Steering as Implicit Low-Rank Finetuning: A Theoretical Framework for Activation Steering in Transformers

Anonymous Authors[1]

Abstract

Activation steering and supervised finetuning are common ways to change language model behavior, yet their mechanistic connection remains unclear. We identify when low-rank finetuning produces steering-like residual shifts and when it does not. We study this question in a controlled linear-attention setting by comparing rank-one LoRA updates to the residual-stream geometry assumed by fixed-direction steering. We separate the value and output pathway, which writes information into the residual stream, from the query and key pathway, which routes attention across context. We prove that rank-one updates to value or output matrices force all residual-stream shifts to lie in a single fixed direction, independent of the data distribution. In contrast, rank-one query and key updates induce context-dependent shift directions, so their resemblance to fixed-direction steering depends on the input distribution. Empirically, value and output updates produce exactly rank-one activation shifts, while query and key updates are less rank-concentrated and require larger weight changes for comparable behavioral effects. When all attention matrices are trainable, component replay shows that query and key routing contributes meaningfully and interacts non-additively with value and output writing.

1. Introduction

Activation steering and supervised finetuning are two common ways to change the behavior of a trained language model. Activation steering modifies the residual stream at inference time by adding a fixed direction u, while leaving the weights unchanged (Turner et al., 2024; Panickssery et al., 2023; Li et al., 2023; Zou et al., 2023). Supervised finetuning instead modifies the weights through gradient descent on a curated dataset, producing changes that persist across inputs (Hu et al., 2022; Ren & Sutherland, 2024). Although these methods act on different objects, activations in one case and weights in the other, they can produce strikingly similar effects. A steered model and a finetuned model may achieve similar accuracy and may even succeed and fail on the same examples. This raises the mechanistic question of when finetuning behaves like activation steering, and when this relationship breaks down.

This question connects to a growing body of work showing that meaningful model behaviors are often represented by approximately linear directions in activation space (Park et al., 2023; Marks & Tegmark, 2023). Work on activation steering uses such directions to intervene on model behavior directly (Turner et al., 2024; Panickssery et al., 2023; Li et al., 2023; Zou et al., 2023; Stolfo et al., 2025), while work on superposition and sparse autoencoders shows that residual stream directions can encode interpretable and overlapping features (Elhage et al., 2022a;b; Bricken et al., 2023; Templeton et al., 2024). At the same time, parameter efficient finetuning methods such as LoRA constrain weight changes to low rank subspaces (Hu et al., 2022). Low rank in weight space, however, does not by itself imply low rank in activation space. Our goal is to identify when such a connection is structurally forced, and when it is not.

We argue that the answer depends on which attention pathway the weight update passes through. In a linear attention layer, the value and output projections W_V and W_O act as write operations. We prove that any rank-one update to either matrix produces a rank-one residual stream displacement, an algebraic consequence of the write structure that holds independently of the task or data distribution. This gives an exact mechanistic explanation for why V and O updates can resemble fixed-direction activation steering.

The query and key projections W_Q and W_K behave differently. Following the transformer circuits distinction between QK and VO mechanisms, QK determines where attention is routed, while VO determines what is written to the residual stream (Elhage et al., 2021; Olsson et al., 2022). Rank-one updates to W_Q and W_K modify attention routing and produce residual shifts whose directions depend on the full

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

input context. Such updates resemble activation steering only when these context-dependent directions are nearly collinear across the dataset, a property that is not guaranteed by algebra and must be verified empirically.

Our central contribution is to identify this structural asymmetry precisely. For the value and output pathway, rank-one updates are algebraically forced to produce rank-one residual stream displacements, regardless of the data distribution. For the query and key pathway, the rank of the induced shifts depends on the data distribution rather than algebra alone. This asymmetry is not visible from behavioral accuracy, since both pathways can solve the same alignment task, and it requires geometric diagnostics to detect. More broadly, the results suggest that steering-like behavior is not a generic consequence of low-rank finetuning. It arises when the learned update changes what is written into the residual stream rather than how information is routed through context.

2. Background and Motivation

We build intuition for the main results by first considering a simple setting where steering can be written as a low-rank weight edit, then explaining why a weaker moment-level notion is more appropriate for attention layers, and finally motivating why the VO and QK pathways implement structurally different mechanisms.

2.1. Equivalence Between Steering and Weight Updates

The most common form of activation steering adds a fixed direction u to the residual stream at inference time,

$$h' = h + \alpha u, \tag{1}$$

where $\alpha \in \mathbb{R}$ is a fixed scalar coefficient (Turner et al., 2024; Panickssery et al., 2023). A more general family allows the coefficient to depend on the input,

$$h' = h + \alpha(h)\, u, \tag{2}$$

where $\alpha(h)$ is a scalar function of the activation. Dynamic projection steering is one instance of this family, in which $\alpha(h) = z - u^\top h$ for a target scalar z and unit direction u, so that the component of h along u is replaced by z (Li et al., 2023; Zou et al., 2023). The general family (2) is the more natural counterpart to VO LoRA updates, which produce shifts of the form $\delta h_i = s_i v$ for a fixed direction v and input-dependent scalar s_i, and it provides the right conceptual frame for the mean equivalence discussion below.

To see why low-rank weight updates can resemble fixed-direction steering, consider an affine layer $h = Wx + b$ with $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and $b, h \in \mathbb{R}^{d_{\text{out}}}$. For dynamic projection steering the intervention is

$$h' = (I - uu^\top)h + zu. \tag{3}$$

By direct substitution, this is exactly equivalent for every input x to the parameter edit

$$W' = (I - uu^\top)W, \qquad b' = (I - uu^\top)b + zu, \tag{4}$$

where the weight change $\Delta W = -uu^\top W$ has rank at most one. Thus in this affine setting a fixed-direction steering intervention can be represented exactly as a rank-one weight edit. This pointwise equivalence is available because the coefficient $\alpha(h) = z - u^\top h$ is affine in h; for a constant coefficient the edit reduces to a bias shift and the weight change is zero, which is less analogous to the bias-free LoRA setting.

Attention layers do not admit this pointwise equivalence directly. Their projections are bias-free, their inputs depend on the full residual stream, and changes to attention routing alter which values are read. We therefore use a weaker notion: a weight update is *mean-equivalent* to a steering intervention if it matches the first moment of the resulting activation distribution,

$$\mathbb{E}[\tilde{h}] = \mathbb{E}[h + \alpha(h)u], \tag{5}$$

where $\tilde{h}$ is the activation produced by the edited weights and the expectation is over the empirical input distribution. Mean equivalence is useful as an explanation for directional alignment: in the VO pathway, the minimum-norm update satisfying (5) for a target mean displacement is rank-one and aligns its left singular direction with u, which predicts positive cosine alignment when gradient descent has an implicit low-norm bias (Arora et al., 2019).

The alignment task studied here requires a conditional version of this target. The intervention should shift activations on negative-control prompts while leaving positive-control prompts approximately unchanged,

$$\mathbb{E}[\Delta h \mid C^-] = \bar{\alpha}\, u, \qquad \mathbb{E}[\Delta h \mid C^+] \approx 0. \tag{6}$$

Unconditional mean equivalence is therefore neither necessary nor sufficient for the alignment objective, and the discussion above is motivational rather than definitional. Because the steering coefficient is chosen from a fixed sweep rather than optimized to match a particular finetuned model, we do not treat steering and LoRA as matched-strength interventions. Our comparison is geometric: we ask whether the residual shifts induced by finetuning lie in a low-dimensional, fixed-direction subspace of the kind that activation steering assumes. The exact results in Section 3 do not rely on affine pointwise equivalence; instead they use structural rank arguments to determine when low-rank weight updates force fixed-direction residual shifts, and when such steering-like behavior depends on the data distribution.

2.2. Two Pathways, Two Mechanisms

The affine analysis motivates looking for rank-one structure in the residual-stream displacements induced by finetuning. In a linear attention layer, however, the two weight-matrix groups (W_V, W_O) and (W_Q, W_K) implement structurally different operations, and this difference has direct consequences for whether their finetuning updates resemble fixed-direction steering (Elhage et al., 2021; Olsson et al., 2022).

The *value/output* pathway is a write operation. W_V computes a feature vector from the attention-weighted input, and W_O projects that feature directly into the residual stream. A rank-one update to either matrix changes a single outer-product factor in this write, which we will show forces all per-prompt residual shifts to lie in a single fixed direction. This is an algebraic consequence of the write structure rather than a distributional property: it holds regardless of the task or the activation distribution.

The *query/key* pathway is a routing operation. W_Q and W_K jointly determine the attention score matrix $A(H) = H^\top W_Q^\top W_K H$, which controls how much each value vector contributes to each output position. A rank-one update to W_Q or W_K reroutes attention, changing which values are read rather than what is written. The induced residual-stream shift depends on which values are now emphasized, and those values vary with the input context. Whether such routing changes resemble fixed-direction steering therefore depends on whether the context-dependent shift directions are nearly collinear across the data distribution, a property that is not algebraically guaranteed and must be tested empirically.

This structural asymmetry between the two pathways is the central organizing idea of the theory that follows. It predicts the rank structure of individual LoRA updates and provides the language for interpreting all-matrices supervised finetuning as an interaction between writing and routing: when all parameters are jointly optimized, the solution coordinates VO writing with QK routing, with the two components interacting non-additively.

3. Steering as Low-Rank Finetuning in Linear Attention

3.1. Setup

We use column vector convention throughout. Let $H \in \mathbb{R}^{d\times T}$ be the residual stream matrix for a sequence of length T, with columns $h_t \in \mathbb{R}^d$. We study a single-head unnormalized linear attention layer (Katharopoulos et al., 2020) with $W_Q, W_K \in \mathbb{R}^{d_k\times d}$, $W_V \in \mathbb{R}^{d_v\times d}$, and $W_O \in \mathbb{R}^{d\times d_v}$. The layer computes

$$Q = W_Q H, \qquad K = W_K H, \qquad V = W_V H,$$

with attention scores

$$A(H) = Q^\top K = H^\top W_Q^\top W_K H,$$

feature matrix $Y = V A(H)$, and residual update

$$H^+ = H + W_O Y.$$

For position t, this is $h_t^+ = h_t + W_O y_t$.

We compare low-rank weight updates to residual stream steering interventions of the form $h_t^+ \mapsto h_t^+ + \alpha(h_t^+)u$, where $u \in \mathbb{R}^d$ is a fixed direction. Two natural choices of u arise from the residual stream activations after the first attention layer. The mean difference direction is

$$u_{\text{mean}} = \frac{\mathbb{E}[h_t^+ \mid C^+] - \mathbb{E}[h_t^+ \mid C^-]}{\|\mathbb{E}[h_t^+ \mid C^+] - \mathbb{E}[h_t^+ \mid C^-]\|},$$

which captures the average displacement between the two behavior classes in activation space (Panickssery et al., 2023). The probe direction $u_{\text{probe}} = w/\|w\|$ is the normalized weight vector of a linear classifier trained to distinguish C^+ from C^- activations, and identifies a direction that is discriminative rather than merely mean-separating (Li et al., 2023). Both directions are fixed at inference time and used to measure how well the learned residual shifts align with interpretable steering targets.

3.2. The Value and Output Pathway

The VO pathway is a write operation. A rank-one update to either W_O or W_V changes a single outer-product factor in this write, which forces the resulting residual shifts to lie in one fixed direction.

Theorem 3.1 (Rank-one VO updates produce rank-one residual shifts)**.** *Let $\Delta W_O = b_O a_O^\top$ be a rank-one update to the output projection, where $b_O \in \mathbb{R}^d$ and $a_O \in \mathbb{R}^{d_v}$. For any feature vectors $y_1, \dots, y_n \in \mathbb{R}^{d_v}$, the induced residual stream shifts satisfy*

$$\delta h_i = \Delta W_O\, y_i = (a_O^\top y_i)\, b_O. \tag{7}$$

Thus $\Delta H = [\delta h_1, \dots, \delta h_n]$ has $\operatorname{rank}(\Delta H) \le 1$, with left singular direction $b_O/\|b_O\|$.

Let $\Delta W_V = b_V a_V^\top$ be a rank-one update to the value projection, where $b_V \in \mathbb{R}^{d_v}$ and $a_V \in \mathbb{R}^d$. If $\bar{h}_i$ denotes the attention-weighted input at the final position for prompt i, then

$$\delta h_i = W_O\, \Delta W_V\, \bar{h}_i = (a_V^\top \bar{h}_i)\, W_O b_V. \tag{8}$$

Thus $\Delta H = [\delta h_1, \dots, \delta h_n]$ has $\operatorname{rank}(\Delta H) \le 1$, with left singular direction $W_O b_V/\|W_O b_V\|$.

Proof. Equations (7) and (8) show that the stacked shift matrices are outer products, $\Delta H = b_O s^\top$ and $\Delta H = W_O b_V r^\top$, with $s_i = a_O^\top y_i$ and $r_i = a_V^\top \bar{h}_i$. Each therefore has rank at most one. □

When W_V and W_O are updated together while QK is fixed, write $y_i = W_V \bar{h}_i$. The total shift is

$$\delta h_i^{V+O} = \Delta W_O y_i + W_O \Delta W_V \bar{h}_i + \Delta W_O \Delta W_V \bar{h}_i. \quad (9)$$

Proposition 3.2 (Joint VO updates have rank at most two)**.** *Assume QK is fixed and let $\Delta W_O = b_O a_O^\top$ and $\Delta W_V = b_V a_V^\top$ be rank-one updates. Then all final-position residual shifts δh_i^{V+O} lie in $\mathrm{span}\{b_O, W_O b_V\}$, and the stacked matrix ΔH^{V+O} has rank at most two. The first and third terms in (9) are scalar multiples of b_O, while the second is a scalar multiple of $W_O b_V$, so the span claim follows immediately.*

Thus individual VO updates are exactly one-dimensional in activation space, while joint VO updates are confined to at most two write directions. Whether the joint update collapses to one dominant direction is an optimization property rather than an algebraic necessity.

3.3. Alignment Between Learned Directions and Steering Vectors

The rank result determines the dimensionality of the VO shift, but not which direction is learned. A simple mean-matching calculation gives a useful explanation. Let $\mu_y = \mathbb{E}[y_t]$ and suppose $\mathbb{E}[\delta h_t] = \bar{\alpha} u$. The minimum Frobenius norm update to W_O satisfying this condition is

$$\Delta W_O^* = \bar{\alpha}\, u \, \frac{\mu_y^\top}{\|\mu_y\|^2}. \quad (10)$$

It is rank-one and its left singular direction is $u/\|u\|$. This does not prove that gradient descent finds this update, but it explains why low-norm VO solutions should tend to align with steering directions when optimization has an implicit low-norm bias (Arora et al., 2019). The mismatch between the minimum-norm update and a pointwise match depends on the dispersion of features around their mean: for W_O updates the relevant features are y_t, while for W_V updates they are the attention-weighted inputs $\bar{h}_i$. A conditional version based on $\mathbb{E}[y_t \mid C^-] - \mathbb{E}[y_t \mid C^+]$ better matches the alignment task and naturally connects the learned direction to u_{mean}.

3.4. The Query and Key Pathway

The QK pathway changes attention routing rather than directly changing what is written (Elhage et al., 2021). Let $M = W_Q^\top W_K$, so that $A(H) = H^\top M H$, and let $C = W_O W_V H$. A perturbation ΔM changes scores by $\Delta A(H) = H^\top \Delta M H$ and induces

$$\Delta H^{\mathrm{QK}} = C\, H^\top \Delta M H. \quad (11)$$

Theorem 3.3 (Rank-one QK updates produce context-dependent shifts)**.** *Suppose $\Delta M = ab^\top$ is rank one. Define the context-dependent direction $d_H = CH^\top a \in \mathbb{R}^d$. Then the final-position residual shift for prompt i is*

$$\delta h_T^{(i)} = (b^\top H^{(i)} e_T)\, d_{H^{(i)}}, \quad (12)$$

where e_T selects the final column of H. For fixed $H^{(i)}$, the shift lies in $\mathrm{span}(d_{H^{(i)}})$.

Proof. Since

$$\Delta H^{\mathrm{QK}} = CH^\top (ab^\top) H = (CH^\top a)(b^\top H),$$

the final column is $d_H (b^\top H e_T)$. □

If both W_Q and W_K receive rank-one updates, then

$$\Delta M = \Delta W_Q^\top W_K + W_Q^\top \Delta W_K + \Delta W_Q^\top \Delta W_K,$$

where ΔM denotes the total score change induced by the QK updates. This has rank at most three, or at most two when the second-order term is ignored. For a fixed sequence, this bounds the token-wise residual shift, but across a dataset the final-position shifts can have higher rank because the directions $d_{H^{(i)}}$ vary with context.

Corollary 3.4 (QK steering-likeness depends on context direction variation)**.** *Define*

$$D_a = \left[(b^\top H^{(1)} e_T)\, d_{H^{(1)}},\ \ldots,\ (b^\top H^{(n)} e_T)\, d_{H^{(n)}}\right].$$

Then $\Delta H^{\mathrm{QK}} = D_a$. If the directions $d_{H^{(i)}}$ are nearly collinear across prompts, the update resembles fixed-direction steering. If they vary substantially, no single steering direction captures the full update.

Thus QK updates can be steering-like, but this is a property of the data distribution rather than an algebraic guarantee.

Applying residual steering $H' = H + us^\top$ also perturbs QK scores to first order as

$$\Delta A \approx H^\top M u s^\top + s u^\top M H. \quad (13)$$

Each term has rank one, so the first-order score perturbation has rank at most two. This already suggests that writing and routing need not remain cleanly separate in a full transformer pass.

3.5. All-Matrix Updates and Interaction Structure

When all first-layer attention matrices are trained jointly, the full residual shift is not generally the sum of an isolated VO shift and an isolated QK shift.

Proposition 3.5 (All-matrices updates contain interaction terms)**.** *Let*

$$A' = A + \Delta A,$$
$$W_O' = W_O + \Delta W_O,$$
$$W_V' = W_V + \Delta W_V,$$

where ΔA denotes the total attention-score change induced by QK updates. Then

$$\Delta H_{\text{all}} = \Delta H_{\text{QK}} + \Delta H_{\text{VO}} + R_{\text{int}}, \tag{14}$$

where

$$\Delta H_{\text{QK}} = W_O W_V H \Delta A,$$
$$\Delta H_{\text{VO}} = \Delta W_O W_V H A + W_O \Delta W_V H A + \Delta W_O \Delta W_V H A,$$

and

$$\begin{aligned} R_{\text{int}} = {} & \Delta W_O W_V H \Delta A \\ & + W_O \Delta W_V H \Delta A \\ & + \Delta W_O \Delta W_V H \Delta A. \end{aligned}$$

Thus the all-matrices shift is not generally additive in the VO and QK components.

The interaction terms arise because ΔA changes which values are routed, while ΔW_V and ΔW_O change what those routed values write into the residual stream. In the experiments, we measure this effect using counterfactual component replay, defined in Section 4.

4. Experimental Setup

We test the predictions of Section 3 in a controlled setting designed to isolate residual-stream geometry in linear attention.

We use a binary behavior-control task with a BOS token, a query token Q, control tokens C^+ and C^-, topic tokens $\{T_i\}_{i=1}^n$, and paired answers $\{A_i^+, A_i^-\}_{i=1}^n$. Each input has the form

$$[\text{BOS}, T_i, C^c, Q], \qquad c \in \{+, -\}.$$

During pretraining,

$$[\text{BOS}, T_i, C^+, Q] \mapsto A_i^+, \qquad [\text{BOS}, T_i, C^-, Q] \mapsto A_i^-.$$

During alignment finetuning, negative-control prompts are retargeted to A_i^+ while positive-control prompts retain target A_i^+. Before any intervention, the pretrained model predicts A_i^- on all negative-control prompts, giving alignment accuracy zero and a clean baseline for measuring behavioral change.

Model. We use a two-layer decoder-only transformer with single-head linear attention, sequence length $T = 4$, no MLPs, and no normalization layers. The main width sweep uses $d \in \{32, 64, 96, 128\}$. Only the first attention layer is trainable during intervention experiments. The second layer, token embeddings, positional embeddings, and unembedding matrix are frozen throughout, ensuring that any behavioral change induced by LoRA must pass through the first-layer residual stream, which is the object of our geometric analyses. We initialize W_V and W_O near the identity with small Gaussian perturbations, and initialize W_Q and W_K from a zero-mean Gaussian.

Training. Pretraining is run on the balanced behavior-control task until accuracy exceeds 0.99, or until 200 epochs. LoRA finetuning is run for up to 200 epochs and is stopped once both alignment and retention accuracy exceed 0.99. The SFT objective is

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{align}} + \lambda \mathcal{L}_{\text{retain}},$$

where $\mathcal{L}_{\text{align}}$ is cross entropy on negative-control prompts with target A_i^+, and $\mathcal{L}_{\text{retain}}$ is cross entropy on positive-control prompts with target A_i^+. The main experiments use $\lambda \in \{0.1, 0.5\}$. Robustness sweeps vary $n \in \{8, 24, 48\}$ and include additional retention settings where noted. Unless otherwise stated, all reported statistics are averaged across 30 random seeds.

Activation steering. We intervene after the first attention layer at the final sequence position by setting

$$H_T^{(1)} \leftarrow H_T^{(1)} + \alpha u,$$

using u_{mean} and u_{probe} as defined in Section 3.1, and sweeping $\alpha \in \{\pm 1, \pm 2, \pm 3, \pm 5, \pm 10, \pm 20\}$. Because α is chosen from a fixed sweep rather than optimized to match a particular LoRA update, steering comparisons assess directional and effect-pattern similarity rather than matched intervention strength.

Rank-one LoRA finetuning. We finetune the first attention layer using rank-one LoRA. For each target matrix W, the update is parameterized as $\Delta W = BA$ with rank-one factors B and A, and the base matrix is frozen. We apply LoRA to selected subsets of $\{W_Q^{(1)}, W_K^{(1)}, W_V^{(1)}, W_O^{(1)}\}$, giving seven configurations: $\{W_O\}$, $\{W_V\}$, $\{W_Q\}$, $\{W_K\}$, $\{W_Q, W_K\}$, $\{W_V, W_O\}$, and $\{W_Q, W_K, W_V, W_O\}$. The last is the all-matrices condition.

Residual-shift matrix. All residual-shift metrics are computed at the final sequence position after the first attention layer on negative-control prompts. For $x_i^- = [\text{BOS}, T_i, C^-, Q]$, define

$$\delta h_i = H_{T,\text{intervention}}^{(1)}(x_i^-) - H_{T,\text{base}}^{(1)}(x_i^-),$$

and stack these shifts as $\Delta H = [\delta h_1, \ldots, \delta h_n]$.

Component replay. For the all-matrices condition, let θ_0 be the pretrained model and $\Delta\theta_{\text{all}}$ the learned update when all first-layer attention matrices are trainable. We replay the learned VO and QK components in isolation,

$$\Delta\theta_{\text{VO}} = \{\Delta W_V, \Delta W_O\}, \qquad \Delta\theta_{\text{QK}} = \{\Delta W_Q, \Delta W_K\},$$

giving three residual-shift matrices,

$$\begin{aligned} \Delta H_{\text{VO|all}} &= H(\theta_0 + \Delta\theta_{\text{VO}}) - H(\theta_0), \\ \Delta H_{\text{QK|all}} &= H(\theta_0 + \Delta\theta_{\text{QK}}) - H(\theta_0), \\ \Delta H_{\text{all}} &= H(\theta_0 + \Delta\theta_{\text{all}}) - H(\theta_0), \end{aligned}$$

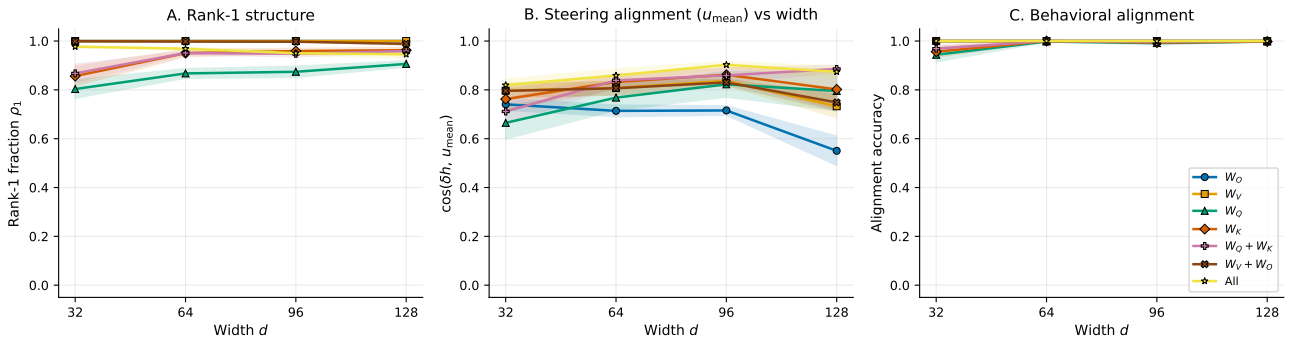

Figure 1. Standalone rank-one LoRA SFT across model width $d \in \{32, 64, 96, 128\}$, averaged over 30 seeds at $n = 24$, $\lambda = 0.1$. **(A)** Rank-one energy fraction ρ_1 of the stacked residual-shift matrix. Individual W_O and W_V updates are pinned at $\rho_1 = 1.000$ across all widths, as predicted by Theorem 3.1. QK configurations are less rank-concentrated, especially W_Q. **(B)** Cosine alignment between the mean SFT-induced shift and u_{mean}. VO alignment degrades with width, more steeply for W_O than W_V, consistent with the minimum-norm alignment argument in Section 3.3. **(C)** Alignment accuracy. All configurations saturate near 1.000, showing that behavioral accuracy alone does not distinguish mechanistic structure. Shaded regions denote standard error across seeds.

and the interaction residual $R_{\text{int}} = \Delta H_{\text{all}} - \Delta H_{\text{VO|all}} - \Delta H_{\text{QK|all}}$.

Metrics. The metrics below separate behavioral success from residual-shift geometry and component-level mechanism, corresponding directly to the theoretical predictions of Section 3.

Let $\bar{\delta h} = n^{-1} \sum_i \delta h_i$ denote the mean residual shift. Alignment accuracy is the fraction of negative-control prompts for which the model predicts A_i^+, and retention accuracy is the corresponding fraction on positive-control prompts. Rank concentration is $\rho_1 = \sigma_1^2 / \sum_j \sigma_j^2$, where σ_j are the singular values of ΔH; a value of $\rho_1 = 1$ means all per-topic shifts lie in a single direction. Steering alignment is $\cos(\delta h, u) = \langle \bar{\delta h}, u \rangle / (\|\bar{\delta h}\| \|u\|)$, reported for both u_{mean} and u_{probe}. Topic-wise effect cosine compares the per-topic logit-margin changes $z(A_i^+) - z(A_i^-)$ induced by steering and LoRA. The total LoRA update norm for a set of matrices $\mathcal{S}$ is $\|\Delta W\|_{\text{total}} = (\sum_{M \in \mathcal{S}} \|\Delta W_M\|_F^2)^{1/2}$.

For component replay, we report the activation norm ratio $\|\Delta H_{\text{component}}\|_F / \|\Delta H_{\text{all}}\|_F$, the parameter norm ratio $\|\Delta W_{\text{component}}\|_{\text{total}} / \|\Delta W_{\text{all}}\|_{\text{total}}$, the normalized interaction magnitude $\|R_{\text{int}}\|_F / \|\Delta H_{\text{all}}\|_F$, and the component cosine

$$\cos_{\text{VO,QK}} = \frac{\langle \text{vec}(\Delta H_{\text{VO|all}}), \text{vec}(\Delta H_{\text{QK|all}}) \rangle}{\|\Delta H_{\text{VO|all}}\|_F \, \|\Delta H_{\text{QK|all}}\|_F}.$$

5. Results & Discussion

We evaluate whether rank-one LoRA updates induce low-rank, steering-like residual shifts, and whether this depends on the attention pathway being updated.

Table 1. Standalone rank-one LoRA results at $d = 64$, $n = 24$, $\lambda = 0.1$, averaged over 30 seeds. ρ_1 is the rank-one energy fraction of the stacked residual-shift matrix. $\cos(\delta h, u_{\text{mean}})$ is the cosine between the mean LoRA-induced shift and the mean-difference steering vector. $\|\Delta W\|$ ratio normalizes each configuration's total update norm by that of the all-matrices update.

Target	Align	ρ_1	$\cos(\delta h, u_{\text{mean}})$	$\|\Delta W\|$	$p(A^+)$
W_O	1.000 ± 0.000	1.000 ± 0.000	0.714 ± 0.105	1.095 ± 0.164	0.984 ± 0.018
W_V	1.000 ± 0.000	1.000 ± 0.000	0.808 ± 0.125	0.958 ± 0.129	0.988 ± 0.017
W_Q	0.998 ± 0.009	0.867 ± 0.088	0.768 ± 0.231	6.602 ± 11.774	0.904 ± 0.070
W_K	1.000 ± 0.000	0.950 ± 0.065	0.832 ± 0.136	5.643 ± 12.961	0.926 ± 0.071
W_Q+W_K	1.000 ± 0.000	0.951 ± 0.032	0.839 ± 0.111	1.577 ± 0.867	0.909 ± 0.107
W_V+W_O	1.000 ± 0.000	0.998 ± 0.004	0.807 ± 0.122	0.919 ± 0.071	0.992 ± 0.009
All	1.000 ± 0.000	0.968 ± 0.027	0.859 ± 0.121	1.000 ± 0.000	0.988 ± 0.011

5.1. Pretraining Gives a Clean Intervention Baseline

Across all widths $d \in \{32, 64, 96, 128\}$, the pretrained model solves the balanced pretraining task (Table 3, Appendix B). Negative-control margins are consistently negative, ranging from approximately -44 to -60, which confirms that the alignment task requires changing the learned negative-control behavior. Activation steering with u_{mean} and u_{probe} improves target probability for intermediate steering strengths, peaking near $\alpha = 5$ to 10, while larger strengths degrade probability despite increasing margins (Appendix C).

5.2. VO LoRA Produces the Fixed-Write Geometry Predicted by Theory

Table 1 and Figure 1 report the primary standalone LoRA results at $d = 64$, $n = 24$, $\lambda = 0.1$, averaged over seeds. Rank-one LoRA on W_O and W_V gives $\rho_1 = 1.000 \pm 0.000$ across all seeds and all widths. This matches Theorem 3.1, since each induced residual shift is a scalar multiple of a fixed direction. Figure 2 confirms zero seed-level variance for both pathways. Thus, for individual VO updates, low rank in parameter space translates directly into low rank in

activation space.

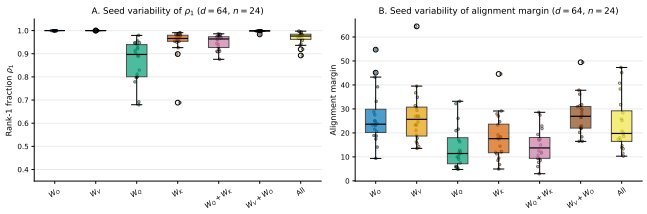

Figure 2. Seed-level variability at $d = 64$, $n = 24$, $\lambda = 0.1$ over 30 seeds. **(A)** ρ_1 distributions. W_O and W_V have zero variance; W_Q-only has the widest spread, with the interquartile range spanning approximately 0.80 to 0.93 and outliers below 0.70. **(B)** Alignment margin distributions. VO configurations produce consistent margins; QK configurations are more variable despite reaching comparable alignment accuracy.

Joint $W_V{+}W_O$ updates are only guaranteed to have rank at most two by Proposition 3.2, but empirically they are almost one-dimensional, with $\rho_1 = 0.998 \pm 0.004$ in the primary setting. The second singular component is around 10^{-3} of total energy for $W_V{+}W_O$, compared with approximately 10^{-2} for QK configurations (Figure 6, Appendix E). This suggests that optimization aligns or concentrates the two possible write directions.

Alignment with u_{mean} is positive but decreases with width (Figure 1B). The decrease is steeper for W_O, from 0.74 at $d = 32$ to 0.55 at $d = 128$, than for W_V, from 0.80 to about 0.73. This is consistent with the minimum-norm alignment argument in Section 3.3, where dispersion of the relevant features weakens pointwise agreement with the mean shift. The same qualitative pattern holds for u_{probe} (Appendix D).

All VO configurations reach alignment accuracy 1.000 in the primary setting. Thus, behavioral success alone does not distinguish an exactly one-dimensional residual shift from a nearly one-dimensional joint write.

5.3. QK Updates Implement Distribution-Dependent Routing

QK updates show the other side of the theory: they can solve the task, but low rank in the updated weights does not force a fixed residual direction. In the primary setting, W_Q gives $\rho_1 = 0.867 \pm 0.088$, while W_K and $W_Q{+}W_K$ give $\rho_1 = 0.950 \pm 0.065$ and $\rho_1 = 0.951 \pm 0.032$ (Table 1). The asymmetry between W_Q and W_K is empirical rather than algebraic, since both enter through $M = W_Q^\top W_K$.

Figure 2 shows that W_Q has the largest seed variability, with a range from approximately 0.80 to 0.93. This matches Corollary 3.4, since steering-like behavior depends on how collinear the context-dependent directions $d_{H^{(i)}}$ are across prompts. The topic-count sweep further supports this view, as W_Q-only ρ_1 decreases with more topics while VO remains fixed at 1.000 (Appendix H).

Table 2. Counterfactual component replay of learned all-matrices updates, aggregated over 481 runs across $d \in \{32, 64, 96, 128\}$, $n \in \{8, 24, 48\}$, and $\lambda \in \{0.1, 0.5\}$. Norm ratio is $\|\Delta H_{\text{component}}\|_F / \|\Delta H_{\text{full}}\|_F$. $\|\Delta W\|$ ratio normalizes each component's parameter norm by the full all-matrices norm.

Component	Norm ratio	ρ_1	Align acc	$\|\Delta W\|$ ratio
W_Q	0.272	0.832	0.062	0.456
W_K	0.450	0.954	0.276	0.534
$W_Q{+}W_K$	0.656	0.919	0.437	0.710
W_V	0.542	1.000	0.444	0.485
W_O	0.465	1.000	0.317	0.477
$W_V{+}W_O$	0.817	0.994	0.791	0.684
All	1.000	0.962	1.000	1.000

QK updates also require larger and more variable parameter changes. The update-norm ratios for W_Q and W_K are 6.60 ± 11.77 and 5.64 ± 12.96, compared with 1.10 and 0.96 for W_O and W_V. This efficiency gap persists across widths (Appendix F). Nevertheless, W_K and $W_Q{+}W_K$ reach perfect alignment accuracy, showing that routing can solve the task even when its residual shifts are less rank-concentrated.

5.4. All-Matrices SFT Coordinates Writing and Routing

The all-matrices update achieves perfect alignment accuracy and remains highly rank-concentrated, with $\rho_1 = 0.968 \pm 0.027$ in the primary setting. However, component replay shows that this is not a purely VO solution.

Table 2 reports counterfactual replay results averaged over all-matrices runs. Replaying only the learned VO factors recovers a large residual norm ratio, 0.817, and reaches alignment accuracy 0.791. Replaying only the learned QK factors gives a smaller but substantial norm ratio, 0.656, and alignment accuracy 0.437. The VO and QK components are positively aligned, with mean vectorized cosine approximately 0.58 at $d = 64$ and above 0.50 across all widths (Figure 11d in Appendix I). Thus, the optimizer coordinates writing and routing rather than assigning the task to one pathway.

Within the VO replay, W_O alone and W_V alone reach accuracies 0.317 and 0.444, while $W_V{+}W_O$ reaches 0.791. Since both individual replays maintain $\rho_1 = 1.000$, this non-additivity is behavioral rather than geometric. The QK-to-VO norm ratio also grows with width, from about 0.57 at $d = 32$ to 1.43 at $d = 128$ in some conditions (Appendix I). This rules out a simple VO-dominance interpretation of the all-matrices solution.

5.5. The Full All-Matrices Update Is Non-Additive

Figure 3 reports the normalized interaction residual

$$\varepsilon_{\text{int}} = \frac{\|\Delta H_{\text{all}} - \Delta H_{\text{VO|all}} - \Delta H_{\text{QK|all}}\|_F}{\|\Delta H_{\text{all}}\|_F}.$$

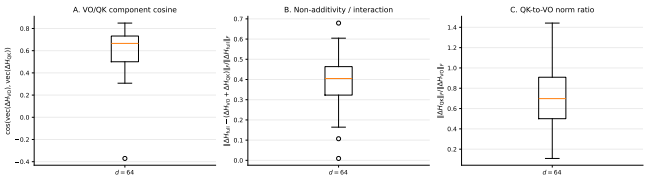

Figure 3. Pairwise diagnostics for counterfactual VO and QK replays at $d = 64$, $n = 24$, and $\lambda = 0.1$, computed over 30 seeds. **(A)** The VO and QK residual-shift matrices are positively aligned, with median vectorized cosine similarity around 0.65. **(B)** The interaction residual is large, with median normalized norm around 0.40, showing that the full update is not well approximated by the sum of independent VO and QK replays, consistent with Proposition 3.5. **(C)** The QK-to-VO norm ratio varies substantially across seeds, indicating that the optimizer balances routing and writing differently across runs.

It is approximately 0.24 at $d = 32$, 0.39 at $d = 64$, and 0.61 at $d = 96$ to 128 (Appendix I). This width-dependent growth shows that the all-matrices update is not well described by adding independent VO and QK replays. Instead, writing and routing interact through the terms identified in Proposition 3.5.

The full residual shift remains highly rank-concentrated despite this interaction, with $\rho_1 \approx 0.962$ averaged across all runs. The most consistent interpretation is that the learned solution combines a strong VO write, a positively aligned QK contribution, and interaction terms that largely preserve the dominant direction.

5.6. Behavioral Accuracy Hides Mechanistic Differences

All seven standalone LoRA configurations reach near-perfect alignment accuracy in the primary setting, but their mechanisms differ sharply. They vary in rank concentration from 0.867 to 1.000, in update norm from 0.92 to 6.60 relative to the all-matrices baseline, and in component replay accuracy from 0.062 to 1.000. The all-matrices solution reaches perfect accuracy although neither VO replay nor QK replay does so alone. These results show that behavioral accuracy must be supplemented with residual-geometry and component-level diagnostics.

5.7. Robustness

The qualitative separation between VO and QK is stable across the robustness sweeps. Varying λ from 0.1 to 0.5 leaves ρ_1 essentially flat. The rank hierarchy is preserved across $n \in \{8, 24, 48\}$, with W_O and W_V fixed at 1.000, W_Q lowest, and W_K intermediate. As n grows, W_Q-only ρ_1 decreases, consistent with greater diversity in routing directions (Appendix H). Across the full (d, n, λ) grid, VO replay consistently outperforms QK replay on alignment accuracy, while the interaction residual grows with width.

5.8. Discussion

The results support three conclusions. First, individual rank-one VO updates produce exactly rank-one residual-shift matrices in this setting, and gradient descent recovers this fixed-write structure. Second, QK updates solve the task through context-dependent routing, which makes their rank structure distribution-dependent and less stable, especially for W_Q. Third, when all matrices are trained jointly, the solution combines VO writing and QK routing in a positively aligned but non-additive way.

This distinction matters for interpreting finetuning through activation steering. If the behavioral change flows through VO, residual shifts can be compared to steering directions using ρ_1 and $\cos(\delta h, u_{\text{mean}})$. If it flows through QK, the shift is more context-dependent and less well summarized by a single vector. Component replay is thus needed to distinguish fixed-direction writing from routing-based change. This suggests why fixed-direction steering may work well for behaviors implemented as residual writes, but less reliably for behaviors requiring context-dependent routing.

The main limitation is the controlled linear-attention setting. We chose this setting because it allows exact characterization of rank structure, isolating the write-route asymmetry without softmax normalization, MLPs, layer normalization, or residual-stream normalization. Whether analogues of the exact VO rank result and the distributional QK picture persist in standard transformers remains an important direction.

The rank-one focus is a starting point. Higher-rank LoRA corresponds naturally to multi-direction steering, and Proposition 3.2 generalizes to rank r with a bound of at most $2r$, but higher-rank alignment behavior remains open. The binary task is also a controlled choice that gives a clean intervention baseline. Extending component replay to multi-behavior settings, where topics or domains may require different steering directions, remains future work.

6. Conclusion

We identified a structural asymmetry between the two attention pathways under low-rank finetuning. Rank-one updates to W_V and W_O are algebraically forced to produce one-dimensional residual shifts regardless of the data distribution, while updates to W_Q and W_K produce context-dependent geometry whose rank concentration depends on the input distribution. When all matrices are trained jointly, component replay shows that QK routing contributes meaningfully and interacts non-additively with VO writing. Together, these findings suggest that fixed-direction steering is most reliable for behaviors implementable as residual writes, and less reliable for behaviors requiring context-dependent routing.

References

Adila, D., Cooper, J., Yun, A., Trost, A., and Sala, F. Weight updates as activation shifts: A principled framework for steering. *arXiv preprint arXiv:2603.00425*, 2026.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., and Wei, F. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019, 2023.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022a.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Drain, T., et al. Superposition, memorization, and double descent. *Transformer Circuits Thread*, 2022b.

Fierro, C. and Roger, F. Steering language models with weight arithmetic. *arXiv preprint arXiv:2511.05408*, 2025.

Hedström, A., Amoukou, S. I., Bewley, T., Mishra, S., and Veloso, M. To steer or not to steer? mechanistic error reduction with abstention for language models. *arXiv preprint arXiv:2510.13290*, 2025.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Li, C., Farkhoor, H., Liu, R., and Yosinski, J. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022a.

Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., and Bau, D. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Ortiz-Jimenez, G., Favero, A., and Frossard, P. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36:66727–66754, 2023.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Ren, Y. and Sutherland, D. J. Learning dynamics of llm finetuning. *arXiv preprint arXiv:2407.10490*, 2024.

Stolfo, A., Balachandran, V., Yousefi, S., Horvitz, E., and Nushi, B. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations*, volume 2025, pp. 55790–55823, 2025.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. 2024.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022.

Xu, Z., Wu, C., Sun, H., Hong, H., Wang, M., Yao, Y., Huang, L., Xue, H., Deng, S., Chu, Z., et al. Why steering works: Toward a unified view of language model parameter dynamics. *arXiv preprint arXiv:2602.02343*, 2026.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Extended Related Work

Our work sits at the intersection of four lines of research: activation steering, transformer representation geometry, low-rank finetuning, and mechanistic interpretability. The common thread is that behavioral change in language models can be described as movement along directions in either activation space or weight space. Our contribution is to identify precisely when these two descriptions coincide and when they diverge.

Activation steering and representation engineering. Turner et al. (2024), Panickssery et al. (2023), Li et al. (2023), and Zou et al. (2023) show that many behavioral properties can be manipulated through approximately linear directions in activation space. Stolfo et al. (2025) demonstrate that such directions transfer across models and compose across multiple instructions simultaneously. Marks & Tegmark (2023) find similar linear structure for truthfulness. Hedström et al. (2025) study when steering vectors become brittle under distribution shift. These works establish the empirical basis for fixed-direction steering, but do not characterize when finetuning produces shifts of the same form. Our paper provides that characterization for the VO pathway and shows it fails for QK.

Linear representation hypothesis and activation geometry. Park et al. (2023) formalize the idea that meaningful features correspond to linear directions and study the geometry of interventions. Elhage et al. (2022a) and Elhage et al. (2022b) show that superposition can entangle these directions, and sparse autoencoder work recovers interpretable features from residual stream activations (Bricken et al., 2023; Templeton et al., 2024). Our paper complements this picture by asking not what directions exist in activation space, but when a weight update is structurally forced to write along a single fixed direction.

Low-rank adaptation and parameter efficient finetuning. Hu et al. (2022) introduce LoRA, constraining weight updates to low-rank factorizations, and broader evidence suggests that finetuning dynamics often have low-dimensional structure (Arora et al., 2019; Li et al., 2018; Ren & Sutherland, 2024). Ilharco et al. (2022) and Ortiz-Jimenez et al. (2023) show that weight updates have meaningful directional geometry through task arithmetic, and Wortsman et al. (2022) exploit this structure via model soups. Our work asks a complementary question: when does low rank in weight space imply low rank in activation space? The answer is pathway-dependent and not guaranteed by the rank of the weight update alone.

Model editing and rank-one weight interventions. Meng et al. (2022a) introduce ROME, showing that rank-one weight edits to MLP layers can produce targeted behavioral changes, and Meng et al. (2022b) extend this to many edits via MEMIT. Our work studies the analogous question for attention projections under LoRA, and finds that the answer depends on whether the updated matrices belong to the write pathway or the routing pathway.

Unifying steering and finetuning. Dai et al. (2023) and Von Oswald et al. (2023) connect in-context learning to implicit gradient descent. Adila et al. (2026) and Xu et al. (2026) propose formal mappings between weight updates and activation-level dynamics. Fierro & Roger (2025) construct behavioral directions directly in weight space. Our work complements these by providing an exact algebraic characterization of when such mappings produce fixed-direction residual shifts, identifying the VO versus QK asymmetry as the key structural condition.

Mechanistic interpretability and transformer circuits. Elhage et al. (2021) decompose attention heads into QK and OV circuits, distinguishing routing from writing. Olsson et al. (2022) use this framework to study induction heads in linear attention. Lindsey et al. (2025) trace behaviors through specific circuits in large models. Our theoretical results give this routing-versus-writing distinction an explicit algebraic meaning under low-rank finetuning: VO updates are algebraically forced to produce fixed-direction shifts, while QK updates are not.

Linear attention as a theoretical substrate. Katharopoulos et al. (2020) introduce linear attention as an efficient alternative to softmax attention. Von Oswald et al. (2023) and Olsson et al. (2022) use linear attention as a controlled substrate for studying transformer computations exactly. We follow this approach, using linear attention because it is bilinear in the relevant quantities and admits exact rank characterization of residual shifts, with softmax attention as the primary direction for future work.

B. Pretraining and Intervention Baselines

Table 3 reports pretraining and pre-intervention statistics for all model widths. In every configuration the pretrained model achieves perfect accuracy on the balanced pretraining task (1.000 for both positive- and negative-control conditions) and zero accuracy on the alignment set before any intervention. Pre-intervention alignment margins are consistently large and negative (approximately -44 to -60 depending on width), confirming that the alignment objective requires genuine behavioral change.

Table 3. Pretraining and pre-intervention statistics, mean $\pm$ std over 30 seeds per width.

Metric	$d = 32$	$d = 64$	$d = 96$	$d = 128$
Pretrain acc	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Align acc (before)	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Align margin (before)	-54.2 ± 30.4	-45.6 ± 21.4	-60.3 ± 37.5	-44.0 ± 18.8
Retain-pos acc	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Neg-original acc	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

C. Steering Sweep Results

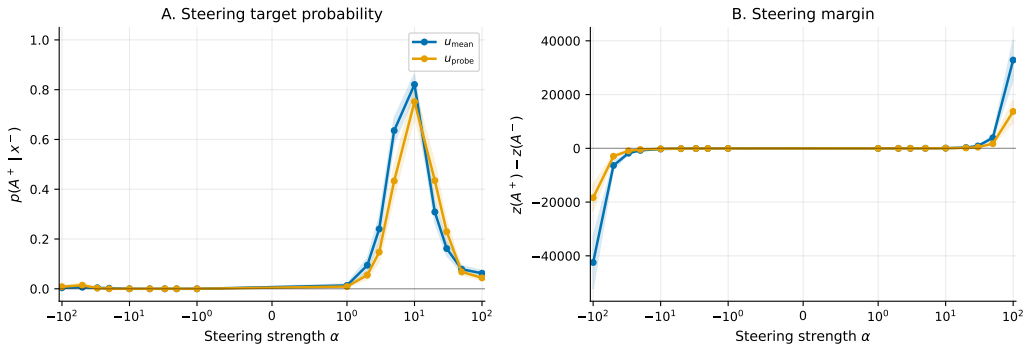

Figure 4. Activation steering sweep at $d = 64$, $n = 24$, over 30 seeds. **(A)** Target probability $p(A^+ \mid x^-)$ as a function of steering strength α for both u_{mean} and u_{probe}. Both directions produce nearly identical curves, peaking near $\alpha = 5$–10. **(B)** Logit margin $z(A^+) - z(A^-)$. Large positive α produces extreme margins but the target probability collapses, reflecting a misalignment between the steering direction and the unembedding geometry at large amplitudes. Shaded regions denote standard error across seeds.

Figure 4 shows the alignment target probability and logit margin as a function of steering strength α for both u_{mean} and u_{probe}. Both steering directions produce nearly identical sweep curves, consistent with their shared origin in the residual-stream geometry described in Section 3.1. The target probability peaks near $\alpha = 5$–10 before collapsing at larger strengths, where the steering vector overwhelms the unembedding geometry. This topic-level sweep (Figure 4) confirms that this pattern holds across individual topics rather than being an artifact of averaging.

D. Probe Cosine Scaling

Figure 5 reports cosine alignment between the mean SFT-induced shift and u_{probe} across widths. The pattern closely mirrors the u_{mean} alignment shown in Figure 1B: VO pathways have moderate and slightly declining alignment, W_Q is consistently the lowest, and the all-matrices and $W_V{+}W_O$ configurations are among the most stable. The near-equivalence of u_{mean} and u_{probe} alignment is consistent with the shared geometric origin of the two vectors in this synthetic task.

E. Singular Spectra

Figure 6 shows the energy fractions $\sigma_j^2 / \sum_k \sigma_k^2$ for the first three singular values of the stacked residual-shift matrix. For W_O and W_V, the second and third singular values are exactly zero (the distributions are point masses at $\sigma_1^2 / \sum = 1.000$). For $W_V{+}W_O$, $\sigma_2^2 / \sum$ is concentrated near 10^{-3}, confirming empirical near-rank-1 structure within the theoretical rank-2 bound. QK configurations have second singular values near 10^{-2}, roughly an order of magnitude larger, and broader

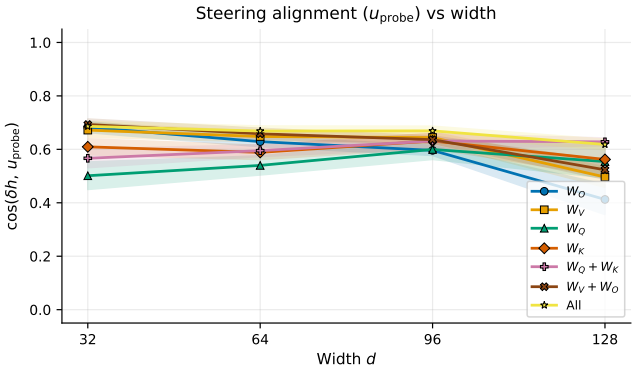

Figure 5. Cosine alignment between the mean SFT-induced residual shift and u_{probe} as a function of model width, averaged over 30 seeds at $n = 24$, $\lambda = 0.1$. The pattern mirrors the u_{mean} alignment shown in Figure 1B, with W_O declining most steeply and VO configurations generally more aligned than QK at small widths. Shaded regions denote standard error across seeds.

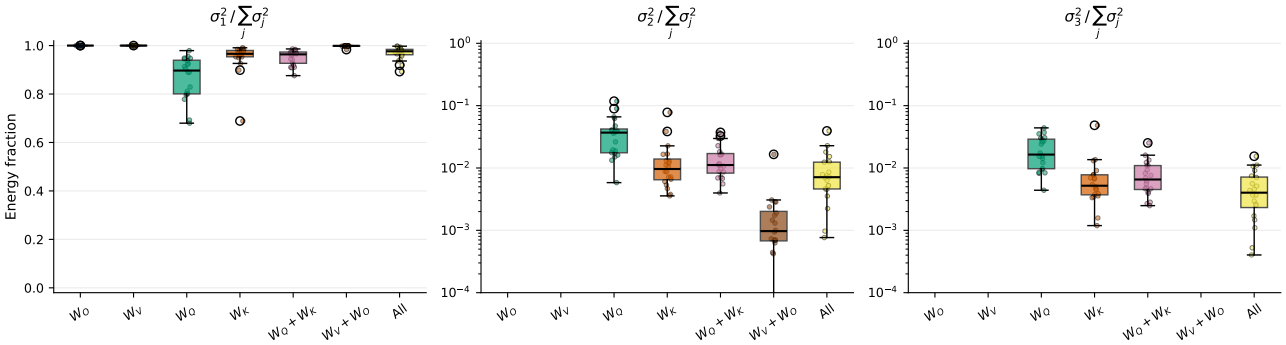

Figure 6. Singular value energy fractions $\sigma_j^2 / \sum_k \sigma_k^2$ for $j \in \{1, 2, 3\}$ of the stacked residual-shift matrix, at $d = 64$, $n = 24$, $\lambda = 0.1$, over 30 seeds. **Left**: σ_1^2 fraction. W_O and W_V are point masses at 1.000. **Center and Right**: σ_2^2 and σ_3^2 fractions on a log scale. $W_V{+}W_O$ second-singular energy is near 10^{-3}; QK configurations have second-singular energy near 10^{-2}, approximately one order of magnitude larger.

distributions reflecting the seed-level variability identified in Section 5.3.

F. Standalone Update Norm Scaling

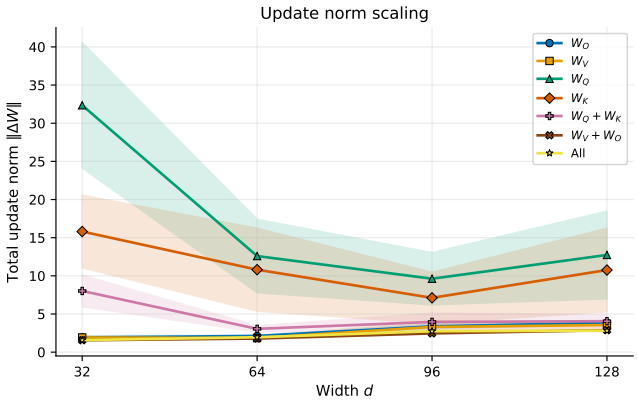

Figure 7. Total LoRA update norm $\|\Delta W\|_F$ as a function of model width for standalone configurations, averaged over 30 seeds at $n = 24$, $\lambda = 0.1$. W_Q-only has the largest and most variable norm at $d = 32$ (32.4 ± 36.8), with the standard deviation exceeding the mean; W_K-only shows a similar but less extreme pattern. VO configurations and the all-matrices configuration maintain compact norms across all widths. Note that these standalone norms differ from the within-all-matrices component norms in Table 2, where QK and VO parameter norms are comparable. Shaded regions denote standard error across seeds.

Figure 7 reports total update norms as a function of model width for standalone LoRA configurations. The W_Q-only condition has the largest and most variable update norm at $d = 32$ (32.4 ± 36.8), with the standard deviation exceeding the mean, indicating bimodal behavior: some seeds find compact QK solutions while others require very large weight changes. W_K-only shows a similar but less extreme pattern. VO configurations (W_O, W_V, W_V+W_O) and the all-matrices configuration maintain compact update norms across all widths. Note that these standalone norms differ from the within-all-matrices component norms reported in Table 2: when all matrices are jointly optimized, QK and VO parameter norms become comparable, with the learned QK components growing toward the VO norm ratio rather than the extreme standalone values.

G. Per-Width Stability Boxplots

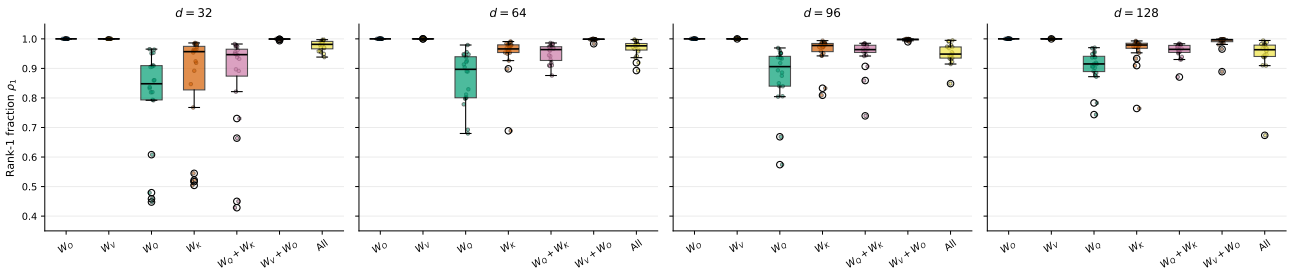

Figure 8. Seed-level ρ_1 distributions across model widths $d \in \{32, 64, 96, 128\}$ at $n = 24$, $\lambda = 0.1$, over 30 seeds. W_O and W_V maintain $\rho_1 = 1.000$ with zero variance at every width. W_Q-only remains the most variable pathway across all widths, though the IQR narrows at $d = 96$–128. Individual points show all 30 seeds.

Figures 8 and 9 extend the seed-variability analysis of Figure 2 across all widths. The rank-1 fraction boxplots confirm that W_O and W_V remain pinned at 1.000 with no variance at every width, while W_Q-only remains the most variable and has the lowest median ρ_1. W_K and W_Q+W_K narrow their IQR at larger widths but retain non-trivial spread. Alignment margin boxplots show that VO configurations produce more consistent margins than QK configurations, though the all-matrices condition produces margins competitive with W_V+W_O in most settings.

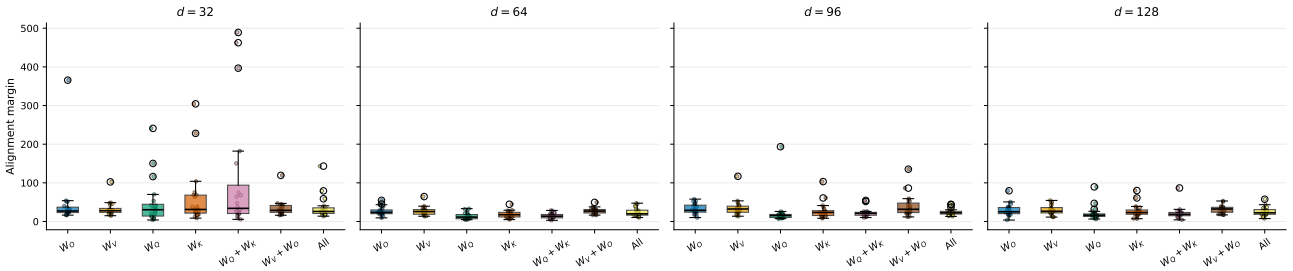

Figure 9. Seed-level alignment margin distributions across model widths $d \in \{32, 64, 96, 128\}$ at $n = 24$, $\lambda = 0.1$, over 30 seeds. Margins compress at larger widths for most configurations. $W_Q{+}W_K$ shows the widest margin distribution at $d = 32$ (median approximately 80, with outliers above 400), reflecting the bimodal optimization behavior identified in Appendix F. Individual points show all 30 seeds.

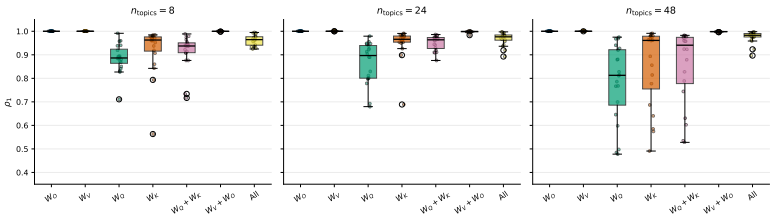

Figure 10. ρ_1 distributions for $n_{\text{topics}} \in \{8, 24, 48\}$ at $d = 64$, $\lambda = 0.1$, over 30 seeds. W_O and W_V remain at $\rho_1 = 1.000$ across all topic counts. W_Q-only median ρ_1 decreases monotonically from approximately 0.88 at $n = 8$ to 0.82 at $n = 48$, consistent with Corollary 3.4: more topics introduce more diverse prompts, increasing the variation in context-dependent routing directions. Individual points show all 30 seeds.

H. Robustness to Topic Count

Figure 10 shows ρ_1 across configurations for $n_{\text{topics}} \in \{8, 24, 48\}$ at $d = 64$. W_O and W_V remain at $\rho_1 = 1.000$ across all topic counts. W_Q-only ρ_1 decreases monotonically as n grows (median approximately 0.88 at $n = 8$, falling to approximately 0.82 at $n = 48$), consistent with Corollary 3.4: more topics introduce more diverse prompts, increasing the variation in context-dependent directions $d_{H^{(i)}}$ and raising the rank of D_a. W_K and $W_Q{+}W_K$ also widen their interquartile range with increasing topic count, though less dramatically than W_Q-only.

I. All-Matrices Component Decomposition: Scaling Results

Figures 11a–11e extend the component replay analysis of Section 5.4 across widths.

Figure 11a shows counterfactual component norm ratios as a function of width. The $W_V{+}W_O$-only ratio starts at approximately 0.85 at $d = 32$ and declines toward 0.71 at $d = 128$, while the $W_Q{+}W_K$-only ratio rises from approximately 0.58 at $d = 32$ to 0.89 at $d = 128$. At $d \geq 96$ the QK component contributes more residual norm than the individual VO component in many seeds, which directly contradicts any simple VO-dominance interpretation of the all-matrices result. Figure 11c shows counterfactual alignment accuracy by width. VO-only replay accuracy declines from approximately 0.76 at $d = 32$ to approximately 0.64 at $d = 128$, while QK-only replay accuracy increases from approximately 0.44 to 0.68. The two curves converge at large widths, suggesting that wider models distribute behavioral effect more evenly between pathways. Individual W_O-only replay accuracy is near zero at $d = 32$ and rises slowly, while W_V-only is moderate and roughly stable; this asymmetry reflects that in the all-matrices solution, b_O is learned in a direction that complements $W_O b_V$ rather than independently carrying the behavioral signal. Figure 11d shows the VO/QK component cosine across widths. The mean cosine is approximately 0.56 at $d = 32$, rises to 0.69 at $d = 96$, and remains above 0.55 at $d = 128$. This persistent positive alignment holds across seeds and widths, indicating that the optimizer consistently produces QK routing that reinforces rather than opposes the VO write direction. Figure 11e shows the interaction residual ε_{int} as a function of width. The mean grows from 0.24 at $d = 32$ to 0.61 at $d = 96$, then plateaus near 0.59 at $d = 128$. This growth pattern suggests that wider models exploit more of the cross-pathway interaction capacity of the architecture, consistent with Proposition 3.5: larger ΔW_O and ΔW_V amplify the interaction terms involving ΔA.

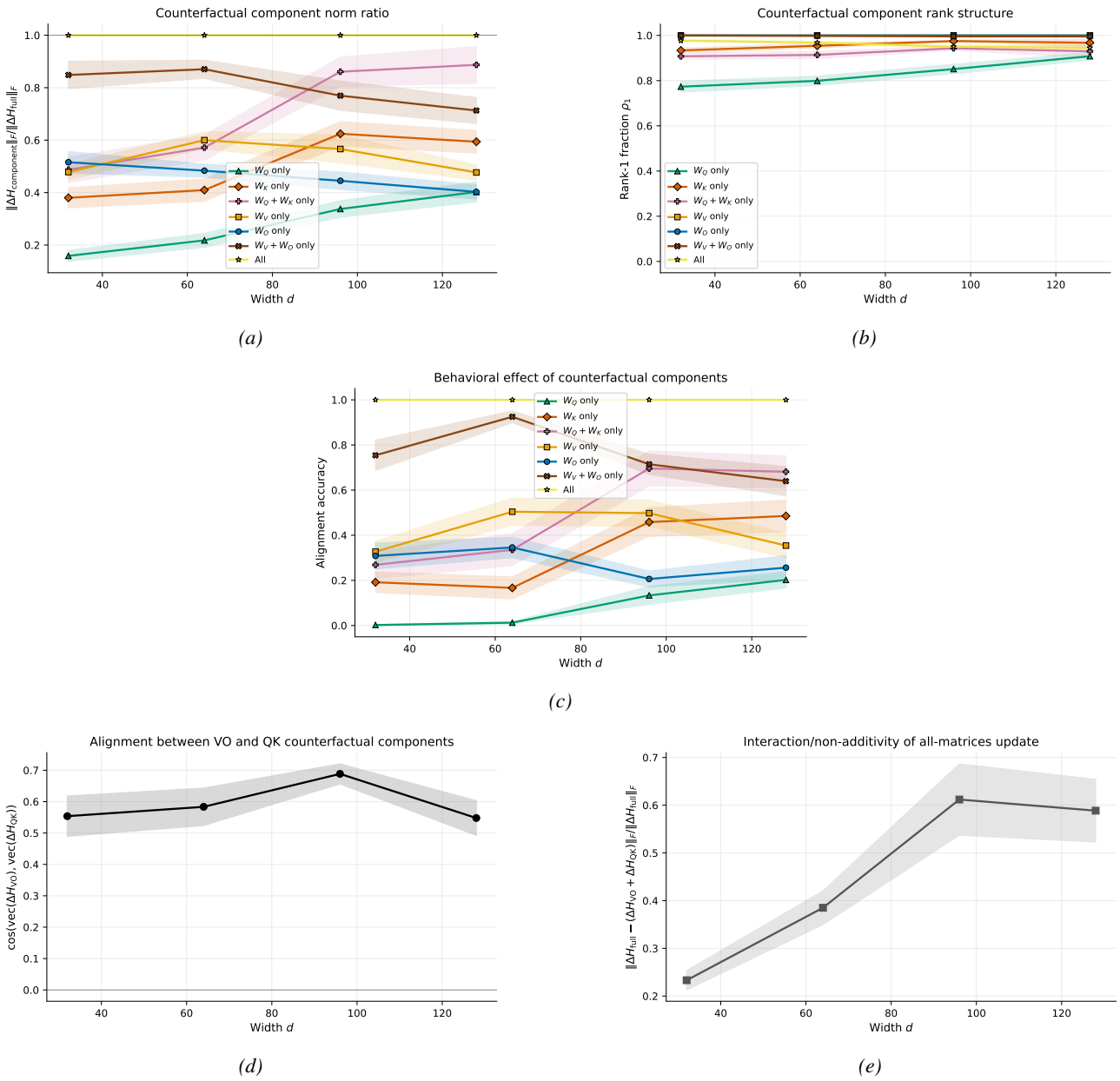

Figure 11. All-matrices component decomposition results as a function of model width $d \in \{32, 64, 96, 128\}$, averaged over seeds at $n = 24$, $\lambda = 0.1$. Shaded regions denote standard error across seeds. **(a)** Counterfactual component norm ratio $\|\Delta H_{\text{component}}\|_F / \|\Delta H_{\text{full}}\|_F$. The $W_V{+}W_O$-only ratio declines from approximately 0.85 at $d = 32$ to 0.71 at $d = 128$, while the $W_Q{+}W_K$-only ratio rises from 0.58 to 0.89; at $d \geq 96$ the QK component contributes more residual norm than VO in many seeds. **(b)** Counterfactual component rank-1 fraction ρ_1. W_V-only and W_O-only components maintain $\rho_1 = 1.000$ across all widths by Theorem 3.1; W_Q-only has the lowest and most variable ρ_1. **(c)** Alignment accuracy of counterfactual component replays. VO-only accuracy declines from 0.76 at $d = 32$ to 0.64 at $d = 128$; QK-only accuracy increases from 0.44 to 0.68, with the two curves converging at large widths. W_O-only replay accuracy is near zero at $d = 32$ and rises slowly, reflecting the asymmetric role of the two write matrices in the learned all-matrices solution. **(d)** Cosine similarity between the vectorized VO and QK component residual-shift matrices. The mean cosine remains between 0.55 and 0.70 across all widths, indicating persistent positive alignment between pathways. **(e)** Interaction residual $\|\Delta H_{\text{all}} - (\Delta H_{\text{VO}} + \Delta H_{\text{QK}})\|_F / \|\Delta H_{\text{all}}\|_F$. The mean grows from 0.24 at $d = 32$ to 0.61 at $d = 96$ then plateaus near 0.59 at $d = 128$, confirming that the full update is increasingly non-additive at larger widths, consistent with Proposition 3.5.