DiPer: Disentangled Personal Direction Discovery for Personalized Generation Steering

Anonymous ACL submission

Abstract

Large Language Model (LLM) personalization aims to adapt generation to individual users by capturing implicit preferences from historical interactions. Steering-based methods offer an inference-efficient solution, but existing work follows a difference-based paradigm that estimates personalized steering directions from representation offsets between contrastive historical examples, thus entangling inherent user preference with preference-agnostic contextual bias. This limitation arises because offset-based estimation uses raw contrastive variation, making the derived direction confounded by contextual bias rather than identifying preference-discriminative variation between user-conditioned and generic behavior. To address this, we propose **Disentangled Personal Direction Discovery (DiPer)**, which formulates personalized steering vector estimation as a discriminative direction discovery problem. DiPer learns layer-wise user-specific directions through supervised metric learning, based on Recursive Feature Machine (Radhakrishnan et al., 2024), over historical activations, to discover preference-discriminative directions while suppressing contextual bias. Theoretical analysis and extensive experiments demonstrate DiPer's efficacy and robustness.

1 Introduction

Large Language Model (LLM) personalization is essential for improving user experience in real-world generation applications (Zhang et al., 2024c; Liu et al., 2025), especially for capturing implicit preferences reflected in historical interactions (Zhao et al., 2026a; Guan et al., 2025). Existing personalization can be broadly divided into non-parameterized approaches, such as role prompting and in-context learning, and parameterized approaches which encode user-specific information into model internal states (Zhang et al., 2024a; Zhu et al., 2024) . Compared with non-parameterized methods, parameterized personalization provides a more generalizable and inference-efficient path(Zhu et al., 2024; Cho et al., 2024) and falls into tuning-based and steering-based.

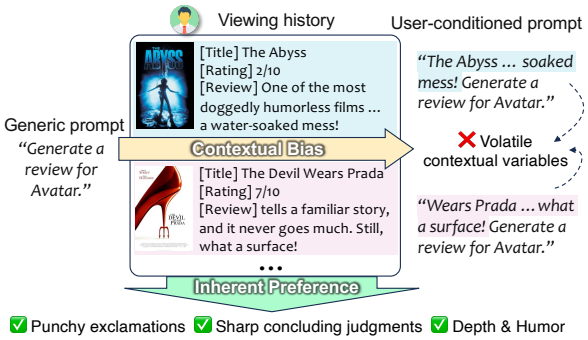

Figure 1: The entanglement challenge in personalized generation steering.

Tuning-based methods intervene in LLM parameter space by converting user histories into modules such as soft prompts or LoRA adapters (Tan et al., 2024b; Zhao et al., 2026b). Yet these modules require non-trivial parameter overhead that results in high computation and storage costs, steering-based methods therefore offer a cost-effective alternative by simply adding a personalized steering vector, derived from contrastive behavioral examples, to the residual stream at all token positions during inference (Bo et al., 2025; Cao et al., 2024).

Current steering work (He et al., 2025; Zhang et al., 2025a), either amplifies distributional shifts induced by user history or averages activation differences between personalized and generic responses, follows a difference-based paradigm (Rimsky et al., 2023) that estimates personalized steering directions from representation offsets between contrastive historical examples. This paradigm, however, is vulnerable in personalized generation, where historical examples inevitably mix preference signals with preference-agnostic bias (Zhao et al., 2025b), *i.e.*, volatile contextual variables,

such as factual content and prompt patterns. The obtained bias-contaminated direction thus undermines the generation quality.

As illustrated in Figure 1, historical contrastive prompts contain two orthogonal factors: volatile contextual variables that vary across samples, and the desired inherent preference that persists across the entire user history and determines steering effectiveness. Since raw offsets fail to distinguish stable preference from volatile variables, the estimated direction is contaminated by contextual bias. To extract directions that better reflect user preference, we reframe steering vector estimation as a discriminative direction discovery problem, where contrastive representations provide supervision. The resulting direction is expected to align with preference-discriminative variation while reducing the influence of contextual bias.

To this end, we propose **Disentangled Personal Direction Discovery (DiPer)**, which learns from historical activations to extract directions that better reflect user preference. Specifically, DiPer utilizes Recursive Feature Machine (RFM) (Radhakrishnan et al., 2024), a feature-learning algorithm that recursively learns a Mahalanobis metric from the gradient signals of a kernel predictor. Using user-conditioned and generic representations derived from history, RFM is applied to learn a metric that captures preference-relevant variation separating contrastive activations while reducing the influence of contextual bias. The personal direction is then extracted as the top eigenvector of the learned metric matrix, corresponding to the dominant direction of preference-relevant variation under this metric. We further extend the metric update with Hessian information to better capture preference patterns induced by joint variations in activation space, complementing the gradient signal used by standard RFM. The resulting layer-wise directions are injected into the corresponding hidden states during inference, steering generation toward the target user's preference.

Theoretical analysis shows that DiPer yields a lower preference-direction estimation error than diff-based estimation as historical samples increase. Extensive experiments validate the efficacy of DiPer, while analyses of layer-wise steering, recursive learning, and Hessian extension further verify that it captures robust preference-discriminative directions. Our main contributions are threefold:

- We formulate personalized generation steering as a discriminative direction discovery problem, revealing the limitation of diff-based estimation under context-biased contrastive representations.
- We propose **DiPer**, a supervised approach that learns layer-wise personal directions from contrastive activations, with a higher-order extension based on Hessian-informed metric updates.
- We provide theoretical and empirical evidence for DiPer, showing its efficacy, and robustness.

2 Related Work

2.1 Parameterized Personalization for LLMs

A common approach personalizes LLMs via user-specific parameterized objects like soft prompts or LoRA modules (Hu et al., 2022; Zhang et al., 2024b, 2025b). These PEFT methods encode user preferences into learnable model parameters for long-term personalization (Zhong et al., 2021; Tan et al., 2024b,a). However, maintaining distinct parameters per user raises scalability challenges as the user base grows, motivating steering-based personalization that applies vector interventions at inference time without parameter updates. The key question thus shifts from training user-specific parameters to constructing accurate, inherent, and generalizable steering vectors from user history.

2.2 Steering for Personalized Generation

Steering primarily controls general model behavior by manipulating directions in latent space. ActAdd (Turner et al., 2023) and UCS (Beaglehole et al., 2025) construct or select internal directions to steer concepts (*e.g.*, refusal, sentiment, truthfulness) without updating model parameters. Recent work extends this paradigm to personalized generation, where user preference is treated as a high-level latent attribute that can be steered at inference time.

Following Contrastive Activation Addition (CAA) (Rimsky et al., 2023), existing personalized steering methods are difference-based, constructing a user-specific direction by contrasting user-conditioned representations with neutral or user-agnostic ones. Context Steering (ContS) (He et al., 2025) compares next-token distributions with and without the entire user history and amplifies the difference during decoding, but treating the whole historical context as a personalization signal may conflate inherent user preference with topic, content, and context-induced bias. Forward-pass activation steering methods, represented by StyleVec-

tor (SV) (Zhang et al., 2025a), compute hidden-state differences between user-authored responses and style-agnostic responses generated from the same history samples, and inject the averaged difference into a Transformer block. It still estimates the user direction as a raw displacement between two activation sets, thus difficult to distinguish dimensions related to user preference from those associated with contextual variation.

This entanglement is related to the heterogeneous nature of user histories, which contain preference-agnostic volatile variables such as factual content and prompt scaffolds (Zhao et al., 2025b; Ye et al., 2024). SteerX (Zhao et al., 2025a) attempts to separate preference-related signals from contextual content by using causal inference to identify preference-driven tokens and rewrite them into coherent descriptions. However, the subsequent steering stage still relies on difference-based backbones, leaving latent activation-level entanglement insufficiently addressed.

A concurrent submission (Anonymous, 2026) proposes a meta-learning framework for cross-user/domain generalization via training-time parameter adaptation, while our work orthogonally focuses on tuning-free inference-time steering which disentangles user preference from contextual bias.

3 Problem Formulation

We study personalization through latent steering, where a frozen LLM G_θ is controlled by adding user-specific direction v to its internal representation. Let $H = \{(x_i, y_i)\}_{i=1}^n$ denote the historical interactions of a user, and let x be a test-time input. The goal is to generate $y \sim G_\theta^v(\cdot \mid x)$, where the output reflects the user's preference while all model parameters remain fixed.

Latent steering constructs pairs of positive and negative prompts of a user behavior: $v = \Phi(\mathcal{A}(p^+), \mathcal{A}(p^-))$, where p^+/p^- denotes the presence/absence of same-user preference evidence for the target user.$\mathcal{A}(\cdot)$ denotes the representation used for steering, and Φ extracts a direction from the contrast. At inference, the direction is injected into latent state with strength s: $\widetilde{\mathcal{A}}(x) \leftarrow \mathcal{A}(x) + sv$.

Bias-contaminated direction estimation. The key challenge lies in estimating v, which should capture the user's inherent preference rather than arbitrary differences between user-conditioned and generic representations. Since contrastive observations for direction estimation is contaminated with preference-agnostic contextual bias, the direction estimation problem should be viewed as recovering an inherent preference direction from confounded contrastive observations. Assume that the population means of the two conditions can be decomposed as $\begin{array}{c}\mathbb{E}[\mathcal{A}(p^+)] = \alpha u + \mathbb{E}[B^+], \\ \mathbb{E}[\mathcal{A}(p^-)] = -\alpha u + \mathbb{E}[B^-],\end{array}$ where $u \in \mathbb{R}^d$ is the inherent preference direction, $\alpha > 0$ denotes the average preference separation between the two conditions, and B^+, B^- denote residual preference-agnostic contextual biases, such as topic variation or template artifacts, after removing the component aligned with u. A difference-based estimator gives $\Delta_{\text{diff}} = \mathbb{E}[\mathcal{A}(p^+)] - \mathbb{E}[\mathcal{A}(p^-)] = 2\alpha u + \Delta_B$, where $\Delta_B = \mathbb{E}[B^+] - \mathbb{E}[B^-]$, and $u^\top \Delta_B = 0$.[1] Thus, the estimated direction is contaminated with preference-agnostic bias.

Proposition. The angular error between the difference-based estimator and the inherent preference direction satisfies

$$\sin \angle(\Delta_{\text{diff}}, u) = \frac{\|\Delta_B\|}{\sqrt{4\alpha^2 + \|\Delta_B\|^2}}.$$

Therefore, $\Delta_{\text{diff}} \propto u$ only when $\Delta_B = 0$.

4 Methodology

The problem formulation suggests that the inherent preference direction u is entangled with preference-agnostic contextual variation. To address this issue, we propose **Disentangled Personal Direction Discovery (DiPer)**, which realizes discriminative direction discovery in activation space.

4.1 Motivation & Overview

As illustrated in Figure 2, DiPer comprises three steps. (1) We construct contrastive samples from user historical interactions, where user-conditioned prompts serve as positive samples and generic prompts serve as negative samples. This contrast provides a supervised signal for distinguishing personalized behavior from generic behavior. (2) For each layer of the frozen LLM, we learn a user-specific direction that separates personalized activations from generic ones. Instead of directly averaging activation offsets, DiPer uses RFM to learn a task-adaptive Mahalanobis metric from a supervised contrastive predictor. The learned metric assigns larger weights to activation-space directions that consistently support this distinction,

[1] This orthogonality follows from the definition of B as the residual component after projecting out the preference u.

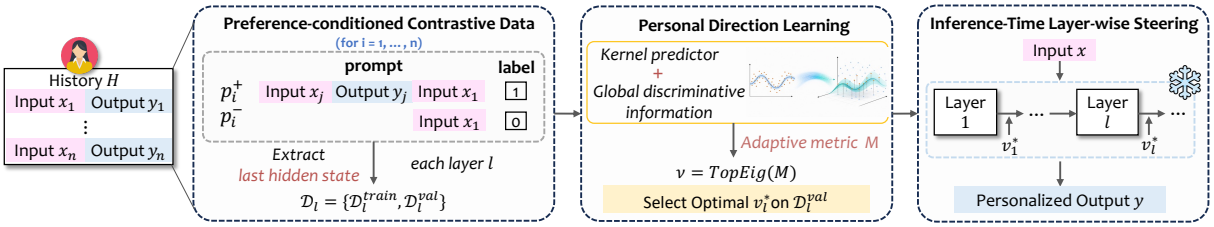

Figure 2: Overview of the proposed Disentangled Personal Direction Discovery (DiPer).

while reducing the influence of preference-agnostic contextual variation. We further enhance RFM with a Hessian-based term to capture preference evidence associated with joint changes across activation dimensions, complementing the gradient signal used by standard RFM. (3) Finally, the learned directions are injected into the frozen LLM during inference to steer generation toward the user's preference.

4.2 Preference-Conditioned Contrastive Construction

Given H, we construct contrastive prompts for each user using only that user's own history, aiming to reduce obvious text-level confounding before activation-space direction learning. For each current input x_i, we keep x_i identical in both the user-conditioned and user-agnostic prompts, so that the target task content is shared. We retrieve one relevant historical interaction (x_j, y_j) from the same user, where $j \neq i$. This reduces unrelated topical variation and excessive heterogeneity while preserving local evidence of the user's preference.

The user-conditioned prompt is $p_i^+ = \mathrm{Concat}([x_j; y_j], x_i)$, where $[x_j; y_j]$ denotes the retrieved input-output pair formatted as user history, followed by x_i. The user-agnostic counterpart keeps the same scaffold but removes the user-specific content: $p_i^- = \mathrm{Concat}([\varnothing], x_i)$. Labels $z(p_i^+) = 1$ and $z(p_i^-) = 0$ indicate whether the current input is conditioned on the target user's preference evidence. To prevent the contrast from becoming trivially separable by prompt format, p^- uses the same history slot and instruction scaffold as p^+, with the user-specific history replaced by an empty placeholder. Thus, DiPer is encouraged to rely on activation patterns induced by user-specific evidence rather than superficial differences.

For each contrastive pair (p_i^+, p_i^-), we extract the last-token hidden states at layer ℓ: $x_{i,\ell}^+ = \mathcal{A}_\ell(p_i^+)$, $x_{i,\ell}^- = \mathcal{A}_\ell(p_i^-)$, where $x_{i,\ell}^+, x_{i,\ell}^- \in \mathbb{R}^d$. We form $\mathcal{D}_\ell = \{(x_{i,\ell}^+, 1), (x_{i,\ell}^-, 0)\}_{i=1}^n$, with $X_\ell \in \mathbb{R}^{2n \times d}$ and $z \in \{0,1\}^{2n}$. This layer-wise construction preserves personalization signals across layers.

4.3 Personal Direction Learning

DiPer learns a supervised activation-space metric and extracts directions that remain discriminative under held-out contrastive validation.

For each user and layer, we split $\mathcal{D}_\ell$ into $\mathcal{D}_\ell^{\text{train}} = (X_\ell^{\text{train}}, z_\ell^{\text{train}})$, $\mathcal{D}_\ell^{\text{val}} = (X_\ell^{\text{val}}, z_\ell^{\text{val}})$ with an 80/20 ratio. The training split fits the kernel predictor and updates the metric, while the validation split only selects the best direction.

Given D_ℓ, DiPer learns [2] a user-specific metric M_ℓ and steering direction v_ℓ. For clarity, we omit the layer subscript ℓ. Starting from $M^{(0)} = I$, we repeat the following steps for T iterations.

Step 1: Fit a kernel predictor. At iteration t, a Mahalanobis Laplacian kernel is defined under $M^{(t)}$: $k_t(x_a, x_b) = \exp\left(-\frac{d_t(x_a, x_b)}{\sigma}\right)$, where $d_t(x_a, x_b) = \sqrt{(x_a - x_b)^\top M^{(t)} (x_a - x_b)}$. Let K_t be the kernel matrix with $(K_t)_{ab} = k_t(x_a, x_b)$. Then solve KRR with a regularization term $\lambda = 10^{-3}$: $\eta_t = (K_t + \lambda I)^{-1} z$, and obtain $f_t(x) = \sum_{a=1}^{|\mathcal{D}^{\text{train}}|} \eta_{t,a} k_t(x_a, x)$. Since f_t predicts whether an activation is user-conditioned or generic, personalization-relevant directions are those along which f_t changes with the contrastive label.

Step 2: Compute first-order sensitivity. RFM computes the input gradient $g_a^{(t)} = \nabla_a f_t(x_a)$, and form the Average Gradient Outer Product (AGOP): $M_G^{(t)} = \frac{1}{|\mathcal{D}^{\text{train}}|} \sum_{a=1}^{|\mathcal{D}^{\text{train}}|} g_a^{(t)} g_a^{(t)\top}$. This first-order metric emphasizes directions to which the contrastive predictor is most sensitive.

Step 3: Compute Hessian term. While the gradient captures the directions along which the con-

[2] For each user and layer, we split $\mathcal{D}_\ell$ into $\mathcal{D}_\ell^{\text{train}} = (X_\ell^{\text{train}}, z^{\text{train}})$, $\mathcal{D}_\ell^{\text{val}} = (X_\ell^{\text{val}}, z^{\text{val}})$ with an 80/20 ratio. The training split fits the kernel predictor and updates the metric, while the validation split only selects the best direction.

trastive predictor changes most rapidly at each sample, it only accounts for the first-order term in the local Taylor expansion. This can be insufficient for personalization, because user preference is often entangled with topic, style, and intent, causing first-order gradients from different history samples to be unstable, dispersed, or even partially canceled. In such cases, the personalization signal may not be fully characterized by the local slope of the predictor, but by how this sensitivity itself changes under activation perturbations. Therefore, DiPer introduces a Hessian-based term to capture the second-order response of the contrastive predictor. Directions with large Hessian energy indicate that the personalized-versus-generic score changes nonlinearly along those directions, revealing preference-relevant interactions that first-order AGOP would miss. Thus, we compute the Hessian of the predictor[3]: $H_a^{(t)} = \nabla_x^2 f_t(x_a)$. The second-order metric is defined as $M_H^{(t)} = \frac{1}{|\mathcal{D}^{\text{train}}|} \sum_{a=1}^{|\mathcal{D}^{\text{train}}|} H_a^{(t)\top} H_a^{(t)}$.

Step 4: Update the metric. We normalize the first- and second-order metrics and blend them: $M^{(t+1)} = (1-\beta)\overline{M}_G^{(t)} + \beta\overline{M}_H^{(t)}$. Here, $\overline{M}_G^{(t)}$ and $\overline{M}_H^{(t)}$ denote normalized matrices, and $\beta \in [0,1]$ controls the contribution of second-order information. This update reshapes the activation geometry so that directions predictive of personalization labels are amplified in later iters.

Step 5: Select the personal direction. After each metric update, we extract the leading eigenvector as a candidate direction: $v^{(t+1)} = \text{TopEig}(M^{(t+1)})$. Here, $\text{TopEig}(\cdot)$ returns the eigenvector associated with the largest eigenvalue under the standard unit-norm eigenvector convention. Thus, the learned object is directly used as a unit-norm direction. DiPer normalizes each selected direction to unit norm and uses a global steering strength during inference, avoiding additional per-user strength tuning. The candidate direction is selected by Pearson correlation with labels on the held-out validation set: $v^\star = \arg\max_{v^{(t)}} \text{corr}\left(X^{\text{val}} v^{(t)},\ z^{\text{val}}\right)$.

4.4 Inference-Time Layer-wise Steering

After direction learning, DiPer obtains a set of layer-wise personal directions $V = \{v_\ell^\star\}_{\ell \in \mathcal{L}}$, where $\mathcal{L} = \{1, \ldots, L_m - 1\}$, with L_m denoting the number of Transformer layers. During generation, the frozen LLM is modified by forward hooks. At layer ℓ, DiPer intervenes on each input token t with steering strength s: $\widetilde{A}_{\ell,t}(x) \leftarrow A_{\ell,t}(x) + s v_\ell^\star$.

Theorem 1. *Let $\widehat{M}$ be the empirical metric selected by the held-out validation in Step 5, and let $v^\star = \text{TopEig}(\widehat{M})$ be its leading eigenvector. $M^\star$ denotes the corresponding population metric, whose leading eigenvector is the inherent preference direction u.* [4] *Assume that $M^\star$ has eigengap $\gamma = \lambda_1(M^\star) - \lambda_2(M^\star) > 0$. Suppose that, with probability at least $1-\delta$, the empirical metric satisfies $\|\widehat{M} - M^\star\|_2 \le C\sqrt{\frac{r+\log(1/\delta)}{|\mathcal{D}^{\text{train}}|}}$, where r is the effective dimension and C depends on the boundedness and smoothness of the kernel predictor. Then, by the Davis–Kahan theorem,*

$$\sin\angle(v^\star, u) \le \frac{C}{\gamma}\sqrt{\frac{r+\log(1/\delta)}{|\mathcal{D}^{\text{train}}|}}.$$

Moreover, combining this bound with Proposition, DiPer admits a smaller angular-error bound than the difference-based estimator whenever

$$|\mathcal{D}^{\text{train}}| > \frac{C^2\big(r+\log(1/\delta)\big)}{\gamma^2} \cdot \frac{4\alpha^2 + \|\Delta_B\|^2}{\|\Delta_B\|^2}.$$

The theorem shows that the difference-based estimator is limited by an irreducible contextual-bias term, whereas DiPer can consistently approach the inherent preference direction as the number of contrastive samples increases.

5 Experiments

In this section, we conduct extensive experiments to answer the following research questions:

- **RQ1:** How does our proposed DiPer perform on the real-world personalized text generation tasks compared to baseline methods?
- **RQ2:** How do the key designs of DiPer contribute to performance?
- **RQ3:** Does DiPer discover more preference-discriminative personal directions than simple contrastive difference vectors?
- **RQ4:** Is DiPer robust across different training iterations and base LLM scales? (Appendix C.1)

5.1 Experimental Setup

Datasets. Following prior studies on personalized text generation, we conduct experiments on

[3] For the Mahalanobis Laplacian kernel, the kernel derivatives are analytic; Appendix A derives the efficient second-order metric form used in our implementation.

[4] This is built on the preference-subspace property analyzed in Appendix A.1 and emperically supported in Sec 5.3.

two distinct tasks from LongLaMP (Kumar et al., 2024): Post Writing on Social Media and Product Review Generation.

Baselines. We compare DiPer with both non-personalized and personalized baselines. (1)**Vanilla** denotes the non-personalized setting, where no user history is provided to the model. (2)**RAG** (Salemi et al., 2024) retrieves relevant historical samples from the user's history as contextual evidence for generation. (3) For tuning-based methods, we include **OPPU** (Tan et al., 2024b), which learns user-specific parameter-efficient modules on a task-tuned base LLM, and two LoRA-based variants: **LoRA-C** and **LoRA-P**. Specifically, LoRA-C learns a collective LoRA module shared across users, while LoRA-P trains personalized LoRA modules for each user. (4) For tuning-free steering, we include **StyleVector** (Zhang et al., 2025a) and **Context Steering (ContS)** (He et al., 2025), two representative steering methods based on contrastive hidden-state differences and logit differences, respectively. We further include **SteerX-SV** and **SteerX-ContS**, which apply SteerX (Zhao et al., 2025a) to the two steering backbones by extracting preference-driven components from user histories. Implementation details and evaluation metrics are reported in Appendix B.

5.2 Main Results (RQ1)

We first evaluate the overall effectiveness of DiPer. Results in Table 1 lead to three observations.

DiPer consistently outperforms existing baselines. Best performance across all metrics indicates that DiPer improves not only lexical matching but also alignment with user preferences. This demonstrates that our proposed mechanism can provide effective personalization signals for generation.

Latent steering provides larger margins over parameter tuning. We attribute this to the limited and heterogeneous per-user supervision in personalized generation task. OPPU and LoRA-P require learning user-specific parameters from sparse histories, which can overfit to local topics or templates and generalize poorly to held-out inputs. LoRA-C is more stable for sharing parameters across users, but this also limits its ability to capture individual preferences. In contrast, steering-based methods estimate lightweight inference-time control signals, which is more suitable when each user has sparse supervision. This highlights the practical advantage of latent steering.

DiPer further improves steering by discovering disentangled directions. Existing diff-based steering baselines entangle inherent user preferences with contextual bias and noisy historical signals. By learning a supervised metric over contrastive historical activations, DiPer identifies directions that are more preference-sensitive and less dominated by preference-agnostic contextual bias.

5.3 Ablation Studies (RQ2)

We compare DiPer with several variants: (1) *w/o Derivative-term*, which alters the derivative update for direction learning; (2) *Positive-construction variants*, which incorporate either top-k relevant or randomly sampled histories. (3) *Other-Eigen*, which uses lower-ranked eigenvectors of M rather than the top eigenvector.

Effectiveness of derivative components. We first examine the contribution of the two derivative-based components. The variant w/o Hessian removes the second-order update ($\beta = 0$) and keeps only the first-order metric, while w/o AGOP removes the first-order update ($\beta = 1$) and keeps only the Hessian-based metric. As shown in Table 2, both variants remain competitive but consistently underperform the full DiPer on two tasks. Results indicate that the derivative-based components provide complementary signals: AGOP captures directions where the contrastive predictor changes locally, while the Hessian term refines the metric with directions where this sensitivity varies across the activation space. Their combination yields the most reliable preference-sensitive direction.

Effect of positive sample construction. We vary the historical interaction used to construct the positive condition. As shown in Table 3, Top-1 achieves the best, while using more histories does not bring consistent improvements. This suggests that the most relevant historical interaction already provides sufficient personalization evidence, while adding histories may introduce redundant context.

We further compare Top-1 with Random-1, where the positive prompt is constructed from a randomly sampled interaction from the same user's history. Random-1 still uses only user-specific information and remains competitive, showing that DiPer can extract useful preference signals from noisy same-user histories. However, its clear drop from Top-1 indicates that random histories introduce more preference-agnostic topical and contex-

Method			Post Writing for Social Media					Product Review Generation				
Category	Type	Name	BLEU	MET.	R-1	BERT	*Avg.* ↑	BLEU	MET.	R-1	BERT	*Avg.* ↑
Non-personalized	–	Vanilla	1.110	0.165	0.238	0.445	–	1.358	0.160	0.313	0.453	–
Personalized	Non-parameterized	RAG	2.558	0.191	0.253	0.459	*38.91%*	2.569	0.186	0.347	0.477	*30.40%*
	Tuning-based	OPPU	0.729	0.109	0.154	0.439	*-26.23%*	1.670	0.140	0.224	0.477	*-3.17%*
		LoRA-C	1.149	0.159	0.250	0.452	*1.62%*	1.013	0.117	0.198	0.458	*-21.98%*
		LoRA-P	1.256	0.176	0.257	0.448	*7.12%*	1.734	0.147	0.252	0.463	*0.57%*
	Steering-based	StyleVector	2.591	0.180	0.268	0.461	*39.68%*	2.901	0.220	0.271	0.494	*36.69%*
		ContS	3.032	0.209	0.263	0.460	*53.42%*	2.496	0.217	0.315	0.484	*31.73%*
		SteerX-SV	2.811	0.181	0.271	0.458	*44.93%*	5.155	0.230	0.292	**0.506**	*82.09%*
		SteerX-ContS	3.210	0.209	0.266	0.462	*57.86%*	3.110	0.217	0.365	0.494	*47.57%*
		DiPer	**7.448**	**0.220**	**0.313**	**0.482**	***160.98%***	**5.699**	**0.241**	**0.377**	**0.506**	***100.61%***

Table 1: Performance comparison of baselines and DiPer on two personalized text generation tasks. Best results are in **bold**. $Avg. \uparrow$ indicates the average improvement of each method over Vanilla ($p < 0.05$).

Method	BLEU	MET.	R-1	BERT	*Avg.* ↑
Post Writing on Social Media					
DiPer	**7.448**	**0.220**	**0.313**	**0.482**	***160.98%***
w/o Hessian	6.969	0.209	0.304	0.476	*147.30%*
w/o AGOP	6.269	0.202	0.301	0.470	*129.82%*
Product Review Generation					
DiPer	**5.699**	**0.241**	**0.377**	**0.506**	***100.61%***
w/o Hessian	5.298	0.241	0.371	0.503	*92.58%*
w/o AGOP	5.145	0.237	0.375	0.504	*89.52%*

Table 2: Ablation of derivative terms in DiPer.

Setting	BLEU	MET.	R-1	BERT	*Avg.* ↑
Top-1	7.448	**0.220**	**0.313**	**0.482**	***160.98%***
Top-2	7.356	0.215	0.308	0.480	*157.57%*
Top-3	**7.460**	0.211	0.309	0.480	*159.41%*
Top-4	7.304	0.214	0.307	0.480	*156.14%*
Random-1	6.872	0.200	0.298	0.466	*142.56%*

Table 3: Effect of positive sample construction on Post Writing. Top-k uses the k most relevant histories, while Random-1 uses a randomly sampled one.

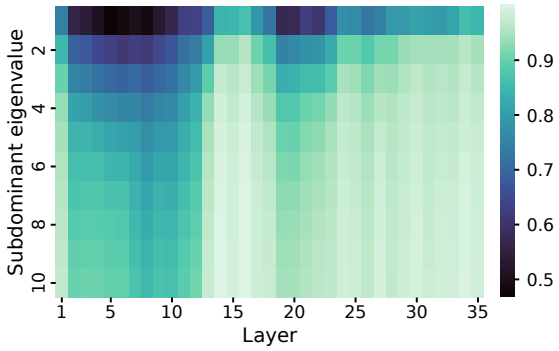

Figure 3: Layer-wise relative eigengap of the learned metric matrix M. Each cell shows the relative gap between the dominant eigenvalue and a subdominant eigenvalue, averaged over users. Larger values indicate stronger spectral separation.

tual variation. These results support our design choice: using a relevant history reduces unnecessary topical heterogeneity at the prompt level, and DiPer further suppresses the remaining activation-space entanglement through supervised direction discovery. Therefore, we adopt the Top-1 setting as it provides the best overall trade-off between performance, simplicity, and bias control.

Why use the top eigenvector? Finally, we analyze the spectrum of the learned metric matrix M to justify the use of its top eigenvector as the steering direction. Since M is positive semi-definite, $v^\top M v$ measures the discriminative energy assigned to a unit direction v under the learned activation geometry. The top eigenvector therefore gives the single direction with the largest preference-sensitive energy.

Figure 3 shows that the leading eigenvalue is clearly separated from the remaining spectrum: the relative gap between the largest eigenvalue and the top-10 subdominant eigenvalues is 89.4% on average. This indicates that lower-ranked directions carry weaker discriminative energy under the learned metric. Rather than assuming personalization is strictly one-dimensional, this result shows that the learned metric contains a dominant direction that compactly summarizes the main contrastive variation, supporting our use of the top eigenvector for layer-wise steering.

5.4 In-depth Analysis (RQ3)

We conduct a probing study to examine whether the learned directions capture stable user-specific signals that generalize beyond the sampled histories. For each user, we split the historical profile into two parts, using the first 90% for direction estima-

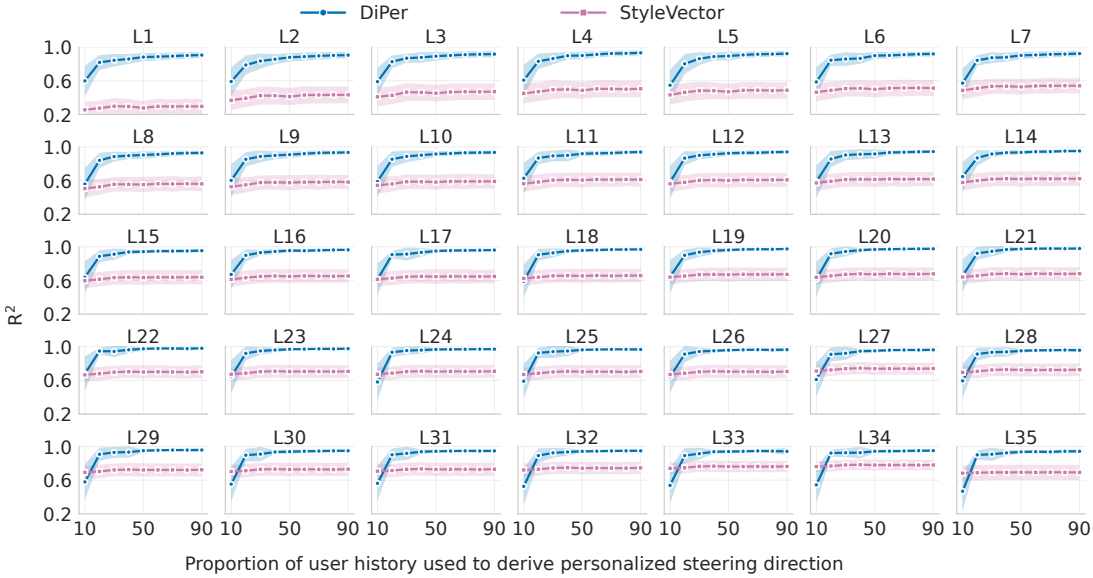

Figure 4: Direction quality analysis across layers. Results are averaged over two tasks.

tion and the remaining 10% for evaluation. From the estimation split, we randomly sample different proportions of histories and learn a layer-wise direction for each method. On the held-out split, we extract positive and negative activations from user-authentic contexts and user-agnostic counterparts, project them onto the learned direction, and compute R^2 between the projected scores and labels. We repeat this process across users and sampled subsets, and report the mean with error bands.

Figure 4 shows that DiPer consistently achieves higher R^2 than StyleVector, a representative contrastive hidden-state difference method, across almost all layers and history ratios. DiPer improves rapidly with about 20% of histories and converges to a high plateau as more histories are used, while its error bands shrink with increasing history size. In contrast, StyleVector remains at a much lower level and shows limited improvement. These results suggest that simple contrastive averaging is more affected by contextual bias, whereas DiPer learns more stable user-specific directions that generalize to held-out histories.

Since the evaluation activations are constructed from samples unseen during direction learning, the higher held-out R^2 of DiPer also indicates that its directions capture reusable user-specific signals rather than memorizing local prompt artifacts. This empirical trend is consistent with Theorem 1, that the supervised direction selected by DiPer has a smaller preference-direction estimation error than difference-based estimators when the number of historical samples is sufficiently large. The decreasing variance and increasing held-out R^2 further provide empirical evidence for this sample-efficiency claim, showing that DiPer converges toward a stable preference-sensitive directions rather than amplifying context-dependent offsets. The layer-wise pattern further justifies our per-layer design: personalization signals are not confined to a single block, as several layers show consistently high distinguishing power while early layers are less stable.

6 Conclusion

In this work, we tackle the critical challenge of disentangled preference modeling for personalized generation steering. We prove that difference-based steering estimators produce bias-contaminated directions because raw contrastive offsets entangle inherent user preference with preference-agnostic contextual bias. To address this, we propose **DiPer**, which formulates personal steering vector estimation as discriminative direction discovery in activation space. DiPer learns layer-wise user-specific metrics from contrastive historical activations, extracts preference-discriminative directions through held-out validation, and incorporates a Hessian-informed extension to capture higher-order preference-discriminative variation. Theoretical analysis and extensive experiments demonstrate that DiPer achieves lower direction-estimation error and consistently improves personalized generation over existing baselines, offering valuable insights for effective tuning-free personalization.

Limitations

DiPer provides an efficient tuning-free approach for personalized generation steering, but several aspects remain to be further studied. First, because DiPer learns user-specific directions from historical interactions, its effectiveness depends on the availability and quality of user history; in extremely sparse settings, the learned direction may be less reliable. The "cold-start" problem for users with no or negligible interaction history falls outside the scope of this work and point to promising directions for future research. Second, our current design extracts a single top-eigenvector direction for each layer independently and uses fixed steering coefficients, leaving top-k eigenvector-based subspace steering, adaptive layer selection, and context-dependent steering strength as future work. Finally, although our experiments cover multiple personalized generation settings, extending DiPer to handle temporal preference drift and multi-turn interactive steering will be a key focus of our subsequent work.

References

Anonymous. 2026. Mcl: Meta-collaborative learning for generalized personalized dialogue generation. *Under review*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Daniel Beaglehole, Adityanarayanan Radhakrishnan, Enric Boix-Adserà, and Misha Belkin. 2025. Toward universal steering and monitoring of ai models.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. *ArXiv*, abs/2306.03819.

Jessica Y Bo, Tianyu Xu, Ishan Chatterjee, Katrina Passarella-Ward, Achin Kulshrestha, and D Shin. 2025. Steerable chatbots: Personalizing llms with preference-based activation steering. *ArXiv*, abs/2505.04260.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *ArXiv*, abs/2406.00045.

Chuanqi Cheng, Quan Tu, Wei Wu, Shuo Shang, Cunli Mao, Zhengtao Yu, and Rui Yan. 2024. "in-dialogues we learn": Towards personalized dialogue without pre-defined profiles through in-dialogue learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10422, Miami, Florida, USA. Association for Computational Linguistics.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. 2024. Heterogeneous lora for federated fine-tuning of on-device foundation models. In *EMNLP 2024*.

Jian Guan, Jun Wu, Jing Li, Chen Cheng, and Wei Wu. 2025. Think-while-generating: On-the-fly reasoning for personalized long-form generation. *arXiv preprint arXiv:2512.06690*.

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L. Schrum, and Anca Dragan. 2025. Context steering: Controllable personalization at inference time. In *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, and Hamed Zamani. 2024. Longlamp: A benchmark for personalized long-form text generation. *ArXiv*, abs/2407.11016.

Paul Pu Liang, Irene Z Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis philippe Morency. 2020. Towards debiasing sentence representations. In *Annual Meeting of the Association for Computational Linguistics*.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.

Jiahong Liu, Wenhao Yu, Quanyu Dai, Zhongyang Li, Jieming Zhu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2026. Perfit: Exploring personalization shifts in representation space of LLMs. In *The Fourteenth International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Misha Belkin. 2024. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383:1461 – 1467.

Nina Rimsky, Nick Gabrieli, Julia Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *ArXiv*, abs/2312.06681.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024a. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6459–6475, Miami, Florida, USA. Association for Computational Linguistics.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024b. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491, Miami, Florida, USA. Association for Computational Linguistics.

Alex Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *Preprint*, arXiv:2308.10248.

Ziang Ye, Zhenru Zhang, Yang Zhang, Jianxin Ma, Junyang Lin, and Fuli Feng. 2024. Disentangling reasoning tokens and boilerplate tokens for language model fine-tuning. In *Annual Meeting of the Association for Computational Linguistics*.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025a. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141, Vienna, Austria. Association for Computational Linguistics.

Kai Zhang, Yejin Kim, and Xiaozhong Liu. 2025b. Personalized llm response generation with parameterized memory injection. *Preprint*, arXiv:2404.03565.

Kai Zhang, Lizhi Qing, Yangyang Kang, and Xiaozhong Liu. 2024a. Personalized llm response generation with parameterized user memory injection. *arXiv preprint arXiv:2404.03565*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2024b. Personalized lora for human-centered text understanding. *Preprint*, arXiv:2403.06208.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024c. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Xiaoyan Zhao, Ming Yan, Yilun Qiu, Haoting Ni, Yang Zhang, Fuli Feng, Hong Cheng, and Tat-Seng Chua. 2025a. SteerX: Disentangled steering for LLM personalization. *Preprint*, arXiv:2510.22256.

Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2025b. NextQuill: Causal preference modeling for enhancing LLM personalization. *Preprint*, arXiv:2506.02368.

Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2026a. Nextquill: Causal preference modeling for enhancing llm personalization. In *ICLR 2026*.

Ziyi Zhao, Chongming Gao, Yang Zhang, Haoyan Liu, Weinan Gan, Huifeng Guo, Yong Liu, and Fuli Feng. 2026b. Don't start over: A cost-effective framework for migrating personalized prompts between llms. *arXiv preprint arXiv:2601.12034*.

Wanjun Zhong, Duyu Tang, Jiahai Wang, Jian Yin, and Nan Duan. 2021. UserAdapter: Few-shot user learning in sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1484–1488, Online. Association for Computational Linguistics.

Jiachen Zhu, Jianghao Lin, Xinyi Dai, Bo Chen, Rong Shan, Jieming Zhu, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. Lifelong personalized low-rank adaptation of large language models for recommendation. *arXiv preprint arXiv:2408.03533*.

A Theoretical Insight and Computation of DiPer

We provide two details omitted from the main text due to space constraints. First, we show why the derivative-based metrics used by DiPer recover preference-discriminative directions while reducing the influence of preference-agnostic contextual bias. Second, we derive the analytic Hessian form used for the Mahalanobis Laplacian kernel predictor.

A.1 Preference-Subspace Deduction

Our orthogonal view follows a common geometric convention in representation analysis and concept erasure. Prior work (Belrose et al., 2023; Liang et al., 2020) represents a target concept or attribute as a linear direction or subspace, and treats concept-agnostic information as lying in the orthogonal complement or nullspace of that subspace. Recent work on LLM personalization further suggests that personalized information can be represented by a low-rank subspace in the model's representation space (Liu et al., 2026). Motivated by these observations, we use B to denote the residual preference-agnostic contextual component after projecting out the preference direction. Motivated by this low-rank view, we treat personalization signals as occupying a low-dimensional region of activation space and view DiPer as recovering a preference-discriminative subspace, from which a dominant user-specific steering direction can be extracted.

Formally, let $U \in \mathbb{R}^{d\times r}$ denote an orthonormal basis of the latent preference subspace, where $U^\top U = I_r$. Assume that the population predictor for the user-conditioned-versus-generic contrastive label depends on the activation only through this subspace:

$$f^\star(a) = \mathbb{E}[z \mid a] = \phi(U^\top a),$$

where $\phi : \mathbb{R}^r \to \mathbb{R}$ is differentiable. The first-order derivative is

$$\nabla f^\star(a) = U\nabla\phi(U^\top a).$$

Therefore, the population first-order metric satisfies

$$M_G^\star = \mathbb{E}\left[\nabla f^\star(a)\nabla f^\star(a)^\top\right] = UC_GU^\top,$$

where

$$C_G = \mathbb{E}\left[\nabla\phi(U^\top a)\nabla\phi(U^\top a)^\top\right].$$

Similarly, the Hessian of the population predictor is

$$\nabla^2 f^\star(a) = U\nabla^2\phi(U^\top a)U^\top.$$

Thus, the population second-order metric becomes

$$M_H^\star = \mathbb{E}\left[(\nabla^2 f^\star(a))^\top\nabla^2 f^\star(a)\right] = UC_HU^\top,$$

where

$$C_H = \mathbb{E}\left[(\nabla^2\phi(U^\top a))^\top\nabla^2\phi(U^\top a)\right].$$

DiPer blends the two metrics:

$$M_{\mathrm{DiPer}}^\star = (1-\beta)M_G^\star + \beta M_H^\star.$$

Substituting the above forms gives

$$M_{\mathrm{DiPer}}^\star = UC_\beta U^\top, \quad C_\beta = (1-\beta)C_G + \beta C_H.$$

Therefore,

$$\mathrm{Range}(M_{\mathrm{DiPer}}^\star) \subseteq \mathrm{span}(U).$$

Consequently, for any preference-agnostic contextual direction b orthogonal to the preference subspace, i.e., $U^\top b = 0$, we have

$$b^\top M_{\mathrm{DiPer}}^\star b = 0.$$

This shows that, under the preference-subspace model, DiPer assigns zero population metric energy to preference-agnostic contextual directions that do not affect the user-conditioned-versus-generic contrastive label. In contrast, difference-based steering directly estimates a raw contrastive offset between two activation means and is therefore contaminated whenever preference-agnostic contextual bias differs across the two conditions.

The role of the second-order term is complementary to the first-order term. The first-order metric emphasizes directions along which the contrastive predictor changes locally, whereas the second-order metric emphasizes directions along which this local sensitivity itself changes. Therefore, the second-order term can capture higher-order preference-discriminative variation that appears through curved decision geometry rather than a globally linear offset.

A.2 Analytic Hessian for the Mahalanobis Laplacian Kernel

We derive the second-order metric used in DiPer. For readability, we omit the iteration index t and denote the current metric by M. Given training activations $\{x_i\}_{i=1}^n$, the kernel ridge predictor is $f(x) = \sum_{i=1}^n \eta_i k(x_i, x)$, where $\eta = (K + \lambda I)^{-1}z$, and the Mahalanobis Laplacian kernel is $k(x_i, x) = \exp\left(-\frac{d_i(x)}{L}\right)$, $d_i(x) = \sqrt{(x-x_i)^\top M(x-x_i)}$. Let $\delta_{ij} = x_j - x_i, q_{ij} = M\delta_{ij}, d_{ij} = \sqrt{\delta_{ij}^\top M\delta_{ij}}, k_{ij} = \exp(-d_{ij}/L)$. The gradient of the distance is

$$\nabla_x d_i(x)\big|_{x=x_j} = \frac{q_{ij}}{d_{ij}}.$$

Therefore, differentiating the Laplacian kernel gives

$$\nabla_x k(x_i, x)\big|_{x=x_j} = -\frac{k_{ij}}{Ld_{ij}} q_{ij}.$$

Taking the second derivative gives the Hessian structure

$$\nabla_x^2 k(x_i, x)\big|_{x=x_j} = k_{ij}\left[\frac{q_{ij}q_{ij}^\top}{L^2 d_{ij}^2} - \frac{M}{Ld_{ij}} + \frac{q_{ij}q_{ij}^\top}{Ld_{ij}^3}\right].$$

The two rank-one terms share the same direction q_{ij}, while the middle term contributes a metric-aligned component. For numerical stability and efficiency, DiPer uses a tractable implementation form that retains the rank-one curvature direction and the metric-aligned component:

$$\widetilde{\nabla_x^2 k}(x_i, x_j) = \frac{k_{ij}}{L^2 d_{ij}^2} q_{ij}q_{ij}^\top - \frac{k_{ij}}{Ld_{ij}} M.$$

The additional rank-one term $\frac{k_{ij}}{Ld_{ij}^3} q_{ij}q_{ij}^\top$ is omitted because it shares the same direction and can be numerically unstable for small distances. When $d_{ij} = 0$, the corresponding ratio terms are set to zero, which is equivalent to excluding the singular self-kernel contribution.

The rank-one part is accumulated through its generating vector. Define

$$A_j = \sum_{i=1}^{n} \eta_i \frac{k_{ij}}{d_{ij}^2} q_{ij}.$$

Since $q_{ij} = Mx_j - Mx_i$, this can be computed efficiently as

$$A_j = \left(\sum_{i=1}^{n} \eta_i \frac{k_{ij}}{d_{ij}^2}\right) Mx_j - \sum_{i=1}^{n} \eta_i \frac{k_{ij}}{d_{ij}^2} Mx_i.$$

This corresponds to the matrix operation used in our implementation. The resulting rank-one curvature contribution is

$$M_{\text{rank}} = \frac{1}{nL^4} \sum_{j=1}^{n} A_j A_j^\top.$$

For the metric-aligned component, define

$$c_j = \sum_{i=1}^{n} \eta_i \frac{k_{ij}}{d_{ij}}.$$

The corresponding contribution is

$$M_{\text{align}} = \frac{1}{nL^2} \sum_{j=1}^{n} c_j^2 M^\top M.$$

The second-order metric used by DiPer is then

$$M_H = M_{\text{rank}} + M_{\text{align}}.$$

Finally, we symmetrize the metric before the update:

$$M_H \leftarrow \frac{1}{2}(M_H + M_H^\top).$$

This computation only differentiates the kernel predictor with respect to activation vectors and does not require backpropagation through the frozen language model.

B Experiment Details

Implementation Details. Consistent with standard protocol, 100 users with the richest interaction histories are selected to construct testsets. For all methods, we use the same base LLM, Qwen3-8B[5], and evaluate them on the same user split. For retrieval-related methods, top-2 historical samples are retrieved by Contriever (Izacard et al., 2022) according to their relevance to the current input and used as user context for fair comparison. For tuning-based methods, we use the same training histories as personalization supervision and set the training epochs to 3 and LoRA rank to 8. For all baselines, we follow the hyperparameter protocols reported in their original papers. For DiPer, we set the kernel bandwidth $\sigma \in \{1, 10, 100\}$ and $\beta \in \{0.25, 0.5, 0.75, 1\}$, which are automatically selected on $\mathcal{D}^{\text{val}}$. Personal directions are discovered layer by layer from the same user historical interactions after $T = 10$ iterations, with each positive sample constructed by concatenating the current input with one retrieved historical instance. We set the steering strenth $s = 1$.

For methods involving sampling, training, or validation-based direction selection, we run three independent trials with different random seeds and report the mean performance.

For automatic evaluation, we compute BLEU with the SacreBLEU toolkit, following its standardized tokenization and scoring protocol. For BERTScore, we use representations from led-base-16384 as the encoder to measure contextual semantic similarity between generated outputs and references.

Here, we clarify the distinction between DiPer and UCS (Beaglehole et al., 2025). UCS intuitively uses RFM to extract representations of

[5] https://huggingface.co/Qwen/Qwen3-8B

broad model-level concepts (*e.g.*, anti-refusal, deception, toxicity) for general concept steering and monitoring, which are shared across inputs and users. In contrast, DiPer targets personalization, where the relevant signal is an inherent user preference entangled with preference-agnostic contextual bias in historical activations. DiPer therefore constructs contrastive historical activations for each user, learns layer-wise user-specific metrics, selects user-specific directions, and extends RFM with Hessian-informed updates to capture higher-order preference-discriminative variation. Our theoretical analysis formalizes why difference-based steering estimators are biased under confounded contrastive observations and why supervised metric learning can better recover the inherent preference direction. Thus, DiPer provides a personalization-specific framework with a higher-order RFM extension and theoretical deduction for disentangled direction discovery.

Token intervention range. Following activation-steering practice (Rimsky et al., 2023), DiPer applies the learned layer-wise direction to all token representations during inference. The direction estimated from last-token activations represents a layer-level preference direction. Applying it to all tokens allows the preference signal to affect subsequent contextualization. Prior results (Zhang et al., 2025a) have shown that intervening on more tokens leads to stronger personalization gains.

Evaluation Metrics. We evaluate generation quality using four standard metrics widely adopted in personalized text generation: BLEU (Papineni et al., 2002), METEOR (MET.) (Banerjee and Lavie, 2005), ROUGE-1 (R-1) (Cheng et al., 2024), and BERTScore (BERT) (Zhang et al., 2019). These metrics measure lexical overlap, semantic matching, and contextual similarity between the generated output and the ground-truth user response. Higher scores indicate better personalized generation performance.

C Additional Analyses

C.1 Robustness Analysis (RQ4)

Iteration sensitivity. We further analyze the iteration at which the validation-selected best direction appears under a maximum budget of $T = 10$. As shown in Figure 5, most best directions are selected in early iterations, with the highest proportion appearing at iteration 2-4 and few selected at later iterations. This indicates that DiPer can usually discover effective user directions within a few recursive updates. Therefore, using a moderate iteration budget is sufficient for stable direction learning.

Figure 5: Distribution of the best iteration selected by validation under a maximum budget of $T = 10$.

Method	BLEU	MET.	R-1	BERT	*Avg.* ↑
Vanilla	1.823	0.195	0.264	0.450	–
RAG	2.474	0.222	0.289	0.470	*15.64%*
OPPU	0.921	0.122	0.161	0.440	*-32.15%*
LoRA-C	1.300	0.193	0.279	0.461	*-5.56%*
LoRA-P	1.470	0.197	0.273	0.443	*-4.28%*
StyleVector	3.192	0.212	0.306	0.468	*25.74%*
ContS	2.787	0.203	0.303	0.466	*18.65%*
SteerX-SV	3.023	0.205	0.301	0.468	*22.06%*
SteerX-ContS	2.729	0.208	0.304	0.464	*18.47%*
DiPer	**7.220**	**0.220**	**0.328**	**0.483**	***84.89%***

Table 4: Performance comparison of baselines and DiPer on Post Writing task with Qwen3-14B.

Generalizability across base LLMs. We further evaluate DiPer on Post Writing task with Qwen3-14B[6]. As shown in Table 4, DiPer still achieves the best overall performance and clearly outperforms baselines. This demonstrates that DiPer can robustly learn effective user steering directions across different base LLMs.

C.2 Sensitivity to Steering Scale

To examine whether the performance depends critically on steering scale, we conduct a sensitivity analysis on the post-writing task by multiplying the learned direction with a global scale s, while keeping all learned directions and other inference settings unchanged:

$$\widetilde{A}_\ell(x) \leftarrow A_\ell(x) + s v_\ell^\star.$$

We evaluate $s \in \{0.25, 0.5, 1.0, 1.5, 2.0\}$, where $s = 1.0$ is the default setting used in DiPer.

[6]https://huggingface.co/Qwen/Qwen3-14B

Scale s	BLEU	MET.	R-1	BERT	Avg. ↑
0.25	7.171	0.202	0.310	0.449	*149.90%*
0.50	7.230	0.218	0.311	0.480	*155.50%*
1.00	**7.448**	**0.220**	**0.313**	**0.482**	***160.98%***
1.50	7.387	0.216	0.311	0.481	*158.79%*
2.00	7.345	0.213	0.310	0.481	*157.29%*

Table 5: Sensitivity to fixed steering scale. We multiply the learned unit-norm direction by a global scale s during inference while keeping all learned directions unchanged. $s = 1.0$ is the default setting used by DiPer.

As shown in Table 5, DiPer remains effective across all tested fixed scales. The default scale $s = 1$ achieves the best performance across all metrics. Smaller scales still improve substantially but weaken the intervention, while larger scales remain competitive but slightly underperform the unit scale, suggesting mild over-steering.

C.3 Shortcut Analysis

A potential concern is that the contrastive labels may be solved by detecting the presence of a history slot or prompt-format artifacts, rather than user-specific preference signals. To address the concern, we introduce a format-only control experiment.

This variant keeps the same instruction scaffold and history slot as DiPer, but replaces the retrieved user history with generic content. We then apply the same direction-learning procedure and the same inference-time layer-wise steering as DiPer. Therefore, the only available signal in this control comes from the prompt format itself, rather than from user-specific historical content. If DiPer mainly benefits from a format shortcut, this control should achieve comparable performance to DiPer.

As shown in Table 6, the format-only control performs substantially worse than DiPer on both tasks. This indicates that preserving the history slot and prompt scaffold alone is insufficient to reproduce DiPer's gains, thus suggests that DiPer's gains are not attributable to superficial history-slot or prompt-format cues; rather, they depend on user-specific historical information captured in the activation space.

C.4 Efficiency Analysis

We further analyze the efficiency of DiPer from two practical perspectives: per-user pre-processing time and user-specific storage. Table 7 reports the empirical cost on Qwen3-8B, averaged over the two tasks. All methods are evaluated on one A100-SXM4-80GB GPU.

Method	BLEU	MET.	R-1	BERT	*Avg.* ↑
Post Writing on Social Media					
Vanilla	1.110	0.165	0.238	0.445	-
Format-only	1.780	0.164	0.280	0.459	*20.14%*
DiPer	**7.448**	**0.220**	**0.313**	**0.482**	***160.98%***
Product Review Generation					
Vanilla	1.358	0.160	0.313	0.453	-
Format-only	1.885	0.184	0.320	0.460	*14.40%*
DiPer	**5.699**	**0.241**	**0.377**	**0.506**	***100.61%***

Table 6: Shortcut analysis. The format-only control keeps the same prompt scaffold and history slot as DiPer, but replaces user-specific history with generic content.

DiPer is efficient in per-user pre-processing. Compared with tuning-based personalization methods, DiPer avoids user-specific gradient updates and only requires frozen-model activation extraction followed by lightweight direction discovery. As a result, DiPer requires 23.88 seconds per user, which is substantially faster than OPPU and LoRA-P. Specifically, DiPer reduces the per-user pre-processing time by 97.48% compared with OPPU and by 85.19% compared with LoRA-P. This shows that DiPer is suitable for scenarios where user profiles are frequently updated or new users need to be personalized at low cost.

DiPer also maintains compact user-specific storage. Unlike OPPU and LoRA-P, which store user-specific parameter modules, DiPer stores only the learned layer-wise steering directions after offline discovery. The intermediate metric used during direction learning is discarded after extracting the final directions. Consequently, DiPer requires only 0.35 MB per user, which is orders of magnitude smaller than OPPU and LoRA-P. Although DiPer stores more vectors than single-direction steering methods such as StyleVector and SteerX-SV, its storage remains below 1 MB per user and is negligible compared with adapter-based personalization.

These results indicate that DiPer preserves the deployment advantage of activation steering while improving direction quality. It pays a moderate offline cost to discover preference-sensitive layer-wise directions, but at deployment time the user representation remains a compact set of vectors rather than a trained parameter module.

D LLM-as-judge Evaluation

To complement the automatic metrics, we use a LLM-as-judge protocol to evaluate the personal-

Metric	PT/User (s)	SS/User (MB)
OPPU	947.25	768
LoRA-P	161.26	364
StyleVector	44.36	0.01
SteerX-SV	17.97	0.01
DiPer	23.88	0.35

Table 7: Empirical cost comparison, averaged over two tasks. PT/User denotes offline training or pre-processing time per user; SS/User denotes user-specific storage.

ization quality of generated responses. We sample 100 instances from test set. For each, the judge model GPT-4o is given the user history, the input query, and the candidate response. The judge assigns a score from 0 to 5, where a higher score indicates better personalization quality. We set the temperature of the LLM to 0 to reduce randomness.

The prompt used for evaluation is shown below.

> You are an impartial evaluator for personalized generation. Given a user history, an input query, and a candidate response, your task is to evaluate whether the candidate response is relevant to the query and consistent with the user's historical preferences.
> User history:
> {USER_HISTORY}
> Input query:
> {INPUT_QUERY}
> Candidate response:
> {CANDIDATE_RESPONSE}
> Please assign an overall score from 0 to 5.
> A higher score indicates that the response is more relevant, coherent, and better aligned with the user's preferences.
> Return your judgment in the following JSON format:
> {
> "score": [SCORE],
> "reason": "[BRIEF_EXPLANATION]"
> }

For each method, we report the average judge score over all instances in Table 8. DiPer obtains the highest judge score, with an average score of 4.2526, outperforming SteerX-SV and SteerX-ContS by 0.9318 and 0.6612, respectively. This result indicates that the responses generated by DiPer are judged to be more consistent with user histories while remaining relevant to the input query. Compared with the two difference-based steering baselines, the improvement suggests that learning preference-sensitive directions from historical activations provides stronger personalization than directly using representation or distribution differences as steering signals. These findings are consistent with the automatic evaluation results, further supporting the effectiveness of DiPer in improving personalization quality.

Method	Judge Score
SteerX-SV	3.3208
SteerX-ContS	3.5914
DiPer	**4.2526**

Table 8: LLM-as-judge evaluation results.

E Case Study

To examine whether the gains come from genuine personalization rather than surface overlap, we provide a case study. This example directly reflects our motivation: user histories contain both preference-relevant review habits and preference-agnostic product content. The user has distinctive profile-level patterns, including historical album framing, track-level evidence, calibrated critical tone, domain-specific music terminology, and signature-style attribution.

Several observations emerge from the comparison. First, DiPer captures structural markers that appear repeatedly in the user's profile, such as bracketed source-style attribution, while the baselines omit this pattern or copy the input description directly. Second, DiPer better calibrates the user's critical tone: the generated review is positive but qualified, using constructions such as "may not have received the attention it deserved" and "may not be a perfect record," whereas SteerContS produces a more uniformly positive and template-like review. Third, DiPer uses domain-appropriate music-review language, including references to soul, R&B, quiet storm, production, and track-level evidence, while preserving relevance to the current album.

These patterns suggest that DiPer does more than increase lexical overlap. By learning directions in activation space from contrastive user histories, it can transfer stable stylistic and preference cues while suppressing preference-agnostic product content. This supports our central claim that activation-space direction learning helps disentangle user preference signals from contextual variation.

Item	Text excerpt	Analysis
Input	Generate a review by a user who gives an overall rating of "4.0" to Margie Joseph's 1975 album *Margie*. The product description emphasizes Joseph's collaboration with Blue Magic on "What's Come Over Me," the album's modest commercial performance, Arif Mardin's production, covers of Carole King and Bill Withers songs, Joseph–Mardin originals such as "Sign of the Times" and "Stay Still," and concludes that *Margie* is a solid Atlantic release that deserved more attention. The provided summary is "Joseph's third for Atlantic re-finds the funk."	Current context that all methods preserve.
History	The user's historical reviews reveal stable personalization cues: historical framing, track-level evidence, nuanced star ratings, and signature-style attribution. Representative profile excerpts include: "Following his 1968 break with fellow Righteous Brother Bobby Hatfield..."; "Medley wrote or co-wrote four of the album's tracks..."; "Issued in 1980, and with band links to Jonathan Richman's Modern Lovers..."; and "4-1/2 stars, if allowed fractional ratings. [2006 hyperbolium dot com]".	Preference signal to learn from history.
RAG	"In 1975, Margie Joseph enjoyed one of her greatest triumphs when she teamed up with Blue Magic for a spectacular remake of the Philadelphia soul ballad 'What's Come Over Me.' ... Although *Margie* falls short of perfect, it is one of the singer's more consistent Atlantic dates and deserved a lot more attention than it received. $\sim$ Alex Henderson, Rovi"	Copies the product description instead of transforming it into the user's review style.
SteerSV	"In 1975, Margie Joseph enjoyed one of her greatest triumphs ... Although *Margie* falls short of perfect, it is one of the singer's more consistent Atlantic dates and deserved a lot more attention than it received. $\sim$ Alex Henderson, Rovi. Joseph's third for Atlantic re-finds the funk."	Preserves input content, but remains close to raw context and summary.
SteerContS	"This album is a great example of Joseph's growth as an artist and a reminder of the quality of music she was capable of producing during her tenure with Atlantic. Overall, *Margie* is a solid and enjoyable listen that deserves more recognition. It's a testament to Joseph's talent and a worthy addition to any collection of 1970s soul and R&B. I would rate this album 4.0 out of 5 stars."	Fluent but generic; the rating expression is explicit and template-like, unlike the user's nuanced review style.
SV	"The album's most notable track, 'Sign of the Times,' is a funky standout that highlights Joseph's growing confidence as a songwriter. The album also features a strong cover of 'What's Come Over Me,' which, while not included on the album, was a hit for Blue Magic in 1974. ... 4.0 stars."	Uses review-like wording, but repeats facts incorrectly and weakly captures the user's historical-review structure.
DiPer	"Margie Joseph's 1975 album *Margie* may not have received the attention it deserved at the time, but in hindsight, it stands as a solid and consistently enjoyable entry in her discography. Produced by the ever-reliable Arif Mardin, the album offers a blend of soul, R&B, and quiet storm sensibilities..."	Frames the album historically and gives a calibrated positive judgment, matching the user's review habit.
	"The album features a strong selection of material, including two Carole King songs – 'Believe in Humanity' and 'After All This Time' ... Bill Withers' 'The Same Love That Made Me Laugh' ... 'Sign of the Times' and the lush, quiet storm-infused 'Stay Still'..."	Transfers the user's track-level evidence style rather than only copying the description.
	"While *Margie* may not be a perfect record, it's one of Joseph's more consistent efforts for Atlantic ... for fans of 1970s soul and R&B, *Margie* is a rewarding listen that re-finds the funk and heart of Joseph's best work. **4.0/5 stars** [2025 hyperbolium dot com]"	Captures nuanced rating language and signature-style attribution from the user profile.

Table 9: Case study on Product Review Generation. The input provides the current product context, while the user profile reveals stable personalization cues. Baseline excerpts highlight failure modes aligned with our motivation: copying input context, relying on raw contextual offsets, or producing generic rating templates. DiPer better extracts profile-level review preferences and applies them to the current item.