$\begin{aligned}\mathbb{E}_{\mathbf{u}\sim\pi_\lambda}\left[\psi(F_{\mathbf{u}}(x))\right] &= \langle \lambda(x), L_\psi \mathbb{E}_{\mathbf{u}\sim\pi_\lambda}[S(\mathbf{u})]\rangle_{\mathcal{H}_S} \\ &= \left\langle L_\psi^*\lambda(x), \varphi(\lambda)\right\rangle_{\mathcal{H}_S} \\ &= \langle \lambda_\psi(x), \varphi(\lambda)\rangle_{\mathcal{H}_S}.\end{aligned}$$

Hence, a target expected QoI level $c \in \mathbb{R}$ defines the affine hyperplane

$$\Phi_{\psi,x}(c) = \{\varphi \in \Phi : \langle \lambda_\psi(x), \varphi\rangle_{\mathcal{H}_S} = c\}.$$

KL divergence as a dual Bregman divergence. For any $\lambda, \lambda_0 \in \Theta$, the KL divergence between two members of the same exponential family satisfies

$$\begin{aligned}D_{\mathrm{KL}}(\pi_\lambda \| \pi_{\lambda_0}) &= \mathbb{E}_{\pi_\lambda}\left[\log \frac{\mathrm{d}\pi_\lambda}{\mathrm{d}\pi_{\lambda_0}}(\mathbf{u})\right] \\ &= \langle \lambda - \lambda_0, \varphi(\lambda)\rangle_{\mathcal{H}_S} - A(\lambda) + A(\lambda_0) \\ &= D_A(\lambda_0, \lambda),\end{aligned}$$

where $D_A(\lambda_0, \lambda) := A(\lambda_0) - A(\lambda) - \langle \nabla A(\lambda), \lambda_0 - \lambda\rangle$ is the Bregman divergence generated by A.

Let A^* denote the convex conjugate of A,

$$A^*(\varphi) := \sup_{\lambda\in\Theta} \{\langle \lambda, \varphi\rangle_{\mathcal{H}_S} - A(\lambda)\}.$$

The Legendre correspondence gives

$$\varphi = \nabla A(\lambda), \qquad \lambda = \nabla A^*(\varphi),$$

and therefore

$$D_{\mathrm{KL}}(\pi_\lambda \| \pi_{\lambda_0}) = D_{A^*}\big(\varphi(\lambda), \varphi(\lambda_0)\big). \tag{21}$$

Projection onto a constant-QoI hyperplane. We now set $\lambda_0 = \lambda(x)$ and consider

$$\widehat{\lambda} \in \arg\min_{\lambda:\, \varphi(\lambda)\in\Phi_{\psi,x}(c)} D_{\mathrm{KL}}\big(\pi_\lambda \| \pi_{\lambda(x)}\big).$$

Using Eq. (21), this is equivalently the Bregman projection

$$\widehat{\varphi} \in \arg\min_{\varphi\in\Phi} D_{A^*}\big(\varphi, \varphi(\lambda(x))\big) \quad \text{subject to} \quad \langle \lambda_\psi(x), \varphi\rangle_{\mathcal{H}_S} = c.$$

Introduce a Lagrange multiplier $\beta \in \mathbb{R}$:

$$\mathcal{L}(\varphi, \beta) = D_{A^*}\big(\varphi, \varphi(\lambda(x))\big) + \beta\left(\langle \lambda_\psi(x), \varphi\rangle_{\mathcal{H}_S} - c\right).$$

The first-order optimality condition with respect to φ is

$$\nabla A^*(\widehat{\varphi}) - \nabla A^*\big(\varphi(\lambda(x))\big) + \beta\lambda_\psi(x) = 0.$$

By Legendre duality,

$$\nabla A^*(\widehat{\varphi}) = \widehat{\lambda}, \qquad \nabla A^*\big(\varphi(\lambda(x))\big) = \lambda(x),$$

and hence $\widehat{\lambda} = \lambda(x) - \beta\lambda_\psi(x)$. Finally, writing $\alpha = -\beta$ gives

$$\widehat{\lambda} = \lambda(x) + \alpha\lambda_\psi(x).$$

C.2 Oracle Representation and Measurement Map

We first write the dynamics in control-affine form. Since the system is driftless,

$$\dot{x} = f_1(x)v + f_2(x)\omega, \qquad f_1(x) = \begin{bmatrix} 1 \\ \theta \\ 0 \end{bmatrix}, \qquad f_2(x) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Observable and exact oracle representation. We consider the QoI $\psi(x) = p_y$. The first-order Lie derivatives are

$$\mathcal{L}_{f_1}\psi(x) = \nabla\psi(x)^\top f_1(x) = \theta, \qquad \mathcal{L}_{f_2}\psi(x) = \nabla\psi(x)^\top f_2(x) = 0.$$

At second order,

$$\mathcal{L}_{f_2}\mathcal{L}_{f_1}\psi(x) = \mathcal{L}_{f_2}\theta = 1,$$

whereas

$$\mathcal{L}_{f_1}\mathcal{L}_{f_1}\psi = \mathcal{L}_{f_1}\mathcal{L}_{f_2}\psi = \mathcal{L}_{f_2}\mathcal{L}_{f_2}\psi = 0.$$

Since the only second-order iterated Lie derivative that is not identically zero is constant, all iterated Lie derivatives of order three or higher vanish identically.

Strictly speaking, under the augmented formulation with $u^0 \equiv 1$, the input should be written as $\mathbf{u} = (1, v, \omega)$. However, since the time component plays no role in the control parametrization, we suppress it for notational simplicity and write $\mathcal{A} = \{1, 2\}$ as the control alphabet, with 1 and 2 corresponding to v and ω, respectively, and let $\mathcal{A}^\star$ denote the set of all finite words over $\mathcal{A}$, including the empty word $\emptyset$. We use the convention

$$\mathcal{L}_{(i_1,\ldots,i_k)}\psi := \mathcal{L}_{f_{i_1}} \cdots \mathcal{L}_{f_{i_k}}\psi, \qquad \mathcal{L}_\emptyset\psi := \psi,$$

Let $S(\mathbf{u}) = \big(S(\mathbf{u})^I_{0,H}\big)_{I\in\mathcal{A}^\star}$ denote the full signature of the control, with $S^\emptyset_{0,H}(\mathbf{u}) = 1$. The only words whose Chen–Fliess coefficient functions are not identically zero are

$$\mathcal{W}_\psi = \{\emptyset, 1, 21\}.$$

In particular,

$$S(\mathbf{u})^1_{0,H} = \int_0^H v(t)\,\mathrm{d}t, \qquad S(\mathbf{u})^{21}_{0,H} = \int_0^H \int_0^{t_2} \omega(t_1)v(t_2)\,\mathrm{d}t_1\,\mathrm{d}t_2.$$

Hence the Chen–Fliess expansion terminates exactly:

$$\psi(F_\mathbf{u}(x)) = \sum_{I\in\mathcal{A}^\star} \mathcal{L}_I\psi(x) S(\mathbf{u})^I_{0,H} = p_y + \theta S(\mathbf{u})^1_{0,H} + S^{21}_{0,H}(\mathbf{u}).$$

Define

$$\lambda_\psi(x) = \big(\lambda^I_\psi(x)\big)_{I\in\mathcal{A}^\star}, \qquad \lambda^I_\psi(x) := \mathcal{L}_I\psi(x).$$

Since $\lambda_\psi(x)$ has finite support $\mathcal{W}_\psi$, its pairing with the full signature is a finite sum:

$$\psi(F_\mathbf{u}(x)) = \langle \lambda_\psi(x), S(\mathbf{u})\rangle_{\mathcal{H}_S}.$$

The only coordinates of $\lambda_\psi(x)$ that are not identically zero are

$$\lambda^\emptyset_\psi(x) = p_y, \qquad \lambda^1_\psi(x) = \theta, \qquad \lambda^{21}_\psi(x) = 1.$$

Let $\Pi_{\mathcal{W}_\psi}$ denote projection onto these three active coordinates. Then

$$\Pi_{\mathcal{W}_\psi} S(\mathbf{u}) = \begin{bmatrix} 1 \\ S(\mathbf{u})^1_{0,H} \\ S(\mathbf{u})^{21}_{0,H} \end{bmatrix}, \qquad \Pi_{\mathcal{W}_\psi}\lambda_\psi(x) = \begin{bmatrix} p_y \\ \theta \\ 1 \end{bmatrix}.$$

With a slight abuse of notation, whenever only these active coordinates are considered, we write

$$S(\mathbf{u}) = \begin{bmatrix} 1 \\ S(\mathbf{u})^1_{0,H} \\ S(\mathbf{u})^{21}_{0,H} \end{bmatrix}, \qquad \lambda_\psi(x) = \begin{bmatrix} p_y \\ \theta \\ 1 \end{bmatrix}.$$

Signature measurement map L_ψ. For a learned VLA representation λ, the particular realization of L_ψ may depend on how the VLA model is trained, and more specifically, on how much physical information is encoded in λ. In particular, our oracle experiment does not assume or learn a specific realization of L_ψ; rather, L_ψ is treated as given, and the experiment is intended to demonstrate the existence of such a linear correspondence. More generally, how the representation dimensionality scales with the number of QoIs, as well as possible intersections among the associated QoI subspaces, requires further consideration.

D EXISTING EMPIRICAL RESULTS WITHIN THE PROPOSED FRAMEWORK

In this section, we show how existing empirical results on linear probing of physical quantities in VLA models (Lu et al., 2025; Zhang et al., 2026; Garg et al., 2026) can be interpreted as special cases of the proposed framework.

Assume the robot state can be written as $x = (p_x, p_y, p_z, \theta_r, \theta_p, \theta_y, g)$, where $(p_x, p_y, p_z) \in \mathbb{R}^3$ represent Cartesian position, $(\theta_r, \theta_p, \theta_y) \in [0, 2\pi)^3$ encode roll, pitch, and yaw orientation, and $g \in [0, 1]$ represents normalized gripper aperture. Time-augmented robot action trajectory $\mathbf{u} = (t, \Delta x) \in \mathcal{U}_H$ corresponds to relative displacements in this state space.

Existing VLA probing studies primarily examine whether two types of physical quantities $\psi : \mathcal{X} \to \mathbb{R}$ can be linearly decoded from frozen representations: (i) position-level quantities, such as $p_x, p_y, p_z, \theta_r, \theta_p, \theta_y$, and (ii) velocity-level quantities, such as $\Delta p_x, \Delta p_y, \Delta p_z, \Delta\theta_r, \Delta\theta_p, \Delta\theta_y$. These quantities are typically probed using a fixed linear readout,

$$\psi(x) \approx \langle W_\psi \lambda(x),\, w_\psi \rangle\,,$$

where W_ψ is a learned linear map and w_ψ denotes the corresponding readout direction in the transformed representation space. Note that our formulation has considered the evolution of the same physical quantity under a candidate action trajectory:

$$\psi(F_{\mathbf{u}}(x)) \approx \langle \lambda(x), \gamma_\psi(\mathbf{u}) \rangle = \langle \lambda(x), L_\psi S(\mathbf{u}) \rangle = \left\langle L_\psi^* \lambda(x), S(\mathbf{u}) \right\rangle.$$

At a high level, when $S(\mathbf{u})$ contains sufficiently rich nonzero coordinates, the role previously played by a single readout direction w_ψ can be generalized across multiple signature coordinates, yielding a substantially richer action-conditioned readout.

D.1 PROBING POSITION VECTORS

We first consider snapshot-level physical quantities, such as the current position (p_x, p_y, p_z) and orientation $(\theta_r, \theta_p, \theta_y)$, where each QoI ψ is associated with a fixed linear probe.

Under our formulation, exact static probing is recovered in the zero-horizon case $H = 0$, for which only the empty-word coordinate remains. A closely related short-horizon setting is obtained for small $H > 0$ with no nontrivial control input:

$$S(\mathbf{u}) = \big(S(\mathbf{u})^{\emptyset}_{0,H} = 1,\, S(\mathbf{u})^{0}_{0,H} = H, 0, \cdots, S(\mathbf{u})^{00}_{0,H} = \frac{1}{2}H^2, \cdots, 0, \cdots\big) \in \mathcal{S}.$$

Consequently, only the empty-word coordinate and the pure time-word coordinates of the form $0 \cdots 0 \in \mathcal{W}$ contribute to the probe. For a fixed horizon H, these coordinates scale as $H^k/k!$ with the word length k, and hence decay rapidly at higher orders. For small H, it is therefore natural to retain only the first Q relevant signature coordinates, so that only the corresponding columns of L_ψ,

$$(L_\psi)_{:,1:Q},$$

are needed to recover the static physical quantity. Specifically,

$$\psi(x) = \langle \lambda(x), \gamma_\psi \rangle = \langle \lambda(x), L_\psi S(\mathbf{u}) \rangle \approx \left\langle \lambda(x), (L_\psi)_{:,1:Q}\, \bar{S}(\mathbf{u}) \right\rangle,$$

where $\bar{S}(\mathbf{u}) \in \mathbb{R}^Q$ collects the retained signature coordinates.

Equivalently, by setting

$$W_\psi := (L_\psi)^*_{:,1:Q}, \qquad w_\psi := \bar{S}(\mathbf{u}),$$

we obtain

$$\psi(x) \approx \langle W_\psi \lambda(x),\, w_\psi \rangle\,,$$

which recovers the conventional linear probe for current position or orientation as a special case of our action-conditioned formulation.

D.2 Probing Velocity Vectors

Next, consider velocity-level quantities such as $\Delta p_x, \Delta p_y, \Delta p_z$ and $\Delta\theta_r, \Delta\theta_p, \Delta\theta_y$. We interpret these quantities as short-horizon changes of the corresponding physical states. Specifically, for a QoI ψ, define

$$\Delta_H\psi(x;\mathbf{u}) := \psi(F_{\mathbf{u}}(x)) - \psi(x),$$

where $\mathbf{u}$ is an action trajectory applied over the horizon $[0, H]$.

For example, consider a constant action applied only along the first action coordinate,

$$\mathbf{u}(t) = (1, \Delta, 0, \dots, 0), \qquad t \in [0, H],$$

where the first coordinate corresponds to the augmented time channel $u^0(t) \equiv 1$, and $u^1(t) \equiv \Delta$. Its signature begins as

$$S(\mathbf{u}) = \left(1,\ H,\ \Delta H,\ \frac{H^2}{2},\ \frac{\Delta H^2}{2},\ \frac{\Delta H^2}{2},\ \frac{\Delta^2 H^2}{2}, \dots\right).$$

More generally, for any word $I \in \{0,1\}^k$,

$$S^I_{0,H}(\mathbf{u}) = \frac{H^k}{k!}\Delta^{n_1(w)},$$

where $n_1(w)$ denotes the number of occurrences of the active action letter 1.

Unlike the static case, the signature now contains nonzero coordinates involving the action letter 1. Hence, the probe can exploit both time-only and action-dependent signature coordinates. Retaining the first Q relevant coordinates gives

$$\psi(F_{\mathbf{u}}(x)) = \langle \lambda(x), L_\psi S(\mathbf{u})\rangle \approx \left\langle \lambda(x), (L_\psi)_{:,\mathcal{I}_Q}\, \bar{S}(\mathbf{u})\right\rangle,$$

where $\mathcal{I}_Q$ indexes the retained nonzero signature coordinates and $\bar{S}(\mathbf{u}) \in \mathbb{R}^Q$ collects their values.

For a fixed short-horizon action trajectory $\mathbf{u}$, the vector $\bar{S}(\mathbf{u})$ is fixed. Therefore, by defining

$$W_\psi := (L_\psi)^*_{:,\mathcal{I}_Q}, \qquad w_\psi := \bar{S}(\mathbf{u}),$$

we recover

$$\psi(F_{\mathbf{u}}(x)) \approx \langle W_\psi \lambda(x), w_\psi\rangle.$$

Accordingly, the corresponding finite-horizon change

$$\Delta_H\psi(x;\mathbf{u}) = \psi(F_{\mathbf{u}}(x)) - \psi(x)$$

also admits a fixed linear readout from $\lambda(x)$ for a fixed trajectory $\mathbf{u}$. Thus, conventional probing of velocity-level quantities such as Δp_x or $\Delta\theta_y$ can be interpreted as a special case of our action-conditioned formulation in which the candidate trajectory and horizon are fixed. Our formulation generalizes this setting by allowing $\mathbf{u}$ to vary, so that the probe direction changes with the contemplated action trajectory.