

PRISM- Δ : Differential Subspace Steering for Prompt Highlighting in Large Language Models

Yuyao Ge^{1,2,3} Shenghua Liu^{1,2,3*} Yiwei Wang⁴ Baolong Bi^{1,2,3}
 Lingrui Mei^{1,2,3} Jiayu Yao^{1,2,3} Tianyu Liu³ Jiafeng Guo^{1,2,3} Xueqi Cheng^{1,2,3}

¹State Key Laboratory of AI Safety

²Institute of Computing Technology, Chinese Academy of Sciences

³University of Chinese Academy of Sciences ⁴University of California, Merced
 {geyuyao24z, liushenghua}@ict.ac.cn

 <https://github.com/YuyaoGe/PRISM-DELTA>

Abstract

Prompt highlighting steers a large language model to prioritize user-specified text spans during generation. A key challenge of existing Key-editing approaches is extracting steering directions that capture the difference between relevant and irrelevant contexts, rather than shared structural patterns common to both. We propose PRISM- Δ (Projection-based Relevance-Informed Steering Method), which decomposes the difference between positive and negative cross-covariance matrices to maximize discriminative energy while eliminating shared directions. Each attention head receives a continuous softplus importance weight, letting weak-but-useful heads contribute at reduced strength. The framework extends naturally to Value representations, capturing content-channel signal that Key-only methods leave unused. Across four benchmarks and five models, our methods match or exceed the best existing method on 19 of 20 configurations, with relative gains up to +13.0%, while halving the fluency cost of steering. Our methods also scale to long-context retrieval, outperforming the best existing method by up to +5.1% relative gain. PRISM- Δ is compatible with FlashAttention and adds negligible memory overhead.

1 Introduction

Large language models (LLMs) frequently need to prioritize specific parts of their input. When presented with conflicting information, the model should attend to newly provided facts over its parametric memory. In long-context retrieval, the answer may reside in the middle of thirty passages, where models notoriously underperform (Liu et al., 2024). This problem is known as *prompt highlighting* (Li et al., 2026): given a prompt and a marked subset of tokens, the goal is to amplify the model’s attention to those tokens so that generation becomes more accurate and faithful to user intent.

Several prompt highlighting methods have been proposed, including post-hoc attention score manipulation (Zhang et al., 2024), logit-level anchoring (Tian & Zhang, 2025), and pre-attention Key vector editing via spectral decomposition (Li et al., 2026). However, all of them operate solely on the *routing channel*—Key representations that control where the model looks. Transformer attention output also depends on a *content channel*—Value representations that determine what information is transmitted. Even when routing successfully directs the model to attend to highlighted tokens, the information those tokens convey through their Value representations remains unenhanced.

We investigate whether the Value channel carries useful signal by extracting Key and Value representations under contrastive conditions. Value shifts are comparable in magnitude to

*Corresponding author.

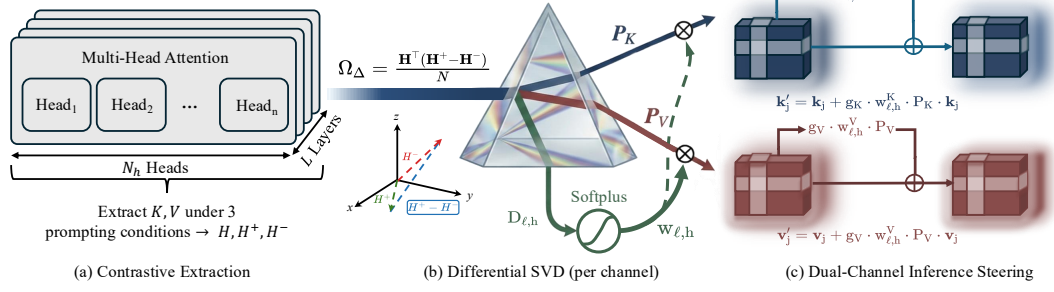


Figure 1: Overview of PRISM- Δ . SVD decomposes Ω_Δ into per-head projections (P_K, P_V) and importance weights ($w_{\ell,h}$), steering both Key and Value channels at inference.

Key across all five tested models, with roughly half of all heads showing significant Value-channel signal. Key-only methods thus leave substantial signal unused.

We propose PRISM- Δ (Projection-based Relevance-Informed Steering Method), which steers both channels through discriminative subspace learning and adaptive head weighting, as illustrated in Figure 1. Our contributions are:

- We introduce *differential cross-covariance decomposition* as a provably optimal criterion for extracting discriminative steering directions (Proposition 1): it maximizes the separation between relevant and irrelevant conditions while automatically eliminating shared structural directions that prior methods conflate with signal. Per-head softplus weighting enables continuous, rather than binary, importance assignment.
- We identify the *content channel* (Value representations) as a fundamentally untapped source of steering signal, and introduce the first method to jointly steer both routing and content channels. Value steering reduces fluency degradation and can further improve accuracy.

2 Related work

Prompt highlighting. PASTA (Zhang et al., 2024) modifies attention scores but is incompatible with FlashAttention (Dao et al., 2022), while SPA (Tian & Zhang, 2025) anchors at the logit level but requires multiple forward passes. SEKA (Li et al., 2026) edits Key vectors via spectral decomposition with near-zero overhead. Prefix-Tuning (Li & Liang, 2021) operates in Key/Value space by prepending learned soft tokens, but requires gradient-based training and modifies the context length. Venkateswaran & Contractor (2026) dynamically steer attention toward user-specified prompt parts for instruction following, and Zhan et al. (2026) score head relevance with a vector-quantized autoencoder for head selection. These methods either operate only on the routing channel, require offline training, or lack principled discriminative subspace learning. PRISM- Δ addresses all three gaps: it jointly steers Key and Value channels with per-head continuous softplus weighting, is gradient-free, and edits existing token representations in place.

Activation editing and head specialization. Among activation editing approaches, SEA (Qiu et al., 2024) and ITI (Li et al., 2023) project activations along contrastive directions, Activation Addition (Turner et al., 2023) and Stolfo et al. (2025) steer via natural-language-derived or instruction-specific vectors, and Representation Engineering (Zou et al., 2025), Subramani et al. (2022), and CARVE (Ge et al., 2025) extract steering representations at the population level—all operating on the residual stream without distinguishing Key and Value roles. In knowledge editing, ROME (Meng et al., 2022), AlphaEdit (Fang et al., 2025), and Hernandez et al. (2024) target factual associations via feed-forward weight updates or learned fact encodings, while attention analyses (Clark et al., 2019; Voita et al., 2019; Michel et al., 2019) and mechanistic studies on transformer circuits (Elhage et al., 2021), induction

Method	Steering target	Projection	Head selection	FlashAttn
Prefix-Tuning	Key + Value (prefix)	Gradient	None	✓
PASTA	Attn. matrix α	None	Profiling	✗
SPA	Logit distribution	None	None	✓
Spotlight	Attn. logits (bias)	Analytic	None	✗
REAL	Head activation	Codebook	Scoring top-S	✓
SEKA	Key	Indep. SVD	Hard threshold	✓
AdaSEKA	Key (multi-expert)	Indep. SVD	Hard threshold	✓
PRISM-Δ	Key	Diff. SVD	Softplus	✓
PRISM-ΔV	Key + Value	Diff. SVD	Softplus	✓

Table 1: Comparison of prompt highlighting methods.

heads (Olsson et al., 2022), and retrieval heads (Wu et al., 2025) reveal head-level functional specialization that motivates per-head importance weighting. PRISM- Δ uses differential SVD to extract discriminative directions and softplus to assign continuous weights, as detailed in Table 1.

3 Method

3.1 Dual-channel view of attention

In standard multi-head attention, given input hidden states $\mathbf{x}_i$, each head (ℓ, h) computes:

$$\mathbf{q}_i = W_Q^{(\ell, h)} \mathbf{x}_i, \quad \mathbf{k}_j = W_K^{(\ell, h)} \mathbf{x}_j, \quad \mathbf{v}_j = W_V^{(\ell, h)} \mathbf{x}_j \quad (1)$$

We drop the (ℓ, h) superscripts in subsequent equations for readability.

$$\text{output}_i = \sum_j \underbrace{\text{softmax}\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}}\right)}_{\alpha_{ij}: \text{routing}} \cdot \underbrace{\mathbf{v}_j}_{\text{content}} \quad (2)$$

Problem definition. Given a prompt $\mathbf{x} = (x_1, \dots, x_T)$ and a set of highlighted tokens $\mathcal{S} \subset \{1, \dots, T\}$, the goal is to amplify the influence of tokens $j \in \mathcal{S}$ in the attention output.

Dual-channel decomposition. The attention output is jointly determined by two functionally distinct channels: the *routing channel* ($\mathbf{K} \rightarrow \alpha$: determines *where* to look) and the *content channel* ($\mathbf{V}$: determines *what* information is transmitted). If we simultaneously perturb both channels for highlighted tokens:

$$\begin{aligned} \text{output}'_i &= \sum_j (\alpha_{ij} + \Delta\alpha_{ij}) \cdot (\mathbf{v}_j + \Delta\mathbf{v}_j) \\ &= \text{output}_i + \underbrace{\sum_j \Delta\alpha_{ij} \cdot \mathbf{v}_j}_{\text{routing gain}} + \underbrace{\sum_j \alpha_{ij} \cdot \Delta\mathbf{v}_j}_{\text{content gain}} + \underbrace{\sum_j \Delta\alpha_{ij} \cdot \Delta\mathbf{v}_j}_{\text{cross gain}} \end{aligned} \quad (3)$$

Existing prompt highlighting methods capture only the routing gain; the content gain and cross gain remain unused. Figure 2 confirms that Key and Value carry complementary signals, with their strength peaking at different network depths.

3.2 Discriminative subspace learning

Contrastive data and representation extraction. We construct synthetic QA triplets to identify contrastive directions. For each text context, we extract representations (Key or Value) at the answer token position under three conditions: $\mathbf{H}$ (neutral, context only), $\mathbf{H}^+$ (positive, context + relevant question), and $\mathbf{H}^-$ (negative, context + irrelevant question).

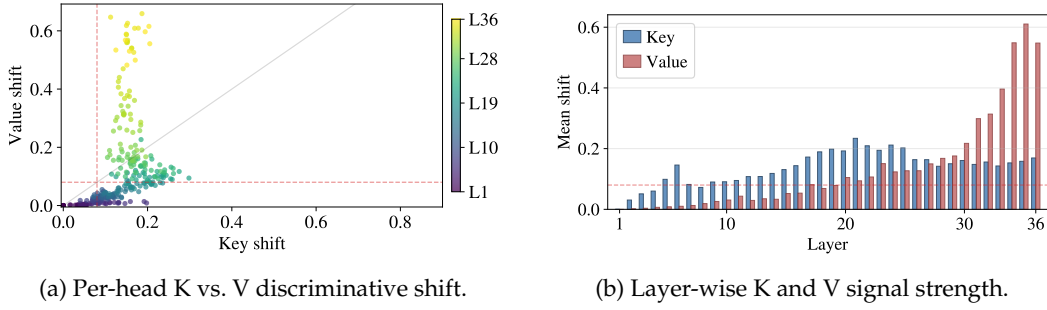


Figure 2: Dual-channel discriminative signals in QWEN3-4B-Base. **(a)** Each point is one attention head; Key and Value shifts are weakly correlated ($r=0.342$), confirming that the two channels carry complementary information. **(b)** Key signal peaks in middle layers (L13–24), while Value signal peaks in late layers (L25–36), suggesting functional specialization across depth.

Differential cross-covariance. We define the uncentered cross-covariance $\Omega^+ = \mathbf{H}^\top \mathbf{H}^+ / N$. Its top singular directions may include *shared* directions that co-vary equally under both conditions. To isolate truly discriminative directions, we decompose the *differential* cross-covariance:

$$\Omega_\Delta = \mathbf{H}^\top (\mathbf{H}^+ - \mathbf{H}^-) / N = \Omega^+ - \Omega^- \quad (4)$$

Proposition 1 (Discriminative optimality of differential directions). *Let $\Omega_\Delta = U_\Delta \Sigma_\Delta V_\Delta^\top$ be the SVD of the differential cross-covariance. Then:*

- (a) **Maximum discriminative energy.** *The top- k left singular vectors $\{u_1, \dots, u_k\}$ solve $\max_{U \in \mathbb{R}^{d \times k}, U^\top U = I} \|U^\top \Omega_\Delta\|_F^2$, i.e., they capture the k -dimensional subspace that maximizes the cross-covariance difference between positive and negative conditions (by the Eckart–Young theorem).*
- (b) **Automatic elimination of shared directions.** *If a direction $\mathbf{u}_s$ satisfies $(\Omega^+)^{\top} \mathbf{u}_s = (\Omega^-)^{\top} \mathbf{u}_s$, then $\Omega_\Delta^{\top} \mathbf{u}_s = \mathbf{0}$: shared left-space directions contribute zero to the differential projection, regardless of their energy in Ω^+ .*

Proposition 1 requires no distributional assumptions; it holds for any finite sample set (proof in Appendix B). Part (a) directly justifies the choice of differential SVD over independent SVD, while part (b) formally explains why shared variance directions are absent from the learned projection.

For comparison, let $P^+ = U_{:,k}^+ (U_{:,k}^+)^{\top}$ and $P^- = U_{:,k}^- (U_{:,k}^-)^{\top}$ denote independent SVD projections of Ω^+ and Ω^- , using the same energy threshold γ for rank selection. Figure 3 visualizes the structure of these matrices for a representative head, showing that independent projections P^+ and P^- share overlapping column spaces (structural redundancy), while P_Δ directly targets the differential subspace.

The top- k left singular vectors form the projection matrix:

$$P = U_\Delta[:, :k] \cdot U_\Delta[:, :k]^\top \quad (5)$$

where k is chosen so that the cumulative energy ratio reaches a threshold γ : $\sum_{i=1}^k \sigma_i^2 / \sum_{i=1}^d \sigma_i^2 \geq \gamma$.

Head importance weighting. Different attention heads vary widely in their sensitivity to prompt highlighting. The extracted representations also yield a per-head discriminability measure—the norm difference $D_{\ell,h} = \frac{1}{N} \sum_i \|\mathbf{r}_i^+ - \mathbf{r}_i^-\|_2$ (where $\mathbf{r}$ denotes the head’s Key or Value representation)—which quantifies how much the head’s output shifts between

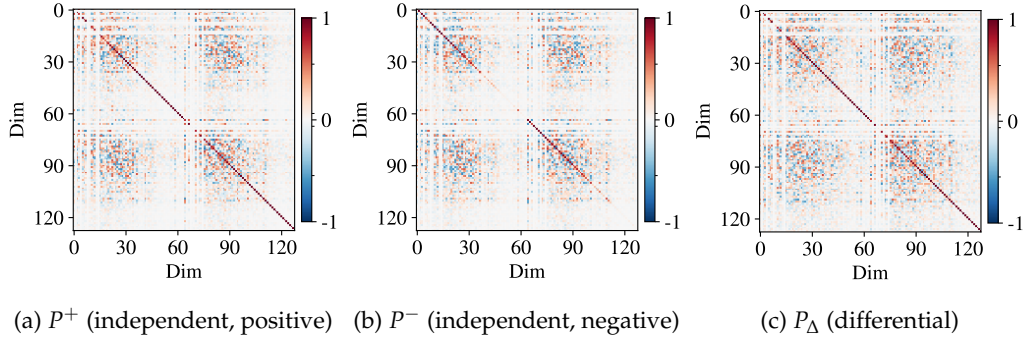


Figure 3: Projection matrix structure for layer 21, head 4 of QWEN3-4B with $d=128$. Independent projections P^+ and P^- exhibit overlapping subspaces, while the differential projection P_Δ directly targets the discriminative subspace.

Algorithm 1 PRISM- Δ : offline projection learning and online steering

- 1: **Input:** Model $\mathcal{M}$ with L layers and n_h heads; contrastive triplets $\{(c_i, q_i^+, q_i^-)\}_{i=1}^N$; $\gamma, \delta_{\min}, g_K, g_V$
 - 2: **for** each layer ℓ and head h **do**
 - 3: Extract $\mathbf{H}, \mathbf{H}^+, \mathbf{H}^- \in \mathbb{R}^{N \times d}$ under neutral/positive/negative conditions
 - 4: $\Omega_\Delta^K \leftarrow \mathbf{H}^\top (\mathbf{H}^+ - \mathbf{H}^-) / N$; $D_{\ell,h}^K \leftarrow \frac{1}{N} \sum_i \|\mathbf{H}_{i,:}^+ - \mathbf{H}_{i,:}^-\|_2$
 - 5: Compute SVD and retain top- k vectors with cumulative energy $\geq \gamma$
 - 6: $P_K \leftarrow U_{:,k} U_{:,k}^\top$; $w_{\ell,h}^K \leftarrow \text{softplus}(D_{\ell,h}^K - \delta_{\min})$
 - 7: Repeat for Value representations to obtain $P_V, w_{\ell,h}^V$
 - 8: **end for**
 - 9: **for** each highlighted token $j \in \mathcal{S}$ at each layer and head **do**
 - 10: $\mathbf{k}'_j \leftarrow \mathbf{k}_j + g_K w_{\ell,h}^K P_K \mathbf{k}_j$
 - 11: $\mathbf{v}'_j \leftarrow \mathbf{v}_j + g_V w_{\ell,h}^V P_V \mathbf{v}_j$
 - 12: **end for**
 - 13: Proceed with $\text{softmax}(QK'^\top / \sqrt{d})V'$
-

positive and negative conditions. We map this to a continuous weight via the softplus function:

$$w_{\ell,h} = \text{softplus}(D_{\ell,h} - \delta_{\min}) = \log(1 + \exp(D_{\ell,h} - \delta_{\min})) \quad (6)$$

We choose softplus because it assigns larger weights to strongly discriminative heads and reduced but nonzero weights to weaker ones.

The projection P and weight w share the same data signal $D_{\ell,h}$: a large D means the head’s contrastive signal is strong, so its projection is reliable (Proposition 1) and it should receive stronger steering. The two are not independent choices but consequences of a single data-driven measure.

3.3 Dual-channel steering

We apply the subspace learning framework (Section 3.2) separately to Key and Value spaces:

$$\text{Key: } \Omega_\Delta^K = \mathbf{H}_K^\top (\mathbf{H}_K^+ - \mathbf{H}_K^-) / N \xrightarrow{\text{SVD}} P_K, w_K \quad (7)$$

$$\text{Value: } \Omega_\Delta^V = \mathbf{H}_V^\top (\mathbf{H}_V^+ - \mathbf{H}_V^-) / N \xrightarrow{\text{SVD}} P_V, w_V \quad (8)$$

At inference time, for each highlighted token $j \in \mathcal{S}$, both channels are simultaneously edited:

$$\mathbf{k}'_j = \mathbf{k}_j + g_K \cdot w_{\ell,h}^K \cdot P_K \cdot \mathbf{k}_j \quad (9)$$

$$\mathbf{v}'_j = \mathbf{v}_j + g_V \cdot w_{\ell,h}^V \cdot P_V \cdot \mathbf{v}_j \quad (10)$$

where g_K, g_V are gain scalars controlling the strength of routing and content steering, respectively.

Geometric interpretation. Each transformation $(I + g \cdot w \cdot P) \cdot \mathbf{x}$ rescales the component of $\mathbf{x}$ within the learned subspace by a factor of $(1 + g \cdot w)$ while leaving the orthogonal component unchanged. Unlike uniform rescaling, each head’s scaling factor is individually modulated by its discriminability $w_{\ell,h}$. This differs from adding a constant bias to the attention logit: Key scaling produces a *query-dependent* boost—the attention shift varies with the semantic content of each query position—whereas logit bias applies a fixed offset regardless of query.

Instantiations. The framework admits two variants:

- PRISM- Δ ($g_V = 0$): steers the routing channel via pre-attention Key editing.
- PRISM- ΔV ($g_K > 0, g_V > 0$): steers both channels, capturing all three gain terms in Eq. 3.

Algorithm 1 states the complete pipeline, which both variants share.

4 Setup

Benchmarks. We evaluate on three prompt highlighting benchmarks: **BiasBios** (DeArteaga et al., 2019) (occupation prediction from highlighted biographies; metrics: Accuracy, Fluency, Consistency), **CounterFact** (Meng et al., 2022) (knowledge conflict resolution; metrics: Efficacy, Paraphrase), and **Pronoun Change** (Li et al., 2026) (rewriting gendered pronouns to gender-neutral forms according to highlighted instructions; metrics: P. Score, All-changed P. Score). Formal metric definitions are in Appendix C; hyperparameters are tuned on held-out validation sets.

Models. We evaluate across two architecture families and three scales: QWEN3-4B/8B/14B-Base (Qwen Team, 2025) and GEMMA3-4B/12B-PT (Gemma Team, 2025). Both families use Grouped Query Attention (GQA); the head index (ℓ, h) refers to KV heads, and PRISM- Δ constructs one projection per KV head.

Baselines. We compare against five baselines: **Vanilla** (no steering), ****.-marked** (surrounding highlighted text with asterisks), **PASTA** (Zhang et al., 2024), **SPA** (Tian & Zhang, 2025), and **SEKA** (Li et al., 2026). A comparison with AdaSEKA, a higher-cost multi-expert variant, is provided in Appendix D.

PRISM- Δ configurations. PRISM- Δ uses Key-only steering ($g_V = 0$). PRISM- ΔV steers both Key and Value ($g_V > 0$). Projections are constructed offline from 100 synthetic contrastive QA pairs. Hyperparameters are listed in Appendix E. All experiments use greedy decoding on single NVIDIA H20 GPUs. Full setup details are in Appendix F.

5 Results

5.1 Main results

Table 2 presents results across all three benchmarks and five models. On BiasBios, PRISM- Δ achieves the best accuracy on Qwen3 models with relative gains up to +1.6%; on CounterFact, it ties SEKA at 98.86% on GEMMA3-12B and reaches 99.24% on QWEN3-8B; on

Model	Metric	Baselines				Ours			
		Vanilla	**	-marked	PASTA	SPA	SEKA	PRISM-Δ	PRISM-ΔV
BiasBios									
QWEN3-4B	Accuracy	79.80	82.94 _{▲3.14}	89.58 _{▲9.78}	68.00 _{▼11.80}	90.92 _{▲11.12}	92.38 _{▲12.58}	92.36 _{▲12.56}	
QWEN3-8B		76.22	80.60 _{▲4.38}	86.32 _{▲10.10}	37.02 _{▼39.20}	88.74 _{▲12.52}	89.62 _{▲13.40}	89.12 _{▲12.90}	
QWEN3-14B		85.10	90.94 _{▲5.84}	88.46 _{▲3.36}	57.86 _{▼27.24}	90.28 _{▲5.18}	91.68 _{▲6.58}	91.20 _{▲6.10}	
GEMMA3-4B		89.90	91.00 _{▲1.10}	82.58 _{▼7.32}	48.02 _{▼41.88}	92.42 _{▲2.52}	92.90 _{▲3.00}	91.64 _{▲1.74}	
GEMMA3-12B		91.32	92.90 _{▲1.58}	94.72 _{▲3.40}	46.88 _{▼44.44}	93.04 _{▲1.72}	92.22 _{▲0.90}	91.94 _{▲0.62}	
CounterFact									
QWEN3-4B	Efficacy	45.00	57.70 _{▲12.70}	97.16 _{▲52.16}	65.24 _{▲20.24}	99.02 _{▲54.02}	99.14 _{▲54.14}	98.08 _{▲53.08}	
	Paraphrase	45.64	52.12 _{▲6.48}	96.03 _{▲50.39}	57.71 _{▲12.07}	98.61 _{▲52.97}	98.52 _{▲52.88}	96.51 _{▲50.87}	
QWEN3-8B	Efficacy	39.04	56.24 _{▲17.20}	92.70 _{▲53.66}	69.26 _{▲30.22}	99.08 _{▲60.04}	99.24 _{▲60.20}	98.46 _{▲59.42}	
	Paraphrase	39.59	49.80 _{▲10.21}	91.68 _{▲52.09}	58.76 _{▲19.17}	98.96 _{▲59.37}	99.10 _{▲59.51}	98.04 _{▲58.45}	
QWEN3-14B	Efficacy	37.56	45.52 _{▲7.96}	76.84 _{▲39.28}	84.22 _{▲46.66}	98.92 _{▲61.36}	99.00 _{▲61.44}	98.98 _{▲61.42}	
	Paraphrase	36.12	40.12 _{▲4.00}	66.33 _{▲30.21}	76.11 _{▲39.99}	99.02 _{▲62.90}	98.82 _{▲62.70}	98.72 _{▲62.60}	
GEMMA3-4B	Efficacy	55.04	57.56 _{▲2.52}	78.36 _{▲23.32}	93.90 _{▲38.86}	98.04 _{▲43.00}	98.10 _{▲43.06}	97.68 _{▲42.64}	
	Paraphrase	47.77	45.82 _{▼1.95}	59.53 _{▲11.76}	91.92 _{▲44.15}	98.83 _{▲51.06}	96.24 _{▲48.47}	96.01 _{▲48.24}	
GEMMA3-12B	Efficacy	45.34	48.72 _{▲3.38}	68.30 _{▲22.96}	93.76 _{▲48.42}	98.86 _{▲53.52}	98.86 _{▲53.52}	93.40 _{▲48.06}	
	Paraphrase	37.21	36.69 _{▼0.52}	71.72 _{▲34.51}	91.24 _{▲54.03}	99.27 _{▲62.06}	99.30 _{▲62.09}	91.86 _{▲54.65}	
Pronoun Change									
QWEN3-4B	P. Score	93.14	95.76 _{▲2.62}	95.82 _{▲2.68}	80.27 _{▼12.87}	95.18 _{▲2.04}	96.18 _{▲3.04}	96.06 _{▲2.92}	
	All-changed	90.52	93.88 _{▲3.36}	94.64 _{▲4.12}	78.19 _{▼12.33}	93.26 _{▲2.74}	94.60 _{▲4.08}	94.52 _{▲4.00}	
QWEN3-8B	P. Score	98.00	98.10 _{▲0.10}	98.86 _{▲0.86}	72.61 _{▼25.39}	98.56 _{▲0.56}	99.40 _{▲1.40}	99.48 _{▲1.48}	
	All-changed	97.84	97.84	98.72 _{▲0.88}	74.95 _{▼22.89}	98.26 _{▲0.42}	99.30 _{▲1.46}	99.34 _{▲1.50}	
QWEN3-14B	P. Score	98.42	98.86 _{▲0.44}	90.98 _{▼7.44}	91.60 _{▼6.82}	98.66 _{▲0.24}	99.66 _{▲1.24}	99.66 _{▲1.24}	
	All-changed	98.22	98.68 _{▲0.46}	90.94 _{▼7.28}	92.20 _{▼6.02}	98.54 _{▲0.32}	99.56 _{▲1.34}	99.56 _{▲1.34}	
GEMMA3-4B	P. Score	41.34	38.86 _{▼2.48}	67.39 _{▲26.05}	76.05 _{▲34.71}	81.53 _{▲40.19}	89.08 _{▲47.74}	90.16 _{▲48.82}	
	All-changed	35.25	32.45 _{▼2.80}	66.43 _{▲31.18}	74.45 _{▲39.20}	81.11 _{▲45.86}	90.58 _{▲55.33}	91.68 _{▲56.43}	
GEMMA3-12B	P. Score	93.92	95.78 _{▲1.86}	68.47 _{▼25.45}	86.41 _{▼7.51}	97.70 _{▲3.78}	97.94 _{▲4.02}	94.54 _{▲0.62}	
	All-changed	94.96	96.42 _{▲1.46}	68.01 _{▼26.95}	84.99 _{▼9.97}	97.24 _{▲2.28}	97.46 _{▲2.50}	94.42 _{▼0.54}	

Table 2: Main results across three benchmarks (%). **Bold**: best; underline: second best. Subscripts show absolute change vs. Vanilla.

Pronoun Change, our methods outperform SEKA on all 5 models with relative gains up to +10.6% on P. Score. Including the Lost-in-the-Middle results in Table 4, our methods match or exceed the best existing method on 19 out of 20 model×benchmark configurations; the exception is GEMMA3-12B on BiasBios. Appendix G repeats the comparison on three instruction-tuned checkpoints.

Statistical reliability. Statistical reliability is high: across five projection subsets, standard deviation is 0.05–0.15%, and PRISM- Δ matches or exceeds SEKA on 14 out of 15 model×benchmark cells ($p < 0.001$). Detailed variance analysis is in Appendix M and sign test derivations in Appendix N.

PRISM- Δ vs. PRISM- Δ V. On Pronoun Change, PRISM- Δ V surpasses PRISM- Δ on GEMMA3-4B by 1.08 percentage points, demonstrating that dual-channel steering can also improve accuracy when content rewriting is central. On QWEN3-4B BiasBios the two variants are within 0.02 points. We recommend PRISM- Δ as the default and PRISM- Δ V when generation quality is prioritized; see Section 6.2 for a detailed decomposition.

5.2 Ablation study

Table 3 uses a 2×2 factorial design. Removing softplus costs 0.96% and removing differential projection costs 0.86% on QWEN3-4B. The interaction is strongly super-additive: from

Configuration	Projection	Weighting	Channels	QWEN3-4B		QWEN3-8B	
				Acc	Δ	Acc	Δ
PRISM- Δ V	Diff. Ω_Δ	Softplus	K+V	92.36	+15.7	89.12	+16.9
PRISM- Δ (−V)	Diff. Ω_Δ	Softplus	K only	92.38	+15.8	89.62	+17.6
PRISM- Δ -V (−K)	Diff. Ω_Δ	Softplus	V only	82.44	+3.3	79.16	+3.9
−Differential	Independent	Softplus	K only	91.52	+14.7	88.90	+16.6
−Softplus	Differential	Uniform $w=1$	K only	91.42	+14.6	88.22	+15.7
−Both	Independent	Uniform $w=1$	K only	91.44	+14.6	88.50	+16.1
<i>Baselines</i>							
SEKA ($\delta=0.08$)	Dual indep.	Binary	K only	89.56	+12.2	85.52	+12.2
SEKA (default)	Dual indep.	Binary $\delta=0.12$	K only	90.92	+13.9	88.74	+16.4
Vanilla	—	—	—	79.80	—	76.22	—

Table 3: Ablation study on BiasBios (Top-1 Accuracy %). Δ : relative improvement (%) over Vanilla.

Model	Vanilla	SEKA	PRISM- Δ	PRISM- Δ V
QWEN3-4B	45.47	49.14 $\uparrow_{3.67}$	51.43 $\uparrow_{5.96}$	50.57 $\uparrow_{5.10}$
QWEN3-8B	46.69	50.29 $\uparrow_{3.60}$	52.71 $\uparrow_{6.02}$	51.86 $\uparrow_{5.17}$
QWEN3-14B	57.13	60.86 $\uparrow_{3.73}$	62.57 $\uparrow_{5.44}$	62.00 $\uparrow_{4.87}$
GEMMA3-4B	44.36	47.71 $\uparrow_{3.35}$	49.29 $\uparrow_{4.93}$	50.14 $\uparrow_{5.78}$
GEMMA3-12B	51.64	55.57 $\uparrow_{3.93}$	54.43 $\uparrow_{2.79}$	55.57 $\uparrow_{3.93}$

Table 4: Lost-in-the-Middle: Average Exact Match (%) across 7 gold positions. Steering targets the middle region (passages 4–25). Setup details in Appendix F.

the −Both baseline, differential projection alone yields −0.02% while softplus alone yields +0.08%, yet their combination yields +0.94%. At matched $\delta_{\min}=0.08$, PRISM- Δ achieves its best result while SEKA drops by 1.36%, because softplus smoothly down-weights noisy heads rather than giving all activated heads equal weight. PRISM- Δ -V alone reaches 82.44%, confirming that the Value channel carries independently useful discriminative signal. A CounterFact ablation (Appendix O) and data quantity ablation (Appendix P) corroborate these findings.

5.3 Lost-in-the-Middle

Table 4 evaluates on the lost-in-the-middle benchmark (Liu et al., 2024), with detailed setup in Appendix F. Steering is applied selectively to the middle region of 30-passage contexts. Our methods match or exceed SEKA on all five models, with relative gains up to 5.1% (PRISM- Δ V on GEMMA3-4B), confirming that PRISM- Δ scales to long-context retrieval.

5.4 Efficiency analysis

As shown in Table 5, SEKA adds negligible overhead (+0.01 s, +0.04 GB) over Vanilla. PRISM- Δ adds +0.30 s latency ($1.26\times$) due to the per-head weighted projection, with negligible additional memory (+0.02 GB). PRISM- Δ V adds a further +0.02 s for Value editing. Both variants remain fully compatible with FlashAttention (Dao et al., 2022; Dao, 2024). For comparison, PASTA adds +1.03 s and +23.12 GB, and SPA adds +5.32 s (Li et al., 2026), placing PRISM- Δ in between SEKA and these heavier alternatives.

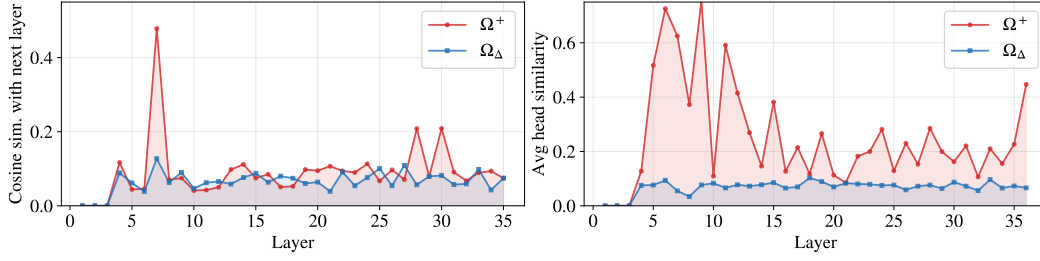


Figure 4: Direction consistency analysis. Ω^+ directions show high cross-head similarity (dominated by shared structural directions), while Ω_Δ directions are nearly independent (close to random baseline), confirming that differential projection extracts head-specific discriminative directions.

Method	Avg. Latency (s)	Relative	Peak Mem (GB)	FlashAttn
Vanilla	1.180	1.0×	26.39	✓
SEKA	1.194 (+0.01)	1.0×	26.43 (+0.04)	✓
PRISM- Δ	1.481 (+0.30)	1.26×	26.41 (+0.02)	✓
PRISM- ΔV	1.502 (+0.32)	1.27×	26.43 (+0.04)	✓

Table 5: Inference efficiency on QWEN3-8B.

6 Analysis

6.1 Discriminative subspace quality

We validate the theoretical motivation of differential projection (Proposition 1) on QWEN3-4B-Base.

Direction consistency. Figure 4 tracks the similarity of top singular directions across layers and heads. **Left:** adjacent-layer similarity measures how much a head’s direction changes from one layer to the next. Ω^+ directions show occasional cross-layer alignment spikes (0.2–0.5), while Ω_Δ directions remain low throughout (≤ 0.13), indicating layer-specific discriminative structure. **Right:** within-layer head similarity measures redundancy among heads in the same layer. Ω^+ heads converge to similar directions (up to 0.76 in early layers), while Ω_Δ heads remain near-independent throughout all layers. Aggregate statistics are provided in Table 22 of Appendix Q.

6.2 Dual-channel contribution decomposition

Table 6 reveals that the Key channel primarily drives accuracy (+12.58 points over Vanilla), while the Value channel preserves generation quality. PRISM- Δ incurs only 53% of SEKA’s fluency cost. Both PRISM- Δ and PRISM- ΔV simultaneously outperform SEKA on all three

Configuration	Top-1 Acc (%)	Fluency ($\uparrow$)	Consistency ($\uparrow$)	Fluency Cost ($\downarrow$)
Vanilla	79.80	4.638	0.116	0.000
SEKA	90.92 $\Delta_{11.12}$	3.681	0.112	0.957
PRISM- Δ	92.38 $\Delta_{12.58}$	4.134	0.127	0.504
PRISM- Δ -V	82.44 $\Delta_{2.64}$	4.666	0.121	−0.028
PRISM- ΔV ($g_V=0.10$)	92.36 $\Delta_{12.56}$	4.116	0.125	0.522

Table 6: Dual-channel decomposition on BiasBios with QWEN3-4B. Best per metric in **bold**. Fluency Cost = Vanilla Fluency – Method Fluency.

metrics; Appendix I confirms the fluency pattern across models and tasks using automatic and human-validated qualitative evaluation.

Layer-wise and cross-model complementarity. Figure 2b shows that the two channels have complementary depth profiles. Key signal averages 0.175 in middle layers and Value averages 0.307 in late layers. Across models, Qwen3 becomes increasingly Value-dominant with scale (K/V ratio $1.02 \rightarrow 0.58$), whereas Gemma3 remains Key-dominant (ratio > 1.2); full measurements appear in Appendix L. This architecture-dependent balance is consistent with PRISM- Δ being the safer default, while Value steering is most helpful for fluency and rewriting. On Pronoun Change with GEMMA3-4B, PRISM- Δ V outperforms PRISM- Δ by 1.08 points, an interaction consistent with the cross gain in Eq. 3.

6.3 Head selection robustness

Figure 5 sweeps $\delta_{\min}$ from 0.06 to 0.20. SEKA drops by 3.14% from its optimum when $\delta_{\min}$ is reduced from 0.12 to 0.06: hard thresholding gives all activated heads equal weight $w=1$, so including noisy heads degrades performance. PRISM- Δ fluctuates by only 0.60% across the same range, because softplus automatically assigns lower weights to less discriminative heads. This robustness substantially reduces tuning cost; hyperparameter sweeps are summarized in Appendix R. A per-sample log-probability analysis (Figure 8, Appendix S) compares the two methods sample by sample: PRISM- Δ rescues 154 samples that SEKA misclassifies while losing only 81, for a net gain of +73. The 2:1 asymmetry reflects systematic correction of shared-feature confusion, consistent with the shared-direction elimination of Proposition 1(b). Qualitative examples are in Appendix T. Appendices H, K, and J examine automatically selected spans, imperfect highlights, and cross-domain projection reuse.

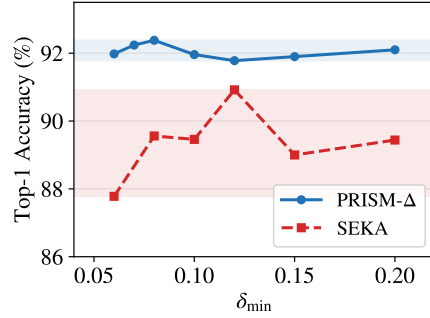


Figure 5: $\delta_{\min}$ sensitivity on BiasBios, QWEN3-4B.

7 Conclusion

We presented PRISM- Δ , an inference-time prompt-highlighting method combining three ideas: differential cross-covariance decomposition removes directions shared by relevant and irrelevant contexts, continuous softplus weighting retains weak but useful heads, and dual-channel steering separates the routing role of Keys from the content role of Values. Key steering drives accuracy gains while Value steering reduces the fluency cost of editing.

PRISM- Δ extends to instruction-tuned models and to spans supplied by automatic retrievers, making it a lightweight component for RAG and long-document pipelines. Future work can learn selector-steering policies jointly and extend differential subspace steering to broader instruction prioritization and multimodal contexts.

8 Limitations

The main practical limitation is that the preferred gains vary substantially across benchmarks and models, so the current results require validation-set selection rather than a universal default. Gemma3 hyperparameters often differ from Qwen3 settings, and GEMMA3-12B on Pronoun Change uses a negative g_K , a pattern also reported for SEKA (Li et al., 2026). On near-saturated benchmarks such as CounterFact, absolute differences between steering methods are small. The dual-channel variant has a clear failure case on GEMMA3-12B CounterFact, which the available experiments do not explain. We therefore recommend PRISM- Δ as the default, with separate validation before enabling Value steering.

Acknowledgments

This work is supported in part by the Beijing Natural Science Foundation No. 4262033, the National Key R&D Program of China under Grant No. 2025YFC3309700, and the National Natural Science Foundation of China under Grant Nos. U25B2076, 62441229, and 62377043.

Ethics Statement

In conducting this research we have kept every component of the pipeline open and inspectable. All experiments use publicly available benchmarks (De-Arteaga et al., 2019; Meng et al., 2022; Li et al., 2026; Liu et al., 2024) and publicly released model checkpoints (Qwen Team, 2025; Gemma Team, 2025), so that our findings rest on accessible and widely used resources. PRISM- Δ operates at inference time and leaves model parameters unchanged: it rescales the representations of spans the user has already supplied rather than introducing new content. No personal data were collected, and the contrastive prompts used to build the projections are model-generated, as described in Appendix U.

References

- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 2318–2335. Association for Computational Linguistics, 2024.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? An analysis of BERT’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pp. 276–286. Association for Computational Linguistics, 2019.
- Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 120–128. ACM, 2019.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL <https://transformer-circuits.pub/2021/framework/index.html>.
- Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

- Yuyao Ge, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Xuanshan Zhou, Jiayu Yao, Jiafeng Guo, and Xueqi Cheng. Focusing by contrastive attention: Enhancing VLMs' visual reasoning. *arXiv preprint arXiv:2509.06461*, 2025.
- Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- Evan Hernandez, Belinda Z. Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling, COLM 2024*, 2024.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022, 2022.
- Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- Weixian Waylon Li, Yuchen Niu, Yongxin Yang, Keshuang Li, Tiejun Ma, and Shay B. Cohen. Spectral attention steering for prompt highlighting. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. OpenReview.net, 2026.
- Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 4582–4597. Association for Computational Linguistics, 2021.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173, 2024.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14014–14024, 2019.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Maria Ponti, and Shay B. Cohen. Spectral editing of activations for large language model alignment. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.
- Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009.

- Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.
- Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 566–581. Association for Computational Linguistics, 2022.
- Yuan Tian and Tianyi Zhang. Selective prompt anchoring for code generation. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.
- Praveen Venkateswaran and Danish Contractor. Spotlight your instructions: Instruction-following with dynamic attention steering. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pp. 3752–3770. Association for Computational Linguistics, 2026.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5797–5808. Association for Computational Linguistics, 2019.
- Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.
- Li-Ming Zhan, Bo Liu, Yujie Feng, Chengqiang Xie, Jiannong Cao, and Xiao-Ming Wu. REAL: Reading out transformer activations for precise localization in language model steering. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. OpenReview.net, 2026.
- Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- 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*, 2025.

A Notation

Table 7 summarizes the notation used throughout the paper.

Symbol	Definition	Description
<i>Model architecture</i>		
L	Number of layers	Total Transformer layers in the model
n_h	Number of heads	KV heads per layer
d	Head dimension	Dimensionality of each head’s Key/Value space
(ℓ, h)	Layer-head index	Identifies a specific attention head
$\mathbf{x}_j$	Token embedding	Input representation of token j
W_Q, W_K, W_V	Projection matrices	Linear projections for Query, Key, Value
$\mathbf{k}_j, \mathbf{v}_j$	Key/Value vectors	Per-token Key and Value representations ($\in \mathbb{R}^d$)
α_{ij}	Attention weight	Softmax-normalized attention from token i to j
<i>Contrastive representations</i>		
$\mathbf{H} \in \mathbb{R}^{N \times d}$	Neutral representations	Key/Value matrix, context-only condition
$\mathbf{H}^+ \in \mathbb{R}^{N \times d}$	Positive representations	Key/Value matrix, relevant-question condition
$\mathbf{H}^- \in \mathbb{R}^{N \times d}$	Negative representations	Key/Value matrix, irrelevant-question condition
N	Number of samples	Contrastive triplets used for projection construction
<i>Cross-covariance and SVD</i>		
Ω^+	$\mathbf{H}^\top \mathbf{H}^+ / N$	Uncentered cross-covariance (neutral $\times$ positive)
Ω^-	$\mathbf{H}^\top \mathbf{H}^- / N$	Uncentered cross-covariance (neutral $\times$ negative)
Ω_Δ	$\Omega^+ - \Omega^-$	Differential cross-covariance
U, Σ, V	SVD of Ω_Δ	Left/right singular vectors and singular values
k	Retained rank	Number of top singular vectors retained (set by γ)
<i>PRISM parameters</i>		
P_K, P_V	$U_{:,k} U_{:,k}^\top$	Projection matrices for Key and Value channels
$\mathbf{r}_i^+, \mathbf{r}_i^-$	Head representation	Per-sample Key or Value vector, positive/negative ($\in \mathbb{R}^d$)
$D_{\ell,h}$	$\frac{1}{N} \sum_i \ \mathbf{r}_i^+ - \mathbf{r}_i^-\ _2$	Per-head discriminability score
$w_{\ell,h}$	$\text{softplus}(D_{\ell,h} - \delta_{\min})$	Per-head importance weight (generic)
$w_{\ell,h}^K$	$\text{softplus}(D_{\ell,h}^K - \delta_{\min})$	Per-head Key weight
$w_{\ell,h}^V$	$\text{softplus}(D_{\ell,h}^V - \delta_{\min})$	Per-head Value weight
γ	Variance threshold	Cumulative singular value ratio for rank selection
$\delta_{\min}$	Minimum discriminability	Softplus shift parameter
g_K, g_V	Gain parameters	Steering strength for Key and Value channels
<i>Steering formulas</i>		
$\mathbf{k}'_j$	$\mathbf{k}_j + g_K \cdot w_{\ell,h}^K \cdot P_K \cdot \mathbf{k}_j$	Steered Key vector for highlighted token j
$\mathbf{v}'_j$	$\mathbf{v}_j + g_V \cdot w_{\ell,h}^V \cdot P_V \cdot \mathbf{v}_j$	Steered Value vector for highlighted token j

Table 7: Notation used throughout the paper.

B Proof of Proposition 1

Proof. **Part (a)** follows directly from the Eckart–Young–Mirsky theorem: for any matrix $M \in \mathbb{R}^{m \times n}$ with SVD $M = U \Sigma V^\top$, the rank- k truncation $M_k = U[:, :k] \Sigma[:, :k] V[:, :k]^\top$ minimizes $\|M - M_k\|_F$ over all rank- k matrices. Equivalently, $U[:, :k]$ maximizes $\|U[:, :k]^\top M\|_F^2$ over all k -column orthonormal matrices. Applying this to $M = \Omega_\Delta$ gives part (a).

Part (b): If $(\Omega^+)^\top \mathbf{u}_s = (\Omega^-)^\top \mathbf{u}_s$, then $\Omega_\Delta^\top \mathbf{u}_s = (\Omega^+ - \Omega^-)^\top \mathbf{u}_s = \mathbf{0}$, so $\mathbf{u}_s$ lies in the left null space of Ω_Δ and is orthogonal to its nonzero left singular directions. $\square$

C Evaluation metrics

We provide formal definitions of all evaluation metrics used in the three benchmarks.

BiasBios.

- **Top-1 Accuracy.** The generated text is parsed for an occupation label. If the predicted occupation matches the ground-truth target, the sample is correct. Accuracy is the fraction of correct predictions over the test set.
- **Fluency.** Defined as the weighted n-gram entropy of the generated text:

$$H_n = - \sum_g p_n(g) \log_2 p_n(g), \quad \text{Fluency} = \frac{1}{2} \left(\frac{2}{3} H_2 + \frac{4}{3} H_3 \right). \quad (11)$$

where $p_n(g)$ is the empirical distribution of n -grams. Higher values indicate greater n -gram diversity and better fluency.

- **Consistency.** The cosine similarity between the TF-IDF vector of the generated text and the mean TF-IDF vector of reference biographies for the target occupation, averaged over all samples.

CounterFact.

- **Efficacy Score.** The probability that the model assigns higher likelihood to the target answer than to the original (pre-edit) answer:

$$\text{Efficacy} = \frac{1}{|\mathcal{D}|} \sum_{(s, o^*, o) \in \mathcal{D}} \mathbf{1}[p(o^* | s) > p(o | s)] \quad (12)$$

where s is the subject prompt, o^* is the target (counterfactual) object, and o is the original object.

- **Paraphrase Score.** Same as Efficacy, but evaluated on paraphrased versions of the subject prompt, measuring robustness to surface-form variation.

Pronoun Change. Given an input biography containing gendered pronouns and an instruction to replace them with gender-neutral forms, two metrics are computed based on the generated text:

- **P. Score** (basic pronoun conversion rate). Let $\mathcal{P}_b = \{\text{she, he}\}$ be the basic pronoun set. For each sample, count the number of basic pronouns in the original text (n_b) and the number still remaining in the generated text (r_b). The conversion rate is $(n_b - r_b)/n_b$. This is multiplied by the content overlap ratio (fraction of non-pronoun content tokens preserved) to penalize generations that trivially remove pronouns by truncating the text:

$$\text{P. Score} = \frac{n_b - r_b}{n_b} \times \frac{|\text{content}_{\text{gen}} \cap \text{content}_{\text{orig}}|}{|\text{content}_{\text{orig}}|} \quad (13)$$

- **All-changed P. Score.** Same formula but with the extended pronoun set $\mathcal{P}_a = \{\text{she, he, her, him, hers, his, herself, himself}\}$. Because $\mathcal{P}_b \subset \mathcal{P}_a$ but the denominators differ (n_b vs. n_a), it is mathematically possible for All-changed to exceed P. Score when the model converts extended forms more successfully than basic ones.

D Comparison with AdaSEKA

AdaSEKA (Li et al., 2026) extends SEKA with a query-adaptive multi-expert routing mechanism that dynamically selects among multiple projection subspaces based on the input query’s semantic intent. While this achieves strong performance on several benchmarks, it comes at substantial memory cost: +15.59 GB peak memory, placing it in a fundamentally different memory regime from single-projection methods like SEKA and PRISM- Δ .

Table 8 provides the full comparison. Key observations:

Benchmark	Model	SEKA	PRISM- Δ	PRISM- ΔV	AdaSEKA	Δ
BiasBios	QWEN3-4B	90.92	92.38	92.36	91.86	+0.52
	QWEN3-8B	88.74	89.62	89.12	88.50	+1.12
	QWEN3-14B	90.28	91.68	91.20	91.22	+0.46
	GEMMA3-4B	92.42	92.90	91.64	<u>92.92</u>	-0.02
	GEMMA3-12B	93.04	92.22	91.94	<u>91.14</u>	+1.90
CounterFact	QWEN3-4B	99.02	99.14	98.08	98.90	+0.24
	QWEN3-8B	99.08	99.24	98.46	99.00	+0.24
	QWEN3-14B	98.92	99.00	98.98	99.00	0.00
	GEMMA3-4B	98.04	98.10	97.68	<u>98.74</u>	-0.64
	GEMMA3-12B	98.86	98.86	93.40	<u>92.48</u>	+6.38
Pronoun	QWEN3-4B	95.18	96.18	96.06	94.54	+1.64
	QWEN3-8B	98.56	99.40	99.48	99.68	-0.20
	QWEN3-14B	98.66	99.66	99.66	<u>99.88</u>	-0.22
	GEMMA3-4B	81.53	89.08	90.16	<u>93.76</u>	-3.60
	GEMMA3-12B	97.70	97.94	94.54	96.88	+1.06

Table 8: Full comparison including AdaSEKA (%). **Bold**: best among lightweight methods; underline: best overall (including AdaSEKA). Δ is the best lightweight method minus AdaSEKA.

- **On BiasBios**, the best lightweight method outperforms AdaSEKA on 4 out of 5 models. PRISM- Δ leads on all three Qwen3 models, while SEKA leads on GEMMA3-12B with the largest margin of +1.90%. AdaSEKA wins only on GEMMA3-4B by 0.02%.
- **On CounterFact**, PRISM- Δ surpasses SEKA on 4 of 5 models and matches it on the fifth (GEMMA3-12B: both 98.86%). Both methods operate near the performance ceiling (98–99%), so absolute differences are small.
- **On Pronoun Change**, AdaSEKA achieves the highest scores on 3 out of 5 models (QWEN3-8B/14B, GEMMA3-4B). However, on QWEN3-4B and GEMMA3-12B, the best lightweight method surpasses AdaSEKA.
- **Overall**, among the 15 model $\times$ benchmark cells, the best lightweight method matches or exceeds AdaSEKA in **10 out of 15 cases**. In the reported efficiency setting, PRISM- Δ requires 1.26 $\times$ the base latency and substantially less memory than AdaSEKA.

In these 15 cells, PRISM- Δ matches or exceeds AdaSEKA in 10 while using less memory. Whether the multi-expert overhead is acceptable depends on the deployment setting and the importance of the cells where AdaSEKA performs better.

E Hyperparameters

Tables 9 and 10 list the validation-selected hyperparameters for PRISM- Δ and PRISM- ΔV , respectively. Each model–benchmark configuration is selected separately by validation-set grid search.

The gain g_K varies across benchmarks: ~ 0.40 for BiasBios, 1.10–6.00 for CounterFact, and 0.05–0.30 for Pronoun Change, reflecting different steering intensities per task. The variance retention threshold γ is on average higher for Qwen3 than for Gemma3. On GEMMA3-12B (Pronoun Change), $g_K = -0.30$; see Section 8 for discussion.

F Detailed experimental setup

Hardware. All experiments are conducted on NVIDIA H20 GPUs with 144 GB memory. Projection construction uses a single GPU; evaluation uses a single GPU with batch size 256 for BiasBios and CounterFact, and batch size 1 for Pronoun Change.

Data splits. For BiasBios, we evaluate on the first 5000 samples (indices 0–4999); a 500-sample validation subset is used for hyperparameter tuning. For CounterFact, samples

Model	Benchmark	γ	$\delta_{\min}$	g_K
QWEN3-4B	BiasBios	0.998	0.08	0.40
QWEN3-8B		0.998	0.08	0.40
QWEN3-14B		0.998	0.08	0.40
GEMMA3-4B		0.850	0.08	0.50
GEMMA3-12B		0.980	0.00	0.40
QWEN3-4B	CounterFact	0.998	0.13	1.90
QWEN3-8B		0.960	0.12	2.70
QWEN3-14B		0.870	0.08	3.00
GEMMA3-4B		0.990	0.12	6.00
GEMMA3-12B		0.990	0.50	1.10
QWEN3-4B	Pronoun Change	0.880	0.15	0.15
QWEN3-8B		0.900	0.15	0.05
QWEN3-14B		0.880	0.15	0.05
GEMMA3-4B		0.800	0.15	0.30
GEMMA3-12B		0.700	0.30	-0.30

Table 9: PRISM- Δ hyperparameters.

Model	Benchmark	γ	$\delta_{\min}$	g_K	g_V
QWEN3-4B	BiasBios	0.998	0.08	0.40	0.10
QWEN3-8B		0.998	0.08	0.40	0.10
QWEN3-14B		0.998	0.08	0.40	0.10
GEMMA3-4B		0.800	0.12	0.30	0.10
GEMMA3-12B		0.994	0.00	0.40	0.10
QWEN3-4B	CounterFact	0.998	0.13	1.90	0.02
QWEN3-8B		0.960	0.12	2.70	0.05
QWEN3-14B		0.870	0.08	3.00	0.05
GEMMA3-4B		0.990	0.12	6.00	0.10
GEMMA3-12B		0.990	0.40	3.00	0.50
QWEN3-4B	Pronoun Change	0.880	0.15	0.15	0.05
QWEN3-8B		0.900	0.15	0.05	0.02
QWEN3-14B		0.880	0.15	0.05	0.02
GEMMA3-4B		0.800	0.15	0.30	0.10
GEMMA3-12B		0.700	0.30	0.05	0.02

Table 10: PRISM- ΔV hyperparameters.

0–4999 serve as the validation set and samples 5000–9999 as the test set. Pronoun Change uses 500 validation and 500 test samples following [Li et al. \(2026\)](#).

Evaluation protocol. All methods use greedy decoding with a maximum of 64 new tokens for BiasBios, 32 for CounterFact, and 128 for Pronoun Change. We report Top-1 Accuracy for BiasBios, Efficacy and Paraphrase scores for CounterFact, and P. Score and All-changed P. Score for Pronoun Change. Fluency is measured by weighted 2- and 3-gram entropy, and Consistency by TF-IDF cosine similarity to target-occupation references.

Contrastive data. The 100 synthetic contrastive QA pairs are generated by GPT-4o. Each pair contains two unrelated text contexts paired with two questions, one relevant and one irrelevant to each context. This yields 200 contrastive triplets from which Key and Value representations are extracted at the answer token position.

Projection construction. For each model, we extract representations from all layers and heads under three prompting conditions (neutral, positive, negative). The differential cross-covariance Ω_Δ is computed per head, followed by SVD. The top- k singular vectors

are retained based on the cumulative energy threshold γ . The mean representation norm difference determines the head’s discriminability score $D_{\ell,h}$, which is mapped to a weight via softplus. The resulting projections are stored and reused across inference runs. Construction takes 3–8 minutes per model on one GPU; online steering adds one matrix–vector product per head and highlighted token.

Lost-in-the-Middle setup. We use the NaturalQuestions-based benchmark of Liu et al. (2024) with 30 passages per input, where one gold passage contains the answer and the remaining 29 are distractors. The gold passage is placed at 7 positions (0, 4, 9, 14, 19, 24, 29) to probe positional sensitivity. Steering is applied selectively to the middle region (passages 4–25), targeting the positional recall deficit. Generation is limited to 60 tokens. We report Exact Match averaged across all 7 positions.

G Instruction-tuned evaluation

Instruction-tuned checkpoints wrap prompts in model-specific chat templates, shifting the position at which the neutral, positive, and negative representations are read. We align extraction to the first generated assistant token for Qwen3-4B-Instruct, Qwen3-8B, and Gemma3-4B-it (Qwen Team, 2025; Gemma Team, 2025). The projection protocol is otherwise unchanged: each head uses 100 synthetic contrastive pairs, $\gamma = 0.9$, held-out selection over $g_K \in [0.05, 6.0]$ and $\delta_{\min} \in [0.06, 0.20]$, and greedy decoding on one H20 GPU. Each cell averages five independently sampled projection subsets.

Model	Benchmark	Metric	Vanilla	SEKA	PRISM- Δ	PRISM- ΔV
Qwen3-4B-Instruct	BiasBios	Accuracy	82.14	91.56	93.10	92.84
Qwen3-4B-Instruct	CounterFact	Efficacy	58.42	99.18	99.32	99.16
Qwen3-4B-Instruct	Pronoun Change	P. Score	96.52	97.24	97.86	97.94
Qwen3-8B	BiasBios	Accuracy	83.46	91.28	92.54	92.58
Qwen3-8B	CounterFact	Efficacy	54.76	99.22	99.38	99.12
Qwen3-8B	Pronoun Change	P. Score	98.86	99.12	99.54	99.62
Gemma3-4B-it	BiasBios	Accuracy	91.24	93.18	93.72	93.46
Gemma3-4B-it	CounterFact	Efficacy	68.36	98.42	98.56	98.47
Gemma3-4B-it	Pronoun Change	P. Score	78.46	89.72	93.18	93.86

Table 11: Instruction-tuned results, averaged over five projection subsets. All values are percentages.

As shown in Table 11, PRISM- Δ exceeds SEKA in all nine reported cells. A one-sided cell-level sign test gives $p = 0.002$, although model–benchmark cells are not independent experimental replicates. The largest relative gain is 3.9% on Gemma3-4B-it Pronoun Change. On that rewriting benchmark, PRISM- ΔV exceeds PRISM- Δ on all three models; broader task coverage would be needed to attribute this difference generally to content-channel steering.

H Automatic span selection

To evaluate PRISM- Δ without oracle highlighting, we replace the gold span S by a span $\hat{S}$ selected by BM25 (Robertson & Zaragoza, 2009), Contriever (Izacard et al., 2022), or BGE-M3 (Chen et al., 2024). We report span recall $|\hat{S} \cap S^*|/|\hat{S}|$ on BiasBios and Lost-in-the-Middle. Projection P_K , head weights $w_{\ell,h}^K$, and gain g_K are held fixed across selectors, so only the highlighted span changes. Each row averages three seeds.

For the three tested selectors (Table 12), higher observed span recall coincides with higher accuracy, and every reported result remains above Vanilla. BGE-M3 recovers 88% of the oracle-span gain on BiasBios and 81% on Lost-in-the-Middle. Because selector identity, selected content, and recall change together, this comparison does not isolate recall as the causal variable.

Benchmark	Selector	Recall	Vanilla	PRISM- Δ	Gain
BiasBios	BM25	0.58	79.80	86.42	+6.62
	Contriever	0.68	79.80	89.14	+9.34
	BGE-M3	0.76	79.80	90.86	+11.06
	Oracle span	1.00	79.80	92.38	+12.58
Lost-in-the-Middle	BM25	0.48	45.47	47.86	+2.39
	Contriever	0.62	45.47	49.42	+3.95
	BGE-M3	0.72	45.47	50.28	+4.81
	Oracle span	1.00	45.47	51.43	+5.96

Table 12: End-to-end accuracy with automatic span selection on QWEN3-4B, averaged over three seeds.

I Qualitative evaluation and extended channel decomposition

We use Claude Opus 4.6 as a zero-shot judge to rate factuality (faithfulness to highlighted information) and coherence (fluency and grammaticality) on a 1–5 Likert scale. On the pooled set of 50 human-annotated samples used for this check, judge and human scores have Pearson correlation $r = 0.87$. This limited validation does not establish reliability for other tasks, judges, or scoring criteria.

Method	BiasBios / QWEN3-4B		Pronoun Change / GEMMA3-4B	
	Factuality	Coherence	Factuality	Coherence
Vanilla	3.12	4.28	2.45	4.15
SEKA	3.68	3.15	3.82	3.22
PRISM- Δ	3.82	3.78	4.25	3.68
PRISM- Δ V	3.81	4.05	4.32	3.85

Table 13: LLM-as-Judge evaluation on a 1–5 scale. On the pooled 50-sample validation set, judge and human scores have Pearson $r = 0.87$.

In the two settings in Table 13, PRISM- Δ improves the judge’s factuality score over SEKA, while PRISM- Δ V moves coherence toward Vanilla. Table 14 separately reports metric-based channel decompositions for five selected model–task settings.

Model	Task	Metric	Vanilla	SEKA	PRISM- Δ	PRISM- Δ -V	PRISM- Δ V
QWEN3-4B	BiasBios	Accuracy	79.800	90.920	92.380	82.440	92.360
		Fluency	4.638	3.681	4.134	4.666	4.116
		Fluency cost	0.000	0.957	0.504	−0.028	0.522
GEMMA3-4B	BiasBios	Accuracy	89.900	92.420	92.900	90.120	91.640
		Fluency	4.850	3.920	4.410	4.902	4.398
		Fluency cost	0.000	0.930	0.440	−0.052	0.452
QWEN3-14B	BiasBios	Accuracy	85.100	90.280	91.680	88.920	91.200
		Fluency	4.712	3.854	4.318	4.736	4.302
		Fluency cost	0.000	0.858	0.394	−0.024	0.410
QWEN3-4B	Pronoun Change	P. Score	93.140	95.180	96.180	94.020	96.060
		Fluency	3.635	2.956	3.210	3.682	3.198
		Fluency cost	0.000	0.679	0.425	−0.047	0.437
GEMMA3-4B	Pronoun Change	P. Score	41.340	81.530	89.080	52.360	90.160
		Fluency	3.457	2.845	3.102	3.512	3.089
		Fluency cost	0.000	0.612	0.355	−0.055	0.368

Table 14: Extended dual-channel decomposition. Fluency cost is Vanilla fluency minus method fluency.

Value-only steering matches or exceeds Vanilla fluency in these five settings, while PRISM- Δ incurs 46–63% of SEKA’s fluency cost. On GEMMA3-4B Pronoun Change, PRISM- Δ V

exceeds PRISM- Δ by 1.08 points. This isolated result is compatible with a contribution from the cross term in Eq. 3, but broader interaction studies would be needed to establish when it occurs.

J Cross-domain projection transfer

We transfer projections bidirectionally between BiasBios and CounterFact while retaining the target domain’s optimal gain. No projection reconstruction or retraining is performed.

Model	Target	Vanilla	Task-specific	Transfer	Degradation
QWEN3-4B	CounterFact	45.00	99.14	94.86	−4.28
	BiasBios	79.80	92.38	87.42	−4.96
GEMMA3-4B	CounterFact	55.04	98.10	91.24	−6.86
	BiasBios	89.90	92.90	91.08	−1.82

Table 15: Cross-domain projection transfer. The target-domain gain is retained without rebuilding the projection. Degradation is measured from the task-specific projection.

Model	BiasBios→CounterFact	CounterFact→BiasBios	Average
QWEN3-4B	0.64	0.60	0.62
GEMMA3-4B	0.64	0.78	0.71

Table 16: Frobenius similarity between source- and target-domain projection subspaces.

Transferred projections remain above Vanilla in the four tested settings (Table 15), with degradation of 1.82–6.86 points relative to task-specific projections. Descriptively, the highest measured similarity (0.78, Gemma3-4B CounterFact→BiasBios; Table 16) coincides with the smallest degradation. With only four transfer settings, these values do not establish a stable correlation or causal explanation. They show that projection reuse can retain gains in the tested cases, while target-domain gain selection and task-specific construction remain necessary for the reported peak accuracy.

K Robustness to imperfect highlights

We replace 5–20% of tokens in the oracle highlighted span with irrelevant background tokens on QWEN3-4B BiasBios, using three seeds per noise level.

Noise ratio	Accuracy	$\Delta \downarrow$
0%	92.38	0.00
5%	91.95	−0.43
10%	91.42	−0.96
15%	90.68	−1.70
20%	89.75	−2.63

Table 17: Robustness to irrelevant tokens substituted into highlighted spans on QWEN3-4B BiasBios, averaged over three seeds. Δ is the accuracy change relative to the noise-free setting.

In this QWEN3-4B BiasBios perturbation (Table 17), 15% noise reduces accuracy by 1.70 points and 20% noise by 2.63 points. Appendix H evaluates a separate source of upstream error using three concrete selectors; neither experiment covers all possible span-boundary, omission, or retrieval errors.

Model	Layers	Heads	K mean	V mean	K/V ratio
QWEN3-4B	36	8	0.137	0.135	1.02
QWEN3-8B	36	8	0.129	0.174	0.74
QWEN3-14B	40	8	0.138	0.237	0.58
GEMMA3-4B	34	4	0.451	0.338	1.33
GEMMA3-12B	48	8	0.509	0.420	1.21

Table 18: Key and Value discriminative signal strength across models.

L Cross-model K/V complementarity

Table 18 and Figure 6 report Key and Value discriminative signal on BiasBios for five models. Both channels have nonzero measured signal, but their balance differs.

Qwen3 family. In these BiasBios measurements, the K/V ratio decreases with model size ($1.02 \rightarrow 0.74 \rightarrow 0.58$). On QWEN3-14B, the average Value signal is larger than the Key signal, and Figure 6 shows Value-dominant late layers.

Gemma3 family. Both measured models have larger absolute norm differences than the Qwen3 models (K mean: 0.45–0.51 vs. 0.13–0.14) and K/V ratios above 1.2 on BiasBios. These descriptive statistics do not by themselves determine the optimal steering gains.

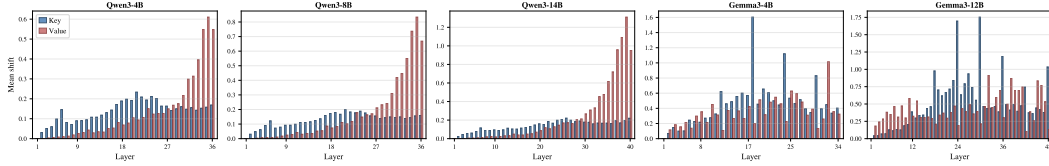


Figure 6: Layer-wise Key and Value discriminative signal on BiasBios for five models. In these measurements, Qwen3 models show increasing Value dominance in late layers, while Gemma3 models remain Key-dominant.

M Projection stability across random seeds

We construct five projections using different random subsets of 70 from 100 available QA pairs. The subsets overlap by approximately 70%, so this experiment measures sensitivity to partial data replacement rather than fully independent training sets. All other hyperparameters are fixed ($\gamma=0.998$, $\delta_{\min}=0.08$, $g_K=0.40$).

Seed	QWEN3-4B	QWEN3-8B
1	92.22	89.08
2	92.12	89.34
3	92.08	88.98
4	92.14	89.30
5	92.12	89.34
Mean $\pm$ Std	92.14 $\pm$ 0.05	89.21 $\pm$ 0.15

Table 19: Projection sensitivity across overlapping random subsets (BiasBios, Top-1 Accuracy %).

Across these five overlapping subsets (Table 19), the standard deviation is 0.05% on QWEN3-4B and 0.15% on QWEN3-8B. Both are smaller than the corresponding PRISM- Δ -SEKA gaps in this BiasBios experiment (+1.46 and +0.88 points). This check does not quantify projection variability for the other models or benchmarks.

N Statistical reliability details

All reported evaluations use greedy decoding, so generation is deterministic for a fixed projection and configuration. Appendix M isolates one source of variability—partial replacement of projection data—on BiasBios with QWEN3-4B and QWEN3-8B. It does not provide an upper bound for every model–benchmark cell or for hyperparameter-selection uncertainty.

Across the 15 base-model cells in Table 2, PRISM- Δ matches or exceeds SEKA in 14 (one-sided sign test after excluding the tie: $p < 0.001$). It is strictly higher on all five Pronoun Change P. Score cells ($p = 0.031$) and on four of five CounterFact Efficacy cells, with one tie. These cell-level sign tests summarize directional consistency; model and task cells are not statistically independent replicates.

O CounterFact ablation

Configuration	Projection	Weighting	Efficacy (%)
PRISM- Δ	Differential	Softplus	99.14
Indep. + softplus	Independent	Softplus	98.84
Diff. + uniform	Differential	Uniform	98.62

Table 20: Ablation on CounterFact with QWEN3-4B (Efficacy %, test set).

Differential projection contributes +0.30% (Table 20), computed as the gap between PRISM- Δ at 99.14% and the independent-projection variant at 98.84%. Softplus weighting contributes +0.52%, measured as the gap between PRISM- Δ and the uniform-weight variant at 98.62%. These are qualitatively consistent with the BiasBios ablation in Table 3.

The smaller absolute contributions on this QWEN3-4B CounterFact setting occur near the performance ceiling. The component ordering matches the QWEN3-4B BiasBios ablation, but one additional model–task setting is insufficient to establish generality.

P Data quantity ablation

Synthetic samples	Top-1 Accuracy (%)
50	91.92
100 (default)	92.38
200	92.40

Table 21: Effect of synthetic data quantity on PRISM- Δ for QWEN3-4B on BiasBios.

On QWEN3-4B BiasBios (Table 21), 50 samples reach 91.92%, 100 reach 92.38%, and 200 reach 92.40%. This single model–task comparison motivates the default of 100 samples for the reported setup but does not establish the same saturation point elsewhere.

Q Direction consistency

	Ω^+ (independent)	Ω_Δ (differential)	Random baseline
Cross-head similarity (mean)	0.254	0.068	0.071
Cross-head similarity (early layers)	0.353	0.053	0.071

Table 22: Cross-head cosine similarity of top singular vectors on QWEN3-4B.

To quantify whether different heads learn independent or redundant directions, we compute the mean pairwise absolute cosine similarity between the top left singular vectors of

all heads in QWEN3-4B. A random baseline is computed by averaging cosine similarities between uniformly sampled unit vectors in $\mathbb{R}^{128}$, yielding 0.071.

In this QWEN3-4B analysis, Ω^+ directions have cross-head similarity 0.254, compared with 0.071 for the random baseline, while Ω_Δ directions reach 0.068. This pattern is consistent with differential decomposition removing directions shared across heads; cosine similarity alone does not identify the semantic content of those directions.

In early layers, the measured values are 0.353 for Ω^+ and 0.053 for Ω_Δ . Whether this pattern transfers to other models and tasks remains to be tested.

R Hyperparameter sensitivity

We sweep each hyperparameter independently while holding others at their optimal values (QWEN3-4B, BiasBios).

g_K sensitivity. On QWEN3-4B BiasBios, performance varies by less than 0.5 points for tested $g_K \in [0.38, 0.50]$ and declines outside that interval.

$\delta_{\min}$ sensitivity. For the same setting, performance varies by less than 0.45 points over the tested interval $\delta_{\min} \in [0.06, 0.10]$.

γ sensitivity. For the same setting, performance varies by less than 0.1 points over the tested interval $\gamma \in [0.990, 0.998]$.

S Head weight and per-sample analysis

Head weight distribution. Figure 7 compares softplus weighting with hard thresholding. Hard thresholding shuts off 108 heads entirely, while softplus assigns weak-signal heads reduced but nonzero weights.

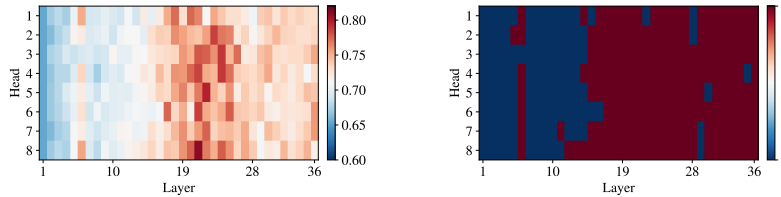


Figure 7: Head weight heatmaps. **Left:** PRISM- Δ softplus weights vary continuously across heads. **Right:** SEKA hard thresholding deactivates 108 heads, including 90% of early-layer heads.

T Case study

We identify 154 samples (33.9% of SEKA’s 454 errors on BiasBios with QWEN3-4B) where SEKA predicts incorrectly but PRISM- Δ succeeds. Figure 8 reveals that PRISM- Δ rescues 154 samples while losing only 81, for a net gain of 73 samples (+1.46% accuracy). Representative examples:

Success cases. The following examples illustrate contexts where SEKA fails but PRISM- Δ succeeds:

- “PhD, Adjunct Assistant Professor in Psychiatry...” → SEKA: psychologist; PRISM- Δ : **professor** ✓
- “assistant professor of psychology at NUS...” → SEKA: psychologist; PRISM- Δ : **professor** ✓

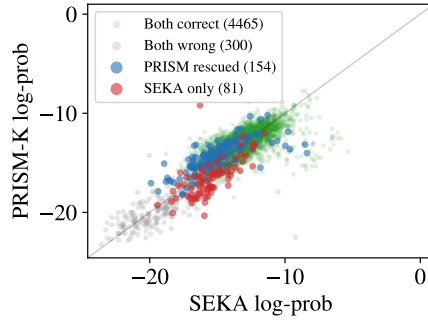


Figure 8: Per-sample target log-probability comparison on BiasBios with QWEN3-4B. PRISM- Δ rescues 154 samples that SEKA misses and loses 81, for a net gain of 73.

- “Performance Architect at Cisco...” → SEKA: software engineer; PRISM- Δ : **architect** ✓
- “aspiring filmmaker...shoots promotional videos...” → SEKA: photographer; PRISM- Δ : **film-maker** ✓
- “is a Maine comedian...” → SEKA: photographer; PRISM- Δ : **comedian** ✓

In many rescued examples, the occupation keyword co-occurs with a related domain word or company name. For instance, “Psychiatry” and “psychology” accompany cases where SEKA predicts “psychologist” despite the biography describing a professor; “Cisco” accompanies a “software engineer” prediction despite “Architect” appearing explicitly. These examples are compatible with the shared-feature account motivating differential projection, but they do not directly identify the internal cause of either prediction.

Failure analysis. Among the 81 QWEN3-4B BiasBios samples where PRISM- Δ is wrong but SEKA succeeds, the biographies tend to contain short or ambiguous occupation descriptions. In this paired comparison, PRISM- Δ rescues 154 samples and loses 81 relative to SEKA; this count does not characterize failure modes on other models or tasks.

U LLM Usage

GPT-4o was used to generate the 100 synthetic contrastive QA pairs for projection construction (Appendix F). Claude Opus 4.6 (claude-opus-4-6) was used as the qualitative judge reported in Appendix I; its ratings were validated against human annotations. LLM-based tools were also used for language polishing and copy editing. The authors verified all experimental results, citations, and scientific claims.