UNLOCKING LATENT PERSONALIZATION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly expected to adapt to individual users, yet effective personalization remains challenging when only limited user-specific samples are available. In this work, we take an alternative perspective: pretrained LLMs may already possess latent capacity for personalization, and a few user samples may therefore suffice to guide the model toward user-aligned behavior with minimal user-specific adaptation. From this perspective, we propose *LatentPersonal*, a framework that formulates personalization as navigation in a shared latent adaptation space. *LatentPersonal* infers a compact latent representation from a few user samples to guide user-specific model adaptation, regularized with a variational information bottleneck to encourage compact preference representations. We instantiate *LatentPersonal* with LoRA, leveraging its low-rank parameterization as a natural low-dimensional adaptation space for personalization. By simply inserting a user-specific guidance vector between the shared low-rank factors, the model can navigate toward personalized adaptations through lightweight inference of this compact representation, without updating the shared LoRA parameters. Experiments across multiple personalization datasets demonstrate that *LatentPersonal* substantially reduces user-specific adaptation overhead while achieving effective personalization from only a few user-specific interactions, with particularly strong performance in the one-shot regime.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable success as general-purpose models and are increasingly being deployed in diverse real-world scenarios (Brown et al., 2020; Wei et al., 2021; Minaee et al., 2024), including personal assistants, domain-specific applications, and task-oriented intelligent systems. However, the transition from broadly capable models to user-centric systems introduces a fundamental challenge: a single pretrained model must accommodate heterogeneous preferences, behaviors, and requirements across users (Houlsby et al., 2019). While general-purpose LLMs provide strong foundational capabilities, effectively tailoring them to individual users remains challenging, particularly when only limited user-specific interactions are available, making direct adaptation both data-inefficient and prone to overfitting (Houlsby et al., 2019; Li & Liang, 2021).

Existing LLM personalization methods leverage user-specific information, including interaction histories, demonstrations, and feedback (Zollo et al., 2025; Aroca-Ouellette et al., 2025), to tailor model behavior to individual users via personalized prompting (Zhang et al., 2024), retrieval-based conditioning (Salemi et al., 2024), user representation learning (Liu et al., 2025), and parameter-efficient adaptation (Tan et al., 2024b). Despite their different mechanisms, these approaches largely follow a similar intuition: personalization is achieved by further tailoring model behavior using user-specific information. Yet large-scale pretraining has endowed LLMs with broad and diverse capabilities, many of which may already support personalized responses but are not necessarily expressed in ways that align with individual users. Nevertheless, many existing personalization methods still rely on limited user-specific data to learn user-aligned behavior anew, potentially relearning behaviors that may already be supported by the pretrained model.

This raises a natural question: does effective personalization necessarily require learning user-specific behavior anew from limited user information? We hypothesize that pretrained LLMs may already encode personalization-relevant behaviors that can be activated with limited user-specific evidence. This intuition is supported by Figure 1(a), where a human-constructed preference prompt substantially improves personalization in a frozen pretrained LLM without user-specific parame-

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

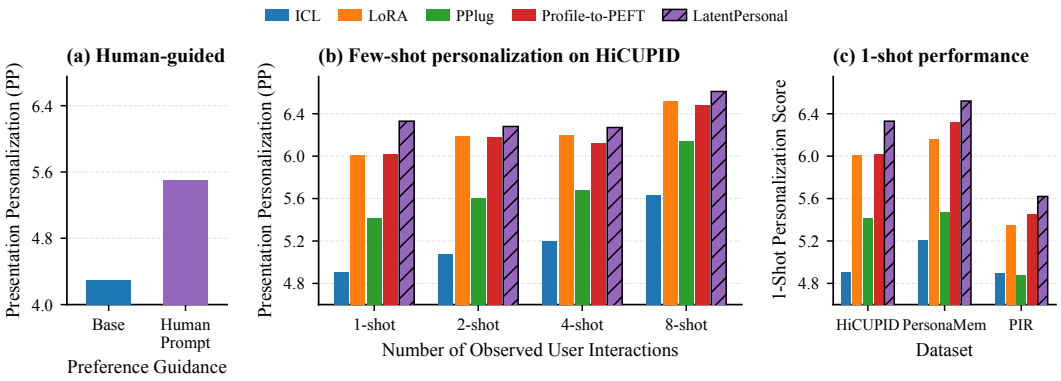


Figure 1: **Personalized behaviors can be elicited from pretrained LLMs with limited user-specific evidence. (a)** Providing a frozen pretrained LLM with a human-identified preference description derived from observed interactions improves personalization without user-specific parameter learning, suggesting that pretrained LLMs can exhibit personalization-relevant behaviors under appropriate guidance. **(b)** On HiCUPID, *LatentPersonal* achieves strong personalization with as few as one observed interaction and remains competitive as additional interactions become available. **(c)** Its strong one-shot performance generalizes across three representative personalization datasets.

ter updates. If the personalization-relevant variation underlying these behaviors can be effectively represented, the resulting representation can serve as guidance for eliciting corresponding model behavior. From this perspective, personalization can be viewed as navigating a general-purpose model toward user-aligned behavior, rather than learning a separate adaptation from scratch for each user.

Parameter-efficient fine-tuning (PEFT) has become a widely adopted paradigm for adapting pretrained LLMs (Houlsby et al., 2019; Hu et al., 2021; Li & Liang, 2021). Importantly, its parameter-efficient adaptation form provides a concrete basis for the navigation perspective above. Rather than modifying the pretrained model as a whole, PEFT introduces a parameter-efficient adaptation that can be written as

$$\theta = \theta_0 + \Delta\theta, \tag{1}$$

where θ_0 denotes the pretrained model and $\Delta\theta$ represents the adaptation applied to it. Building on this view, if personalization-relevant characteristics can be effectively represented within the space of model adaptations $\Delta\theta$, then only a few user-specific samples may be sufficient to locate an appropriate adaptation for a target user.

To operationalize this navigation perspective, we introduce *LatentPersonal*, a personalized adaptation framework that represents personalization-relevant variations in a shared low-dimensional latent space and uses user-specific interactions to navigate this space for personalized adaptation.

Our contributions can be summarized as follows:

- **i)** We introduce a new perspective on LLM personalization by formulating user adaptation as navigation within a shared latent preference space toward user-specific adaptations. Building on this perspective, we propose *LatentPersonal*, a latent-conditioned framework for personalized PEFT that learns compact user representations and uses them to navigate shared adaptation structures toward user-specific parameter updates.
- **ii)** We instantiate *LatentPersonal* across PEFT architectures, with a LoRA-based realization that directly integrates the user latent representation into the intermediate low-rank space, allowing the latent representation to modulate shared low-rank adaptation directions and thereby produce user-specific parameter updates within a shared LoRA structure. We further generalize this design to other PEFT architectures, including Adapter and Prompt Tuning.
- **iii)** We empirically demonstrate the effectiveness of *LatentPersonal* across multiple personalization datasets, PEFT architectures, and varying amounts of user-specific evidence. Our results show that *LatentPersonal* is particularly effective in low-shot regimes, achieving strong personalization from only a few observed interactions. Moreover, our analyses reveal consistent and user-specific adaptation patterns in the learned latent space, providing empirical support for the latent navigation perspective.

As illustrated in Figure 1(b,c), *LatentPersonal* achieves effective personalization from only a few observed user interactions, supporting our view of personalization as navigating toward user-aligned behavior. Code is available at: `https://anonymous.4open.science/r/LatentPersonal-25632`.

2 Related Work

Parameter-Efficient Fine-Tuning of LLMs. Parameter-efficient fine-tuning (PEFT) (Lialin et al., 2023) has emerged as an effective alternative to full fine-tuning by adapting pretrained models with only a small number of trainable parameters. Existing approaches realize parameter-efficient adaptation through different forms of parameterization. Adapters (Houlsby et al., 2019) introduce lightweight trainable modules into pretrained models, while Prefix-Tuning (Li & Liang, 2021) and Prompt Tuning (Lester et al., 2021) optimize continuous representations with the pretrained parameters kept frozen. Intrinsic-dimension methods instead explore efficient adaptation by restricting optimization to low-dimensional subspaces of the full parameter space (Aroca-Ouellette et al., 2025). LoRA (Hu et al., 2021) takes a different approach by parameterizing weight updates with low-rank decomposition and has become one of the most widely adopted PEFT methods for LLMs.

Personalized Adaptation of LLMs. Personalization has become increasingly important as LLMs are deployed to serve users with diverse preferences and needs. Early works such as LaMP (Salemi et al., 2024) incorporated user-specific information into the model context through personalized prompting and retrieval from user histories. Subsequent approaches moved beyond explicit retrieval by learning compact user representations from historical interactions to condition model generation (Ning et al., 2025; Liu et al., 2025; Cao et al., 2026). PerFit (Liu et al., 2026) further explores low-rank personalization shifts in the hidden representation space and performs user-specific adaptation through representation-level interventions. In parallel, parameter-efficient fine-tuning has enabled personalization directly through model parameters. OPPU (Tan et al., 2024b) learns a separate PEFT module for each user, while subsequent approaches such as Personalized Pieces (Tan et al., 2024a) and PROPER (Zhang et al., 2025) improve the scalability of personalized adaptation through parameter composition and progressive group-to-individual adaptation, respectively. More recently, Profile-to-PEFT (Tan et al., 2026) generates personalized PEFT parameters directly from user profiles through a hypernetwork, avoiding per-user optimization at deployment time.

Latent Representations for Model Adaptation. Latent-variable modeling provides a general mechanism for exposing low-dimensional structure underlying high-dimensional observations, and has become a fundamental modeling principle across representation learning (Bengio et al., 2013) and generative modeling (Kingma & Welling, 2013). Latent representations have also been widely explored in language models, including for language modeling (Fang et al., 2019), text generation (Hu et al., 2022), and controllable generation (Gu et al., 2023). Similar ideas have recently appeared in personalized adaptation. Profile-to-PEFT (Tan et al., 2026) encodes user profiles into compact representations and uses them to condition a hypernetwork that generates selected user-specific LoRA parameters. However, this formulation treats user adaptation primarily as a parameter prediction problem and does not explicitly characterize the shared low-dimensional structure underlying personalization across users.

3 Methodology

In this section, we introduce *LatentPersonal*, a framework for efficient personalization from limited user interactions. Our core intuition is that pretrained LLMs may already encode behavioral variations relevant to individual preferences, while these variations are not necessarily expressed for a particular user. We therefore model personalization as navigating user-relevant variations in a shared latent adaptation space, rather than independently optimizing a separate adaptation for each user.

We organize the methodology as follows. Section 3.1 formulates personalized adaptation as navigation in a shared low-dimensional latent preference space. Section 3.2 introduces an information-bottlenecked latent adaptation mechanism that infers compact user-specific representations from observed interactions and maps them to personalized model updates. Finally, Section 3.3 presents LoRA as a representative parameter-efficient instantiation of *LatentPersonal*.

3.1 Personalization as Latent-Space Navigation

We formulate personalized adaptation in a low-dimensional latent preference space. Given a pretrained LLM with parameters θ_0, each user u provides a set of interactions $D_u = \{(x_i, y_i)\}_{i=1}^{K}$, where x_i denotes the input, y_i denotes the corresponding user-specific target response, and K is the number of available interaction samples. The objective is to obtain a personalized model that adapts the pretrained model to the user's preferences.

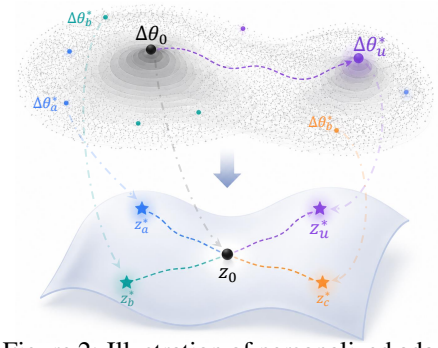


Figure 2: Illustration of personalized adaptation as navigation in the latent preference space of *LatentPersonal*.

Instead of independently learning user-specific adaptations, we model personalization-relevant variations through a shared low-dimensional latent preference space. Specifically, we introduce a latent variable

$$z_u \in \mathbb{R}^d, \tag{2}$$

where d is substantially smaller than the dimensionality of the parameter adaptation, and z_u encodes the preference characteristics inferred from the user's observed interactions. This formulation is motivated by the intuition that personalization primarily involves capturing user-specific variations while largely preserving the pretrained model's general capabilities.

Under this formulation, personalized adaptation is expressed through a latent-to-parameter mapping:

$$\theta_u = \theta_0 + g(z_u), \tag{3}$$

where $g(\cdot)$ maps the latent preference representation to the corresponding model adaptation. Personalization can therefore be formulated as locating an optimal user-specific representation in the latent preference space:

$$z_u^* = \arg\min_z \mathcal{L}(D_u, \theta_0 + g(z)). \tag{4}$$

This formulation provides a natural interpretation of personalization as navigation: the observed user interactions guide the model toward an appropriate location z_u^* in the shared latent preference space, which in turn determines the corresponding personalized adaptation through $g(\cdot)$.

3.2 Information-Bottlenecked Latent Adaptation

Given the latent-space formulation above, the remaining challenge is to infer a compact preference representation z_u from the observed interactions D_u and map it into an effective personalized adaptation. *LatentPersonal* models this process through a latent-conditioned mapping:

$$\Delta\theta_u = g_\phi(z_u), \tag{5}$$

where $g_\phi(\cdot)$ denotes a shared latent-to-adaptation mapping learned across users. Rather than optimizing independent adaptation parameters for each user, this shared mapping produces user-specific adaptations conditioned on their latent preference representations.

Adaptation-sufficient preference representation. During training, for each user u, we partition the available interactions into a support set D_u^{sup} and a disjoint query set D_u^{qry}. Although the support interactions D_u^{sup} provide evidence about user preferences, they also contain interaction-specific variations, such as task content and local contextual details, that are not necessarily relevant to personalization. Meanwhile, the pretrained LLM encodes broad linguistic and reasoning capabilities in a high-dimensional parameter space. In contrast, user-specific preferences correspond to only a limited subset of the model behaviors that need to be adapted.

To obtain a compact preference representation, we introduce an information bottleneck (IB) (Tishby et al., 2000) over the latent variable z_u. The bottleneck constrains the information encoded in z_u to focus on the factors relevant to personalized adaptation. Let $\Delta\theta_u^*$ denote the optimal user-specific adaptation for user u. Ideally, z_u should retain all information in D_u^{sup} that is relevant to $\Delta\theta_u^*$. An ideal representation would satisfy the following adaptation-sufficiency criterion:

$$I(D_u^{\text{sup}}; \Delta\theta_u^* \mid z_u) = 0. \tag{6}$$

Among such adaptation-sufficient representations, we further prefer compact representations that discard interaction-specific information unnecessary for personalization.

Motivated by this ideal representation, we introduce a learnable preference encoder $q_\psi(z_u \mid D_u^{\text{sup}})$ to characterize the latent representation inferred from user interactions. We seek to maximize the information retained by z_u about the desired user-specific adaptation, while simultaneously restricting its dependence on the observed interactions. This encourages the latent variable to capture adaptation-relevant preference factors rather than indiscriminately encoding the interaction history. This formulation serves as a design principle rather than an exact characterization of the objective optimized in practice, since $\Delta\theta_u^*$ is neither observed nor uniquely defined. Motivated by this ideal, we use the following information-bottleneck formulation as a conceptual objective:

$$\max_{\psi} \quad \underbrace{I(z_u; \Delta\theta_u^*)}_{\text{adaptation preservation}} \quad -\beta \quad \underbrace{I(z_u; D_u^{\text{sup}})}_{\text{interaction compression}} \quad , \tag{7}$$

where $\beta > 0$ controls the strength of the information bottleneck and balances adaptation preservation against interaction compression.

Direct optimization of Eq. (7) is intractable in practice, as the optimal user-specific adaptation $\Delta\theta_u^*$ is not directly available and both mutual information terms are difficult to evaluate. We therefore derive tractable surrogate objectives for the two terms below.

Adaptation preservation objective. We first consider the adaptation preservation term $I(z_u; \Delta\theta_u^*)$, which encourages the latent preference representation to preserve information relevant to personalized adaptation. However, the optimal adaptation $\Delta\theta_u^*$ is not directly available during training, making direct supervision or explicit estimation of this mutual information infeasible. We therefore exploit the functional role of the latent representation established above: if z_u preserves the information required for personalized adaptation, the adaptation induced by z_u should enable the personalized model exhibit user-specific behaviors on held-out interactions from the same user.

To operationalize this criterion, for each training user u, we use the support interaction set D_u^{sup} to infer a compact latent preference representation through the preference encoder:

$$z_u \sim q_\psi(z_u \mid D_u^{\text{sup}}), \tag{8}$$

The resulting representation is then translated into a user-specific adaptation through the shared latent-to-adaptation mapping:

$$\Delta\theta_u = g_\phi(z_u), \tag{9}$$

where g_ϕ is the shared latent-to-adaptation mapping introduced above.

Rather than matching the unavailable $\Delta\theta_u^*$ directly, we optimize the induced personalized model on the query interactions:

$$\mathcal{L}_{\text{adapt}} = \mathbb{E}_{z_u \sim q_\psi(z_u \mid D_u^{\text{sup}})} \left[\mathcal{L}\left(D_u^{\text{qry}}; \theta_0 + g_\phi(z_u)\right) \right]. \tag{10}$$

Here, $\mathcal{L}(D_u^{\text{qry}}; \theta_u)$ denotes the task loss under the personalized model, and $\mathcal{L}_{\text{adapt}}$ serves as the adaptation-preservation objective. Minimizing $\mathcal{L}_{\text{adapt}}$ encourages the latent representation inferred from D_u^{sup} to preserve information that is useful for inducing personalized behavior reflected in the observed interactions. We use this task loss as a tractable surrogate for the ideal adaptation-preservation term in Eq. 7, rather than as a formal bound on $I(z_u; \Delta\theta_u^*)$.

Interaction compression objective. While the adaptation preservation objective encourages z_u to retain information necessary for generating effective personalized adaptations, the latent representation may still encode redundant interaction-specific variations that do not contribute to personalized adaptation. We therefore introduce a compression objective by minimizing the mutual information between the latent representation and the observed interactions, i.e., $I(z_u; D_u^{\text{sup}})$.

Direct optimization of $I(z_u; D_u^{\text{sup}})$ is intractable because it depends on the marginal distribution of the latent variable. We therefore introduce a variational prior $r(z)$ to upper-bound this mutual information and instantiate it as a standard Gaussian $r(z) = \mathcal{N}(0, I)$, yielding a tractable KL regularization term. The full derivation is provided in Appendix A.1:

$$\mathcal{L}_{\text{comp}} = \mathbb{E}_u \left[D_{\text{KL}} \left(q_\psi(z_u | D_u^{\text{sup}}) \| r(z) \right) \right]. \tag{11}$$

Minimizing $\mathcal{L}_{\text{comp}}$ restricts the interaction-specific information encoded in z_u, encouraging the latent representation to focus on preference factors relevant to personalized adaptation.

Learning the latent preference space. Combining the adaptation preservation and interaction compression objectives, we jointly optimize the preference encoder q_ψ and the shared latent-to-adaptation mapping g_ϕ across training users. This joint training learns a shared latent preference space in which compact preference representations inferred from limited user interactions induce the corresponding personalized adaptations. The resulting objective is

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{adapt}} + \beta \mathcal{L}_{\text{comp}} \\ &= \underbrace{\mathbb{E}_{z_u \sim q_\psi(z_u \mid D_u^{\text{sup}})} \left[\mathcal{L} \left(D_u^{\text{qry}}; \theta_0 + g_\phi(z_u) \right) \right]}_{\text{adaptation preservation}} + \beta \underbrace{\mathbb{E}_u \left[D_{\text{KL}} \left(q_\psi(z_u \mid D_u^{\text{sup}}) \| r(z) \right) \right]}_{\text{interaction compression}}. \end{aligned} \tag{12}$$

We train LatentPersonal in a supervised manner, jointly optimizing the preference encoder q_ψ and the shared adaptation mapping g_ϕ across training users using their observed interactions. The adaptation preservation term provides the learning signal for both q_ψ and g_ϕ, while the interaction compression term regularizes q_ψ by constraining the information encoded in z_u. Consequently, q_ψ learns to infer compact preference representations from limited user interactions, while g_ϕ learns to generate the corresponding personalized adaptations. At deployment, ***LatentPersonal*** adapts to an unseen user by inferring z_u from a small set of observed interactions and generating the corresponding adaptation through g_ϕ, without requiring user-specific optimization.

3.3 LoRA Instantiation of *LatentPersonal*

To instantiate the proposed ***LatentPersonal*** framework, we adopt Low-Rank Adaptation (LoRA) (Hu et al., 2021) as a representative PEFT technique (Houlsby et al., 2019; Ben Zaken et al., 2022). The detailed implementation and training procedure are provided in Appendix B. LoRA provides a natural adaptation space for ***LatentPersonal*** due to its intrinsic low-rank parameterization. Given a pretrained weight matrix $W_0^{(l)}$, LoRA represents the adaptation as

$$\Delta W^{(l)} = B^{(l)} A^{(l)}. \tag{13}$$

where $A^{(l)}$ and $B^{(l)}$ are trainable low-rank factors. Instead of learning independent adapters for different users, ***LatentPersonal*** uses the latent preference representation z_u to modulate this adaptation space and generate user-conditioned updates:

$$\Delta W_u^{(l)} = B^{(l)} \operatorname{diag}(z_u) A^{(l)}. \tag{14}$$

Here, $\{A^{(l)}, B^{(l)}\}$ parameterize the shared adaptation subspace learned from the user population, while z_u determines the user-specific modulation within this space. In this instantiation, the LoRA structure itself realizes the latent-to-adaptation mapping g_ϕ, avoiding the need to introduce an additional parameter generator. Compared with standard LoRA, the only structural modification is the insertion of the latent modulation $\operatorname{diag}(z_u)$ between the two low-rank factors. During training, the shared LoRA factors $\{A^{(l)}, B^{(l)}\}$ are learned jointly with the lightweight preference encoder q_ψ, which maps the support interactions D_u^{sup} to the latent representation z_u. Thus, extending standard LoRA to ***LatentPersonal*** requires only a lightweight preference encoder and a corresponding modification to the training objective.

At inference time, the shared LoRA factors remain fixed, and a new user's adaptation is obtained solely by inferring z_u from the available interactions through q_ψ. No user-specific optimization of $A^{(l)}$ or $B^{(l)}$ is required. Consequently, ***LatentPersonal*** can construct user-specific parameter updates from only a few interactions by leveraging the shared adaptation structure learned during training.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate ***LatentPersonal*** on four personalization benchmarks: HiCUPID, PRISM, PersonaMem and Persona-Iterative-Responses (PIR).

- **HiCUPID**: HiCUPID (Mok et al., 2025) is a conversational benchmark for personalized AI assistants. It evaluates personalized response generation based on user-specific information captured through prior conversations.

- **PRISM**: PRISM (Kirk et al., 2024) is a human-feedback dataset that captures individual preferences through real-world interactions with diverse LLMs. Each participant engages in multiple conversations and provides personalized ratings and fine-grained feedback on model responses.
- **PersonaMem**: PersonaMem (Jiang et al., 2025) is a benchmark for personalized response generation from long-term user interactions. We adopt PersonaMem-v2, which emphasizes implicit user preferences embedded in interaction histories rather than explicitly provided user profiles.
- **PIR**: Persona-Iterative-Responses (PIR) is a large-scale persona-conditioned preference dataset containing diverse questions paired with persona descriptions and alternative responses. The dataset captures how different persona characteristics induce distinct response preferences, providing a complementary setting for evaluating whether personalized models can infer and reproduce user-specific response tendencies from limited observations.

Baselines. We compare *LatentPersonal* with the base model without personalization and four representative personalization baselines. Unless otherwise specified, *LatentPersonal* refers to its LoRA-based instantiation used throughout the main experiments.

- **ICL**: In-context learning (ICL) (Brown et al., 2020) performs few-shot personalization by conditioning the base model on user-specific demonstrations without parameter updates.
- **LoRA**: We directly fine-tune LoRA parameters on user-specific samples, providing a standard parameter-efficient baseline for personalized adaptation.
- **PPlug**: Persona-Plug (PPlug) (Liu et al., 2025) learns a user-specific representation from user samples and conditions the LLM on the resulting personal embedding for personalized generation.
- **Profile-to-PEFT**: Profile-to-PEFT (Tan et al., 2026) encodes user-specific information into a user representation and employs a shared hypernetwork to generate personalized PEFT parameters for efficient adaptation.

All experiments are conducted on a single NVIDIA A100-SXM4-80GB GPU. Additional implementation details, including training hyperparameters and model configurations, are provided in Appendix C.1, while detailed definitions of the evaluation metrics are provided in Appendix C.2.

4.2 Results and Analysis

Personalization Performance. We report the main results using Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as the backbone and select three representative metrics for a concise comparison. Results with additional backbone models and the complete set of evaluation metrics are provided in Appendix D.1. We first report the zero-shot performance of the base model in Table 1, followed by the main personalization results under different numbers of user-specific observations in Table 2.

Overall, *LatentPersonal* achieves strong performance across all four benchmarks, with particularly consistent advantages in the few-shot regime. As more user observations become available, the gap to competing methods generally narrows, further highlighting the effectiveness of latent navigation when user-specific evidence is limited.

Component Analysis. We ablate three key designs of *LatentPersonal*. with the main results reported in Table 3. For *w/o Latent*, we remove the latent modulation z_u from *LatentPersonal* while retaining the same optimization framework, such that personalization is performed directly through the LoRA components. For *w/o Navigation*, we retain the inferred user latent representation z_u but keep it fixed during personalization, and instead update the LoRA components, thereby replacing latent-space navigation with direct parameter adaptation. For *w/o IB*, we remove the information bottleneck regularization by setting $\beta = 0$.

Removing either the latent representation or the information bottleneck leads to substantial performance degradation, while replacing latent-space navigation with direct LoRA adaptation results in

Table 1: Performance of the frozen base model without personalized adaptation.

Dataset	ROUGE-1	METEOR	Content Personalization
HiCUPID	0.1614	0.2166	3.57
PRISM	0.2558	0.2125	4.18
PersonaMem	0.2368	0.2103	5.01
PIR	0.2954	0.1966	4.08

Table 2: Main results across four personalization benchmarks under different numbers of observed user interactions. R-1, MET., and CP denote ROUGE-1, METEOR, and Content Personalization, respectively. Best results are shown in **bold** and second-best results are underlined.

Dataset	Method	1-shot			2-shot			4-shot			8-shot		
		R-1	MET.	CP	R-1	MET.	CP	R-1	MET.	CP	R-1	MET.	CP
HiCUPID	ICL	0.1713	0.2377	3.85	0.1682	0.2294	4.25	0.1782	0.2429	4.64	0.1934	0.2730	5.52
	LoRA	0.3886	0.3829	5.57	0.3993	0.3940	5.73	0.3912	0.3923	5.61	0.3984	0.4027	**6.26**
	PPlug	0.3018	0.2696	5.15	0.3116	0.2813	5.37	0.3276	0.3018	5.63	0.3315	0.3105	5.95
	Profile-to-PEFT	0.4015	0.4022	5.64	0.4018	0.4052	5.78	0.4039	0.4014	5.75	0.3975	0.4078	6.17
	LatentPersonal	**0.4162**	**0.4133**	**5.95**	**0.4160**	**0.4142**	**5.92**	**0.4156**	**0.4137**	**5.94**	**0.4162**	**0.4135**	6.24
PRISM	ICL	0.2929	0.2372	4.11	0.2917	0.2233	4.71	0.2998	0.2263	4.82	0.3009	0.2276	**5.84**
	LoRA	0.2843	0.2046	4.05	0.2649	0.1940	4.63	0.2568	0.1873	4.97	0.2638	0.1926	5.23
	PPlug	0.2875	0.2033	4.11	0.2806	0.1951	4.55	0.2853	0.1950	4.97	0.2889	0.1945	5.54
	Profile-to-PEFT	0.2895	0.2200	4.15	0.2888	0.2141	4.42	0.2949	0.2221	5.28	0.2928	0.2187	5.68
	LatentPersonal	**0.3081**	**0.2421**	**4.28**	**0.3036**	**0.2380**	**4.82**	**0.3043**	**0.2385**	**5.30**	**0.3049**	**0.2391**	5.79
PersonaMem	ICL	0.2514	0.2352	4.87	0.2682	0.2457	5.51	0.2618	0.2457	6.07	0.2759	0.2618	6.39
	LoRA	0.3448	0.2767	5.97	0.3442	0.2740	6.09	0.3447	0.2800	6.25	0.3446	0.2782	6.40
	PPlug	0.2949	0.2265	5.17	0.2902	0.2216	5.19	0.2939	0.2238	5.53	0.2953	0.2226	5.88
	Profile-to-PEFT	0.3398	0.2749	5.74	0.3480	0.2806	5.70	0.3523	0.2827	5.51	0.3561	0.2859	6.37
	LatentPersonal	**0.3534**	**0.2803**	**6.14**	**0.3631**	**0.2930**	**6.27**	**0.3647**	**0.2996**	**6.35**	**0.3718**	**0.3031**	**6.42**
PIR	ICL	0.3028	0.2066	4.19	0.3090	0.2114	5.16	0.3125	0.2129	5.89	0.3082	0.2082	6.51
	LoRA	0.3805	0.2962	4.28	0.3795	0.2931	5.25	0.3836	0.2983	5.72	0.3846	0.2977	**6.67**
	PPlug	0.3711	0.2495	3.96	0.3633	0.2401	5.16	0.3645	0.2436	5.86	0.3646	0.2409	6.56
	Profile-to-PEFT	0.3891	0.3066	4.08	**0.3952**	**0.3111**	5.13	0.3857	0.3024	5.75	**0.3880**	0.3045	6.63
	LatentPersonal	**0.3932**	**0.3087**	**4.38**	0.3926	0.3094	**5.30**	**0.3912**	**0.3084**	**5.90**	0.3852	**0.3163**	6.58

Table 3: Ablation study of key components in *LatentPersonal* on HiCUPID with Llama-3.1-8B-Instruct. R-1, MET., and CP denote ROUGE-1, METEOR, and Content Personalization.

Variant	1-shot			2-shot			4-shot		
	R-1	MET.	CP	R-1	MET.	CP	R-1	MET.	CP
LatentPersonal	0.4162	0.4133	5.95	0.4160	0.4142	5.92	0.4156	0.4137	5.94
w/o Latent	0.3550	0.3667	4.72	0.3499	0.3841	4.86	0.3308	0.3423	5.29
w/o Navigation	0.3812	0.3777	5.66	0.3910	0.3929	5.71	0.3909	0.3992	5.92
w/o IB	0.2684	0.3423	4.57	0.2703	0.3422	4.65	0.2695	0.3444	4.74

a smaller but consistent drop, supporting the contribution of all three components. Detailed ablation results are provided in Appendix D.2, with additional sensitivity analyses on the bottleneck coefficient β and latent dimension d_z in Appendix D.3.

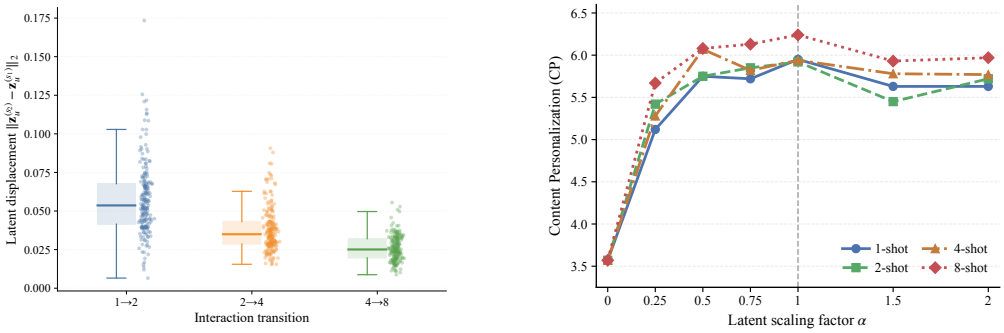
Test-Time Personalization Efficiency. We further evaluate the efficiency of test-time adaptation across personalization methods, with the main results reported in Table 4. Personalization latency measures the time required to process the observed user interactions and obtain the user-specific adaptation before response generation. *LatentPersonal* substantially reduces personalization overhead with only 1-4 observed interactions, and its cost increases moderately as more observations are incorporated. Detailed results on both personalization and generation latency are provided in Appendix D.4.

Table 4: Test-time personalization latency (s) on HiCUPID.

Method	1-shot	2-shot	4-shot	8-shot
LoRA	0.3718	0.3622	0.3692	0.3640
Profile-to-PEFT	0.0539	0.0538	0.0539	**0.0544**
LatentPersonal	**0.0219**	**0.0304**	**0.0404**	0.0676

Structure of the Learned Latent Space. We further examine the learned latent space in Figure 3b. We scale the inferred user representation as $\mathbf{z}'_u = \alpha \mathbf{z}_u$, where α controls the strength of latent guidance. Latent updates progressively decrease as more user observations are incorporated, suggesting that user representations gradually stabilize with increasing evidence. Meanwhile, removing the latent signal ($\alpha = 0$) substantially degrades personalization, while moderate scaling maintains strong performance. We further analyze user-specific navigation in Figure 4. Matched user representations

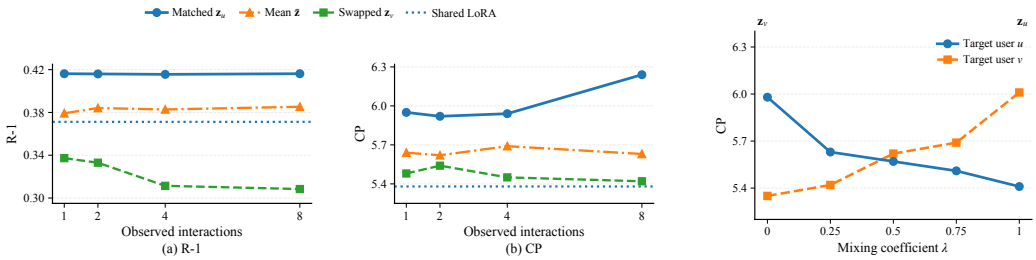
consistently outperform swapped, population-mean, and shared LoRA alternatives, while interpolating between two user representations smoothly shifts personalization performance from one target user toward the other. Together, these results suggest that the learned latent space is not only structured and stable, but also supports continuous user-specific navigation. More extensive analyses are provided in Appendix D.5.



(a) Shot-wise latent displacement.

(b) Effect of latent scaling.

Figure 3: **Structure and functional behavior of the learned latent space on HiCUPID.** (a) Shot-wise latent displacement decreases with additional user observations. (b) Personalization performance under different latent scaling factors α.



(a) User-specificity analysis.

(b) Inter-user interpolation.

Figure 4: **User specificity and continuous navigation in the learned latent adaptation space on HiCUPID.** (a) Matched user representations outperform swapped, population-mean, and shared LoRA alternatives. (b) Interpolation between two representative users smoothly shifts personalization performance from one target user toward the other.

Generalization across PEFT Architectures. Finally, we evaluate Adapter-based and Prompt Tuning-based instantiations of *LatentPersonal*, with results reported in Appendix D.6. The consistent improvements across different PEFT architectures demonstrate that *LatentPersonal* is not specific to its LoRA-based instantiation, but can be naturally extended to different parameter-efficient adaptation mechanisms. Implementation details of these variants are provided in Appendix E.

5 Conclusion

In this work, we revisit LLM personalization through the lens of latent navigation, motivated by the hypothesis that pretrained LLMs may support a broad range of user-specific behaviors that can be elicited with limited user-specific evidence. From this perspective, we introduce *LatentPersonal*, which infers compact user representations and performs personalization through a shared latent adaptation space. Our results show that user-specific adaptations can be effectively represented within this shared space, where representations inferred from only a few interactions enable effective personalization for unseen users. Experiments across multiple personalization benchmarks further demonstrate that *LatentPersonal* is particularly effective in the few-shot regime, while analyses of its components, efficiency, and learned latent space provide empirical support for the proposed formulation. Moreover, its extension across different PEFT architectures suggests that the latent-navigation formulation is not tied to a particular adaptation mechanism. Overall, our results highlight latent navigation as a promising perspective for efficient few-shot personalization of LLMs.

AI USE STATEMENT

Generative AI tools were used to assist with language refinement, code development, and debugging. All AI-assisted code and research suggestions were independently reviewed and validated by the authors. The authors retained full responsibility for the research methodology, experimental design, analysis, conclusions, and final content of the work.

REFERENCES

Stéphane Aroca-Ouellette, Natalie Mackraz, Barry-John Theobald, and Katherine Metcalf. Aligning llms by predicting preferences from user writing samples. *arXiv preprint arXiv:2505.23815*, 2025.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9, 2022.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Heng Cao, Fan Zhang, Jian Yao, Yujie Zheng, Changlin Zhao, Lu Hao, Yuxuan Wei, Wangze Ni, Huaiyu Fu, Yuqian Sun, et al. Beyond retrieval: Learning compact user representations for scalable llm personalization. *arXiv preprint arXiv:2606.04547*, 2026.

Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. Implicit deep latent variable models for text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3946–3956, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. Controllable text generation via probability density estimation in the latent space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12590–12616, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jinyi Hu, Xiaoyuan Yi, Wenhao Li, Maosong Sun, and Xing Xie. Fuse it more deeply! a variational transformer with layer-wise latent variable inference for text generation. In *Proceedings of the 2022 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pp. 697–716, 2022.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, et al. Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Hannah R Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

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 (volume 1: Long papers)*, pp. 4582–4597, 2021.

Vladislav Lialin, Vijeta Deshpande, Xiaowei Yao, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.

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

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. Llms+ persona-plug= personalized llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9373–9385, 2025.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Jisoo Mok, Ik-hwan Kim, Sangkwon Park, and Sungroh Yoon. Exploring the potential of llms as personalized assistants: Dataset, evaluation, and analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10212–10239, 2025.

Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. User-llm: Efficient llm contextualization with user embeddings. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 1219–1223, 2025.

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

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

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

Zhaoxuan Tan, Zixuan Zhang, Haoyang Wen, Zheng Li, Rongzhi Zhang, Pei Chen, Fengran Mo, Zheyuan Liu, Qingkai Zeng, Qingyu Yin, et al. Instant personalized large language model adaptation via hypernetwork. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23557–23580, 2026.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. Proper: A progressive learning framework for personalized large language models with group-level adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16399–16411, 2025.

Yuansen Zhang, Xiao Wang, Tianze Chen, Jiayi Fu, Tao Gui, and Qi Zhang. P4: Plug-and-play discrete prompting for large language models personalization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9129–9144, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Thomas Zollo, Andrew Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personalllm: Tailoring llms to individual preferences. In *International Conference on Learning Representations*, volume 2025, pp. 66949–66971, 2025.

A METHOD SUPPLEMENT

A.1 DERIVATION OF VARIATIONAL PREFERENCE COMPRESSION

We provide the derivation of the variational upper bound used for the interaction compression objective. Under the joint distribution $p(D_u^{\text{sup}})q_\psi(z_u \mid D_u^{\text{sup}})$, the mutual information between the observed interactions D_u^{sup} and the latent preference representation z_u can be written as

$$I(z_u; D_u^{\text{sup}}) = \mathbb{E}_{p(D_u^{\text{sup}})} \left[D_{\text{KL}} \left(q_\psi(z_u|D_u^{\text{sup}}) \| q_\psi(z_u) \right) \right], \tag{15}$$

where the aggregated posterior is defined as

$$q_\psi(z_u) = \int p(D_u^{\text{sup}}) q_\psi(z_u|D_u^{\text{sup}}) dD_u^{\text{sup}}. \tag{16}$$

Since the aggregated posterior $q_\psi(z_u)$ is generally intractable, we introduce a tractable variational prior $r(z)$. The corresponding KL term can be decomposed as

$$\begin{aligned}
&\mathbb{E}_{p(D_u^{\text{sup}})} \left[D_{\text{KL}} \left(q_\psi(z_u|D_u^{\text{sup}}) \| r(z) \right) \right] \\
&= \mathbb{E}_{p(D_u^{\text{sup}}) q_\psi(z_u|D_u^{\text{sup}})} \left[\log \frac{q_\psi(z_u|D_u^{\text{sup}})}{r(z)} \right] \\
&= \mathbb{E}_{p(D_u^{\text{sup}}) q_\psi(z_u|D_u^{\text{sup}})} \left[\log \frac{q_\psi(z_u|D_u^{\text{sup}})}{q_\psi(z_u)} + \log \frac{q_\psi(z_u)}{r(z)} \right] \\
&= I(z_u; D_u^{\text{sup}}) + D_{\text{KL}} \left(q_\psi(z_u) \| r(z) \right).
\end{aligned} \tag{17}$$

Since $D_{\text{KL}}(q_\psi(z_u)\|r(z)) \geq 0$, we obtain the variational upper bound

$$I(z_u; D_u^{\text{sup}}) \leq \mathbb{E}_{p(D_u^{\text{sup}})} \left[D_{\text{KL}} \left(q_\psi(z_u|D_u^{\text{sup}}) \| r(z) \right) \right]. \tag{18}$$

For implementation, we adopt a diagonal Gaussian posterior

$$q_\psi(z_u|D_u^{\text{sup}}) = \mathcal{N}\left(\mu_u, \text{diag}(\sigma_u^2) \right), \tag{19}$$

where $\mu_u \in \mathbb{R}^d$ and $\sigma_u \in \mathbb{R}_+^d$ denote the mean and standard deviation vectors predicted by the preference encoder, respectively. We use a standard Gaussian prior $r(z) = \mathcal{N}(0, I)$. The resulting KL regularization admits the closed-form expression:

$$D_{\text{KL}} \left(q_\psi(z_u|D_u^{\text{sup}}) \| \mathcal{N}(0, I) \right) = \frac{1}{2} \sum_{j=1}^{d} \left(\mu_{u,j}^2 + \sigma_{u,j}^2 - \log \sigma_{u,j}^2 - 1 \right). \tag{20}$$

B Detailed LoRA-based Implementation of *LatentPersonal*

Following the latent adaptation formulation introduced in Section 3.2, we provide a detailed implementation of *LatentPersonal* based on LoRA. Although *LatentPersonal* is not restricted to a specific parameter-efficient adaptation mechanism, we choose LoRA as a representative instantiation due to its wide adoption and computational efficiency.

LoRA parameterizes the update of a pretrained weight matrix as a low-rank decomposition:

$$\Delta W^{(l)} = B^{(l)} A^{(l)}, \tag{21}$$

where $B^{(l)} \in \mathbb{R}^{d_{\text{out}} \times r}$ and $A^{(l)} \in \mathbb{R}^{r \times d_{\text{in}}}$, with $r \ll \min(d_{\text{out}}, d_{\text{in}})$. This low-rank formulation provides a natural parameterization for realizing the latent-to-adaptation mapping in *LatentPersonal*.

In the following, we describe how user preference representations are incorporated into the LoRA adaptation process and how the resulting latent-conditioned LoRA model is jointly trained across users.

B.1 Latent-conditioned LoRA Adaptation

The low-rank structure of LoRA provides a natural adaptation space for realizing the latent-to-adaptation mapping in *LatentPersonal*. Specifically, the rank r determines the dimensionality of the intermediate adaptation space, while the low-rank factors induce a set of shared adaptation directions. For the l-th transformer layer, the LoRA factors can be expressed as

$$B^{(l)} A^{(l)} = \sum_{i=1}^{r} b_i^{(l)} a_i^{(l)T}, \tag{22}$$

where each rank-one component $b_i^{(l)} a_i^{(l)T}$ defines an adaptation direction shared across users. Based on this decomposition, we align the latent preference representation with the LoRA rank space, i.e., $z_u \in \mathbb{R}^r$, and interpret z_u as a user-specific coordinate vector that modulates these shared adaptation directions.

Therefore, rather than parameterizing an independent LoRA update for each user, *LatentPersonal* constructs user-conditioned adaptations by modulating the contribution of the shared low-rank directions:

$$\Delta W_u^{(l)} = B^{(l)} \operatorname{diag}(z_u) A^{(l)}. \tag{23}$$

Equivalently, the personalized update can be written as

$$\Delta W_u^{(l)} = \sum_{i=1}^{r} z_{u,i} b_i^{(l)} a_i^{(l)T}. \tag{24}$$

Here, the LoRA factors $\{b_i^{(l)} a_i^{(l)T}\}_{i=1}^{r}$ define a shared adaptation basis learned from the population of users, while $z_{u,i}$ specifies the contribution of each basis direction for user u. Consequently, *LatentPersonal* does not require independent user-specific adapters; instead, personalization is realized by inferring a compact latent representation that determines the user-specific combination of shared low-rank adaptation directions.

B.2 Preference Representation Inference

To infer a user preference representation from a variable-sized set of support interactions, we first transform $D_u^{\text{sup}} = \{(x_i, y_i)\}_{i=1}^{K}$ into a fixed-dimensional representation. Specifically, we use the frozen pretrained LLM backbone as the interaction encoder. For each interaction (x_i, y_i), we extract the final-layer hidden states and apply mean pooling over non-padding tokens to obtain an

interaction-level representation h_i. The representations of the K observed interactions are then aggregated by mean pooling,

$$h_u = \frac{1}{K} \sum_{i=1}^{K} h_i, \tag{25}$$

yielding a fixed-dimensional user representation independent of the number of observed interactions. The resulting representation h_u is fed into the two-layer MLP preference encoder described in Table 6, which predicts the parameters of the variational posterior,

$$(\mu_u, \log \sigma_u^2) = q_\psi(h_u), \qquad q_\psi(z_u \mid D_u^{\mathrm{sup}}) = \mathcal{N}\big(\mu_u, \mathrm{diag}(\sigma_u^2)\big)\,. \tag{26}$$

This simple parameter-free aggregation naturally accommodates varying numbers of observed interactions while keeping preference inference lightweight.

B.3 Training with Latent-Conditioned LoRA

During training, LatentPersonal jointly learns the preference encoder and the shared LoRA adaptation space across multiple users. For each training user u, we partition the available interactions into a support set D_u^{sup} and a disjoint query set D_u^{qry}. The support interactions are used to infer the latent preference representation:

$$z_u \sim q_\psi(z_u | D_u^{\mathrm{sup}}), \tag{27}$$

The inferred latent representation is then used to generate the user-conditioned LoRA updates described in Section B.1. The resulting personalized model is optimized on D_u^{qry}, providing the learning signal for both the preference encoder and the shared LoRA adaptation space.

$$\theta_u = \theta_0 + \Delta\theta_u^{\mathrm{LoRA}}(z_u). \tag{28}$$

The resulting personalized model is optimized on the query interactions D_u^{qry}, providing the learning signal for both the preference encoder and the shared LoRA adaptation space. The support interactions D_u^{sup} are used for latent inference, while the disjoint query interactions D_u^{qry} provide supervision for learning the shared adaptation space. Specifically, instantiating the general latent-to-adaptation mapping in Eq. (9) with the latent-conditioned LoRA parameterization leads to the following training objective:

$$\mathcal{L}_{\mathrm{total}} = \mathbb{E}_{z_u \sim q_\psi(z_u | D_u^{\mathrm{sup}})} \left[\mathcal{L}\left(D_u^{\mathrm{qry}}; \theta_0 + \Delta\theta_u^{\mathrm{LoRA}}(z_u)\right) \right] + \beta \mathbb{E}_u \left[D_{\mathrm{KL}}\left(q_\psi(z_u \mid D_u^{\mathrm{sup}}) \| r(z)\right) \right]. \tag{29}$$

which corresponds to the general variational objective in Eq. (12) under the LoRA instantiation.

The preference encoder parameters ψ and the shared LoRA parameters $\{A^{(l)}, B^{(l)}\}_{l=1}^{L}$ are jointly optimized by minimizing $\mathcal{L}_{\mathrm{total}}$.

Different from conventional personalized fine-tuning, *LatentPersonal* does not optimize independent LoRA parameters for each user. Instead, it learns a shared LoRA adaptation space across users, while user-specific personalization is achieved by inferring a compact latent preference representation from a small number of observed interactions.

The overall training procedure of *LatentPersonal* with latent-conditioned LoRA is summarized in Algorithm 1.

Algorithm 1 Training Procedure of *LatentPersonal* with Latent-Conditioned LoRA

Require: Pretrained model θ_0, training users $\mathcal{U}_{\text{train}}$

Initialize preference encoder q_ψ
Initialize shared LoRA parameters $\phi = \{A^{(l)}, B^{(l)}\}_{l=1}^{L}$
while not converged **do**
  Sample a batch of users $\mathcal{B} \subset \mathcal{U}_{\text{train}}$
  Initialize $\mathcal{L}_{\text{batch}} \leftarrow 0$
  for each user $u \in \mathcal{B}$ **do**
    Retrieve interactions D_u and split into D_u^{sup} and D_u^{qry}
    Infer preference posterior: $(\mu_u, \sigma_u) \leftarrow q_\psi(D_u^{\text{sup}})$
    Sample latent representation: $z_u = \mu_u + \sigma_u \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$
    Generate latent-conditioned LoRA adaptations:
      $\Delta W_u^{(l)} = B^{(l)} \operatorname{diag}(z_u) A^{(l)}, \quad \forall l = 1, \ldots, L$
    Compute adaptation loss: $\mathcal{L}_{\text{adapt},u} = \mathcal{L}\left(D_u^{\text{qry}}; \theta_0 + \Delta\theta_u^{\text{LoRA}}(z_u)\right)$
    Compute compression loss: $\mathcal{L}_{\text{comp},u} = D_{\text{KL}}\left(\mathcal{N}(\mu_u, \operatorname{diag}(\sigma_u^2)) \| \mathcal{N}(0, I)\right)$
    Accumulate training objective: $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \mathcal{L}_{\text{adapt},u} + \beta\mathcal{L}_{\text{comp},u}$
  end for
  Update ψ and ϕ: $(\psi, \phi) \leftarrow (\psi, \phi) - \eta\nabla_{\psi,\phi}\mathcal{L}_{\text{batch}}$
end while
return q_ψ and shared LoRA parameters ϕ

C Implementation Details

C.1 Model and Training Configuration

Table 5: Training hyperparameters.

Hyperparameter	Value
LoRA rank r	16
LoRA scaling α	32
LoRA dropout	0.05
IB coefficient β	0.01
Batch size	1
Training epochs	3
Learning rate	2×10^{-4}

Table 6: Configuration of the preference encoder.

Component	Configuration
Hidden dimension	512
Number of hidden layers	2
Activation function	GELU
Dropout	0.1
Mean head	Linear ($512 \to d_z$)
Log-variance head	Linear ($512 \to d_z$)
Latent dimension d_z	16

Unless otherwise specified, we use the default configurations of the corresponding pretrained models. For response generation, we use deterministic greedy decoding with `do_sample=False`, `num_beams=1`, and `max_length=512`. Training and test users are strictly disjoint within each dataset. For each test user, 1, 2, 4, 8, or 16 interactions constitute the observed support set for personalization, while 4 additional held-out interactions are reserved exclusively for evaluation. Each experimental setting is evaluated on at least 200 held-out interactions across test users. We first average the evaluation scores across the held-out interactions for each user and then average the resulting user-level scores across all test users.

For each benchmark, we train all methods independently using only the corresponding training split and evaluate them on the held-out test users from the same dataset. This benchmark-specific protocol is applied consistently to all compared methods. Accordingly, the learned latent adaptation space is shared across users within each benchmark, rather than across heterogeneous datasets.

Training support–query construction. To align training with the few-shot personalization setting at test time, we adopt an episodic support–query training procedure. For each sampled training user u, we randomly sample a support size K_{sup} from $\{1, 2, 4, 8, 16\}$ and construct a support set D_u^{sup} containing K_{sup} interactions. We then sample $K_{\text{qry}} = 4$ additional interactions from the same user, disjoint from the support set, to form the query set D_u^{qry}. The support set is used exclusively to

infer the user-specific latent representation $z_u \sim q_\psi(z_u \mid D_u^{\text{sup}})$, while the query set is used to compute the adaptation loss $\mathcal{L}(D_u^{\text{qry}}; \theta_0 + g_\phi(z_u))$. The information-bottleneck regularization is applied to the posterior inferred from D_u^{sup}. Support and query interactions are resampled across training episodes, encouraging the inferred latent representation to capture user-specific preference information that generalizes beyond the particular interactions used for latent inference.

Baseline Selection. Our experiments focus on few-shot personalization for unseen users, where only a small support set is available at test time. We therefore select baselines that can be directly evaluated under this protocol. Methods such as OPPU (Tan et al., 2024b) and TAP-PER (Cao et al., 2026), which primarily rely on persistent user histories or user-specific states, target a different personalization setting and are therefore not included as direct baselines. PerFit (Liu et al., 2026) similarly considers a history-rich personalization setting, selecting users with extensive interaction histories and performing a second-stage optimization of user-specific representation interventions. In contrast, our setting evaluates adaptation to unseen users from only 1–8 observed interactions, without user-specific optimization at deployment. Its reported results are therefore not directly comparable under our few-shot evaluation protocol.

Persona-Iterative-Responses (PIR). We use PIR from the publicly available Hugging Face repository `sher222/persona-iterative-responses`.[1]

C.2 Evaluation Metrics

We evaluate model performance using five metrics that capture both reference-based generation quality and user-specific personalization. Specifically, we report **ROUGE-1**, **ROUGE-L**, and **METEOR** for generation quality, and introduce **Content Personalization (CP)** and **Presentation Personalization (PP)** to assess personalization from complementary perspectives (Zheng et al., 2023). Higher values indicate better performance for all metrics.

ROUGE-1 and ROUGE-L. We use ROUGE-1 and ROUGE-L to measure the lexical and sequential overlap between the generated response and the reference response. ROUGE-1 evaluates unigram-level overlap, while ROUGE-L measures similarity based on the longest common subsequence.

METEOR. METEOR measures the similarity between the generated and reference responses through token-level alignment, providing a complementary evaluation of lexical and semantic consistency beyond n-gram overlap.

Content Personalization (CP). CP evaluates whether the content of a generated response is aligned with the preferences reflected in the user's historical interactions. It considers the selection of topics, examples, information, emphasis, and ordering, and penalizes responses that are generic or inconsistent with the user's inferred interests and goals. Given the user history H_u, current input x, reference response y, and generated response $\hat{y}$, we define the sample-level CP score as

$$s_{\text{CP}} = \mathcal{E}_{\text{CP}}\left(H_u, x, y, \hat{y}\right), \qquad s_{\text{CP}} \in \{1, \ldots, 10\}, \tag{30}$$

where $\mathcal{E}_{\text{CP}}$ denotes the LLM-based evaluator. A higher CP score indicates stronger alignment between the generated content and the user's preferences.

Presentation Personalization (PP). PP evaluates whether the presentation of a generated response matches the patterns reflected in the user's interaction history, including tone, writing style, organization, level of detail, and formatting. Similarly, the sample-level PP score is defined as

$$s_{\text{PP}} = \mathcal{E}_{\text{PP}}\left(H_u, x, y, \hat{y}\right), \qquad s_{\text{PP}} \in \{1, \ldots, 10\}, \tag{31}$$

where $\mathcal{E}_{\text{PP}}$ denotes the corresponding LLM-based evaluator. A higher PP score indicates better alignment with the user's preferred presentation style.

[1] `https://huggingface.co/datasets/sher222/persona-iterative-responses`

For CP and PP, the evaluator is provided with the user history, current input, reference response, and generated response. We report the average score over all evaluation samples:

$$\mathrm{CP} = \frac{1}{N}\sum_{i=1}^{N} s_{\mathrm{CP}}^{(i)}, \qquad \mathrm{PP} = \frac{1}{N}\sum_{i=1}^{N} s_{\mathrm{PP}}^{(i)}, \tag{32}$$

where N denotes the number of evaluated samples.

LLM-based Evaluation Protocol. For CP and PP, we use `gpt-4.1-mini` as the LLM-based evaluator. For each evaluation sample, the evaluator receives the users historical interactions H_u, current input x, reference response y, and model-generated response $\hat{y}$. The same evaluator model, system prompt, and evaluation procedure are used for all methods and datasets. The evaluator independently assigns integer scores from 1 to 10 for content and presentation personalization, where higher scores indicate stronger personalization. We use deterministic evaluation to reduce randomness in judge outputs. The complete evaluator instruction is provided below for reproducibility.

System Prompt.

You are an evaluator for personalized text generation. Evaluate the generated response with respect to the user's historical information and the reference response.

Score the response on two dimensions from 1 to 10.

1. Content Personalization (CP).

Evaluate whether the selected content, topics, examples, emphasis, and ordering align with the user's explicit or inferred interests, preferences, and goals.

A score of 1 indicates that the response is generic, user-agnostic, or substantially inconsistent with the user's preferences; a score of 5 indicates moderate personalization, where some content reflects the user's preferences but substantial portions remain generic; and a score of 10 indicates that the content is strongly and consistently tailored to the user's preferences and goals.

2. Presentation Personalization (PP).

Evaluate whether the response matches the user's preferred tone, writing style, organization, level of detail, formatting, and other presentation patterns reflected in the user's history.

A score of 1 indicates little or no alignment with the user's presentation preferences; a score of 5 indicates partial alignment with noticeable generic or mismatched presentation choices; and a score of 10 indicates strong and consistent alignment with the user's preferred presentation style.

Use the user history as the primary evidence for personalization and the reference response as additional evidence of the desired response for the current input. Do not reward a response merely for being generally high quality if it does not reflect user-specific preferences.

Return only valid JSON in the following format:

```
{
  "content_personalization": <integer from 1 to 10>,
  "presentation_personalization": <integer from 1 to 10>
}
```

Table 7: Full personalization results on HiCUPID under different few-shot settings.

Method	Shot	R-1	R-L	MET.	CP	PP
Llama-3.1-8B-Instruct						
Base	/	0.1614 ± 0.0124	0.1097 ± 0.0083	0.2166 ± 0.0136	3.57 ± 0.30	4.29 ± 0.23
ICL	1	0.1713 ± 0.0148	0.1086 ± 0.0089	0.2377 ± 0.0158	3.85 ± 0.30	4.91 ± 0.21
	2	0.1682 ± 0.0152	0.1064 ± 0.0088	0.2294 ± 0.0169	4.25 ± 0.34	5.08 ± 0.25
	4	0.1782 ± 0.0152	0.1185 ± 0.0103	0.2429 ± 0.0166	4.64 ± 0.36	5.20 ± 0.26
	8	0.1934 ± 0.0189	0.1352 ± 0.0147	0.2730 ± 0.0209	5.52 ± 0.43	5.63 ± 0.34
	16	0.2114 ± 0.0221	0.1634 ± 0.0210	0.3090 ± 0.0272	6.55 ± 0.44	6.72 ± 0.38
LoRA	1	0.3886 ± 0.0412	0.3390 ± 0.0401	0.3829 ± 0.0423	5.57 ± 0.39	6.01 ± 0.26
	2	0.3993 ± 0.0406	0.3541 ± 0.0396	0.3940 ± 0.0416	5.73 ± 0.39	6.19 ± 0.30
	4	0.3912 ± 0.0400	0.3468 ± 0.0398	0.3923 ± 0.0413	5.61 ± 0.43	6.20 ± 0.32
	8	0.3984 ± 0.0392	0.3434 ± 0.0393	0.4027 ± 0.0393	6.26 ± 0.39	6.52 ± 0.31
	16	0.4223 ± 0.0415	0.3729 ± 0.0426	0.4196 ± 0.0431	6.55 ± 0.44	6.83 ± 0.34
PPlug	1	0.3018 ± 0.0310	0.2395 ± 0.0278	0.2696 ± 0.0294	5.15 ± 0.41	5.42 ± 0.30
	2	0.3116 ± 0.0319	0.2529 ± 0.0294	0.2813 ± 0.0303	5.37 ± 0.42	5.60 ± 0.32
	4	0.3276 ± 0.0356	0.2690 ± 0.0327	0.3018 ± 0.0349	5.63 ± 0.40	5.68 ± 0.32
	8	0.3315 ± 0.0372	0.2715 ± 0.0331	0.3105 ± 0.0347	5.95 ± 0.38	6.14 ± 0.30
	16	0.3763 ± 0.0405	0.3137 ± 0.0373	0.3515 ± 0.0387	6.38 ± 0.40	6.37 ± 0.31
Profile-to-PEFT	1	0.4015 ± 0.0372	0.3440 ± 0.0370	0.4022 ± 0.0397	5.64 ± 0.40	6.02 ± 0.29
	2	0.4018 ± 0.0404	0.3482 ± 0.0399	0.4052 ± 0.0416	5.78 ± 0.40	6.18 ± 0.32
	4	0.4039 ± 0.0390	0.3481 ± 0.0400	0.4014 ± 0.0407	5.75 ± 0.40	6.12 ± 0.31
	8	0.3975 ± 0.0380	0.3455 ± 0.0387	0.4078 ± 0.0405	6.17 ± 0.44	6.48 ± 0.32
	16	0.4338 ± 0.0426	0.3804 ± 0.0434	0.4316 ± 0.0439	6.48 ± 0.39	6.71 ± 0.33
LatentPersonal	1	0.4162 ± 0.0408	0.3637 ± 0.0400	0.4133 ± 0.0404	5.95 ± 0.41	6.33 ± 0.31
	2	0.4160 ± 0.0415	0.3656 ± 0.0404	0.4142 ± 0.0412	5.92 ± 0.40	6.28 ± 0.30
	4	0.4156 ± 0.0412	0.3639 ± 0.0403	0.4137 ± 0.0412	5.94 ± 0.43	6.27 ± 0.33
	8	0.4162 ± 0.0412	0.3652 ± 0.0405	0.4135 ± 0.0408	6.24 ± 0.43	6.61 ± 0.32
	16	0.4323 ± 0.0428	0.3851 ± 0.0428	0.4280 ± 0.0424	6.44 ± 0.43	6.72 ± 0.36
Qwen2.5-7B-Instruct						
Base	/	0.1712 ± 0.0124	0.1072 ± 0.0072	0.2171 ± 0.0127	3.84 ± 0.31	4.90 ± 0.23
ICL	1	0.1696 ± 0.0147	0.1086 ± 0.0091	0.2347 ± 0.0160	3.77 ± 0.31	4.69 ± 0.22
	2	0.1713 ± 0.0149	0.1074 ± 0.0086	0.2275 ± 0.0151	4.29 ± 0.32	5.06 ± 0.23
	4	0.1761 ± 0.0150	0.1175 ± 0.0109	0.2440 ± 0.0167	4.68 ± 0.38	5.24 ± 0.28
	8	0.1977 ± 0.0195	0.1382 ± 0.0161	0.2772 ± 0.0223	5.36 ± 0.44	5.58 ± 0.34
	16	0.2185 ± 0.0202	0.1675 ± 0.0193	0.3179 ± 0.0268	6.35 ± 0.42	6.16 ± 0.37
LoRA	1	0.3762 ± 0.0422	0.3203 ± 0.0414	0.3721 ± 0.0410	5.82 ± 0.43	6.16 ± 0.33
	2	0.3852 ± 0.0409	0.3321 ± 0.0413	0.3796 ± 0.0391	5.99 ± 0.45	6.34 ± 0.33
	4	0.3904 ± 0.0376	0.3376 ± 0.0375	0.3880 ± 0.0370	6.06 ± 0.46	6.39 ± 0.37
	8	0.3838 ± 0.0415	0.3344 ± 0.0404	0.3799 ± 0.0409	6.33 ± 0.45	6.54 ± 0.37
	16	0.4169 ± 0.0383	0.3607 ± 0.0395	0.4058 ± 0.0393	6.16 ± 0.51	6.50 ± 0.39
PPlug	1	0.2226 ± 0.0256	0.1638 ± 0.0191	0.2020 ± 0.0247	4.46 ± 0.36	5.01 ± 0.27
	2	0.2478 ± 0.0255	0.1852 ± 0.0199	0.2131 ± 0.0228	4.53 ± 0.38	4.83 ± 0.31
	4	0.2594 ± 0.0221	0.1953 ± 0.0181	0.2260 ± 0.0216	4.79 ± 0.38	5.01 ± 0.33
	8	0.2583 ± 0.0233	0.1989 ± 0.0196	0.2329 ± 0.0223	5.50 ± 0.41	5.64 ± 0.30
	16	0.2878 ± 0.0271	0.2177 ± 0.0233	0.2626 ± 0.0282	5.69 ± 0.39	5.71 ± 0.30
Profile-to-PEFT	1	0.3812 ± 0.0408	0.3320 ± 0.0417	0.3791 ± 0.0405	5.80 ± 0.40	6.17 ± 0.27
	2	0.3841 ± 0.0418	0.3349 ± 0.0411	0.3825 ± 0.0406	5.94 ± 0.38	6.25 ± 0.27
	4	0.3971 ± 0.0420	0.3434 ± 0.0422	0.3934 ± 0.0413	6.03 ± 0.42	6.25 ± 0.30
	8	0.3942 ± 0.0401	0.3433 ± 0.0395	0.3903 ± 0.0390	6.22 ± 0.42	6.64 ± 0.35
	16	0.4116 ± 0.0391	0.3609 ± 0.0394	0.3968 ± 0.0394	6.35 ± 0.40	6.77 ± 0.32
LatentPersonal	1	0.3906 ± 0.0420	0.3425 ± 0.0415	0.3951 ± 0.0401	6.01 ± 0.42	6.44 ± 0.26
	2	0.3817 ± 0.0409	0.3431 ± 0.0396	0.3815 ± 0.0390	6.09 ± 0.43	6.46 ± 0.31
	4	0.3843 ± 0.0413	0.3482 ± 0.0406	0.3898 ± 0.0388	5.91 ± 0.45	6.31 ± 0.33
	8	0.3844 ± 0.0405	0.3478 ± 0.0397	0.3840 ± 0.0386	6.39 ± 0.43	6.84 ± 0.31
	16	0.4111 ± 0.0387	0.3562 ± 0.0400	0.3979 ± 0.0383	6.34 ± 0.43	6.72 ± 0.34

Table 8: Full personalization results on PRISM under different few-shot settings.

Method	Shot	R-1	R-L	MET.	CP	PP
Llama-3.1-8B-Instruct						
Base	/	0.2558 ± 0.0229	0.1524 ± 0.0137	0.2125 ± 0.0176	4.18 ± 0.42	4.08 ± 0.36
ICL	1	0.2929 ± 0.0209	0.1703 ± 0.0149	0.2372 ± 0.0199	4.11 ± 0.34	4.73 ± 0.24
	2	0.2917 ± 0.0204	0.1637 ± 0.0156	0.2233 ± 0.0194	4.71 ± 0.33	5.07 ± 0.24
	4	0.2998 ± 0.0201	0.1706 ± 0.0146	0.2263 ± 0.0189	4.82 ± 0.33	4.98 ± 0.26
	8	0.3009 ± 0.0200	0.1719 ± 0.0153	0.2276 ± 0.0198	5.84 ± 0.35	5.65 ± 0.27
	16	0.3146 ± 0.0204	0.1916 ± 0.0150	0.2528 ± 0.0205	6.56 ± 0.36	6.41 ± 0.27
LoRA	1	0.2843 ± 0.0253	0.1762 ± 0.0159	0.2046 ± 0.0213	4.05 ± 0.43	4.95 ± 0.36
	2	0.2649 ± 0.0236	0.1672 ± 0.0151	0.1940 ± 0.0218	4.63 ± 0.47	5.10 ± 0.38
	4	0.2568 ± 0.0245	0.1653 ± 0.0153	0.1873 ± 0.0208	4.97 ± 0.45	5.24 ± 0.37
	8	0.2638 ± 0.0243	0.1712 ± 0.0158	0.1926 ± 0.0221	5.23 ± 0.46	5.56 ± 0.36
	16	0.2782 ± 0.0263	0.1832 ± 0.0192	0.2015 ± 0.0213	5.75 ± 0.51	5.95 ± 0.43
PPlug	1	0.2875 ± 0.0253	0.1830 ± 0.0167	0.2033 ± 0.0212	4.11 ± 0.43	4.76 ± 0.36
	2	0.2806 ± 0.0235	0.1761 ± 0.0154	0.1951 ± 0.0205	4.55 ± 0.44	4.87 ± 0.38
	4	0.2853 ± 0.0235	0.1822 ± 0.0163	0.1950 ± 0.0203	4.97 ± 0.42	5.26 ± 0.37
	8	0.2889 ± 0.0225	0.1835 ± 0.0165	0.1945 ± 0.0198	5.54 ± 0.43	5.69 ± 0.37
	16	0.2730 ± 0.0205	0.1736 ± 0.0130	0.1851 ± 0.0170	5.65 ± 0.47	5.75 ± 0.38
Profile-to-PEFT	1	0.2895 ± 0.0214	0.1794 ± 0.0144	0.2200 ± 0.0192	4.15 ± 0.42	5.04 ± 0.32
	2	0.2888 ± 0.0223	0.1786 ± 0.0151	0.2141 ± 0.0190	4.42 ± 0.47	5.13 ± 0.38
	4	0.2949 ± 0.0226	0.1839 ± 0.0146	0.2221 ± 0.0200	5.28 ± 0.44	5.73 ± 0.36
	8	0.2928 ± 0.0232	0.1807 ± 0.0153	0.2187 ± 0.0206	5.68 ± 0.48	5.98 ± 0.39
	16	0.2941 ± 0.0226	0.1911 ± 0.0158	0.2145 ± 0.0167	6.19 ± 0.47	6.50 ± 0.40
LatentPersonal	1	0.3081 ± 0.0231	0.2026 ± 0.0161	0.2421 ± 0.0191	4.28 ± 0.43	5.18 ± 0.37
	2	0.3036 ± 0.0229	0.1996 ± 0.0155	0.2380 ± 0.0190	4.82 ± 0.44	5.46 ± 0.34
	4	0.3043 ± 0.0231	0.2009 ± 0.0155	0.2385 ± 0.0177	5.30 ± 0.46	5.75 ± 0.36
	8	0.3049 ± 0.0234	0.2021 ± 0.0155	0.2391 ± 0.0182	5.79 ± 0.48	6.06 ± 0.36
	16	0.3181 ± 0.0207	0.2048 ± 0.0164	0.2469 ± 0.0182	6.12 ± 0.47	6.43 ± 0.39
Qwen2.5-7B-Instruct						
Base	/	0.2713 ± 0.0219	0.1571 ± 0.0135	0.2236 ± 0.0186	4.16 ± 0.41	4.37 ± 0.36
ICL	1	0.2893 ± 0.0195	0.1695 ± 0.0132	0.1994 ± 0.0181	4.16 ± 0.43	4.64 ± 0.32
	2	0.2899 ± 0.0196	0.1732 ± 0.0118	0.1943 ± 0.0170	4.60 ± 0.43	4.93 ± 0.32
	4	0.2974 ± 0.0196	0.1701 ± 0.0118	0.2041 ± 0.0160	4.98 ± 0.40	4.95 ± 0.30
	8	0.3051 ± 0.0193	0.1751 ± 0.0122	0.2051 ± 0.0157	5.78 ± 0.38	5.59 ± 0.30
	16	0.3087 ± 0.0188	0.1763 ± 0.0129	0.2130 ± 0.0165	6.25 ± 0.38	6.08 ± 0.31
LoRA	1	0.2488 ± 0.0218	0.1557 ± 0.0145	0.1841 ± 0.0182	3.57 ± 0.38	4.29 ± 0.34
	2	0.2486 ± 0.0236	0.1517 ± 0.0150	0.1776 ± 0.0195	3.85 ± 0.43	4.46 ± 0.38
	4	0.2351 ± 0.0249	0.1540 ± 0.0176	0.1712 ± 0.0208	3.83 ± 0.45	4.32 ± 0.41
	8	0.2454 ± 0.0240	0.1511 ± 0.0146	0.1738 ± 0.0204	4.56 ± 0.47	4.97 ± 0.38
	16	0.2420 ± 0.0259	0.1553 ± 0.0167	0.1743 ± 0.0205	4.79 ± 0.53	5.12 ± 0.48
PPlug	1	0.3044 ± 0.0208	0.1812 ± 0.0144	0.2141 ± 0.0176	4.14 ± 0.43	4.65 ± 0.38
	2	0.2979 ± 0.0201	0.1797 ± 0.0139	0.2084 ± 0.0180	4.45 ± 0.44	4.69 ± 0.36
	4	0.2908 ± 0.0201	0.1751 ± 0.0144	0.1992 ± 0.0162	4.75 ± 0.44	5.01 ± 0.35
	8	0.2912 ± 0.0213	0.1760 ± 0.0151	0.1962 ± 0.0172	5.45 ± 0.46	5.56 ± 0.38
	16	0.2970 ± 0.0209	0.1865 ± 0.0150	0.2082 ± 0.0186	5.57 ± 0.48	5.68 ± 0.38
Profile-to-PEFT	1	0.2891 ± 0.0228	0.1719 ± 0.0138	0.2223 ± 0.0193	4.25 ± 0.42	5.05 ± 0.36
	2	0.2925 ± 0.0244	0.1760 ± 0.0150	0.2163 ± 0.0193	4.69 ± 0.45	5.19 ± 0.37
	4	0.2853 ± 0.0223	0.1724 ± 0.0129	0.2126 ± 0.0187	5.14 ± 0.44	5.53 ± 0.36
	8	0.2809 ± 0.0208	0.1716 ± 0.0129	0.2050 ± 0.0177	5.57 ± 0.45	5.85 ± 0.34
	16	0.2907 ± 0.0192	0.1825 ± 0.0130	0.2192 ± 0.0166	5.90 ± 0.47	6.21 ± 0.38
LatentPersonal	1	0.3008 ± 0.0229	0.1801 ± 0.0134	0.2251 ± 0.0194	4.42 ± 0.41	5.27 ± 0.34
	2	0.2946 ± 0.0216	0.1794 ± 0.0142	0.2209 ± 0.0192	4.79 ± 0.43	5.28 ± 0.33
	4	0.3020 ± 0.0227	0.1837 ± 0.0138	0.2255 ± 0.0193	5.33 ± 0.44	5.76 ± 0.34
	8	0.2977 ± 0.0223	0.1800 ± 0.0134	0.2247 ± 0.0188	5.59 ± 0.48	5.86 ± 0.36
	16	0.2994 ± 0.0204	0.1869 ± 0.0144	0.2363 ± 0.0185	6.31 ± 0.46	6.36 ± 0.36

Table 9: Full personalization results on PersonaMem under different few-shot settings.

Method	Shot	R-1	R-L	MET.	CP	PP
Llama-3.1-8B-Instruct						
Base	/	0.2368 ± 0.0099	0.1336 ± 0.0061	0.2103 ± 0.0101	5.01 ± 0.44	5.15 ± 0.30
ICL	1	0.2514 ± 0.0136	0.1421 ± 0.0096	0.2352 ± 0.0130	4.87 ± 0.34	5.21 ± 0.28
	2	0.2682 ± 0.0131	0.1522 ± 0.0089	0.2457 ± 0.0121	5.51 ± 0.33	5.65 ± 0.27
	4	0.2618 ± 0.0134	0.1512 ± 0.0091	0.2457 ± 0.0124	6.07 ± 0.35	6.18 ± 0.23
	8	0.2759 ± 0.0132	0.1613 ± 0.0088	0.2618 ± 0.0122	6.39 ± 0.32	6.49 ± 0.25
	16	0.2738 ± 0.0127	0.1659 ± 0.0084	0.2634 ± 0.0119	7.32 ± 0.32	7.29 ± 0.25
LoRA	1	0.3448 ± 0.0193	0.2311 ± 0.0174	0.2767 ± 0.0184	5.97 ± 0.38	6.16 ± 0.28
	2	0.3442 ± 0.0200	0.2286 ± 0.0156	0.2740 ± 0.0202	6.09 ± 0.38	6.32 ± 0.27
	4	0.3447 ± 0.0188	0.2253 ± 0.0150	0.2800 ± 0.0186	6.25 ± 0.39	6.52 ± 0.26
	8	0.3446 ± 0.0204	0.2302 ± 0.0176	0.2782 ± 0.0198	6.40 ± 0.36	6.69 ± 0.28
	16	0.3627 ± 0.0176	0.2367 ± 0.0136	0.2955 ± 0.0188	7.66 ± 0.32	7.16 ± 0.26
PPlug	1	0.2949 ± 0.0155	0.1860 ± 0.0114	0.2265 ± 0.0139	5.17 ± 0.36	5.47 ± 0.26
	2	0.2902 ± 0.0139	0.1837 ± 0.0101	0.2216 ± 0.0136	5.19 ± 0.34	5.45 ± 0.29
	4	0.2939 ± 0.0152	0.1860 ± 0.0117	0.2238 ± 0.0152	5.53 ± 0.36	5.76 ± 0.28
	8	0.2953 ± 0.0141	0.1909 ± 0.0109	0.2226 ± 0.0140	5.88 ± 0.32	5.97 ± 0.24
	16	0.2892 ± 0.0139	0.1818 ± 0.0096	0.2148 ± 0.0120	6.39 ± 0.28	6.37 ± 0.24
Profile-to-PEFT	1	0.3398 ± 0.0189	0.2269 ± 0.0162	0.2749 ± 0.0178	5.74 ± 0.36	6.32 ± 0.25
	2	0.3480 ± 0.0184	0.2286 ± 0.0160	0.2806 ± 0.0195	5.70 ± 0.37	6.51 ± 0.26
	4	0.3523 ± 0.0182	0.2325 ± 0.0157	0.2827 ± 0.0186	5.51 ± 0.35	6.69 ± 0.26
	8	0.3561 ± 0.0189	0.2390 ± 0.0169	0.2859 ± 0.0197	6.37 ± 0.33	6.42 ± 0.24
	16	0.3616 ± 0.0184	0.2397 ± 0.0151	0.2880 ± 0.0199	6.21 ± 0.30	7.55 ± 0.22
LatentPersonal	1	0.3534 ± 0.0181	0.2336 ± 0.0161	0.2803 ± 0.0200	6.14 ± 0.37	6.52 ± 0.25
	2	0.3631 ± 0.0184	0.2536 ± 0.0164	0.2930 ± 0.0202	6.27 ± 0.37	6.67 ± 0.27
	4	0.3647 ± 0.0178	0.2469 ± 0.0157	0.2996 ± 0.0208	6.35 ± 0.37	6.59 ± 0.27
	8	0.3718 ± 0.0180	0.2542 ± 0.0156	0.3031 ± 0.0201	6.42 ± 0.38	6.48 ± 0.29
	16	0.3646 ± 0.0180	0.2597 ± 0.0147	0.3196 ± 0.0193	6.64 ± 0.28	6.78 ± 0.25
Qwen2.5-7B-Instruct						
Base	/	0.2598 ± 0.0129	0.1405 ± 0.0070	0.2303 ± 0.0125	5.64 ± 0.44	5.69 ± 0.29
ICL	1	0.2245 ± 0.0121	0.1152 ± 0.0082	0.2043 ± 0.0122	4.93 ± 0.38	5.26 ± 0.25
	2	0.2376 ± 0.0141	0.1235 ± 0.0100	0.2192 ± 0.0127	5.46 ± 0.36	5.61 ± 0.23
	4	0.2484 ± 0.0123	0.1334 ± 0.0084	0.2294 ± 0.0111	6.07 ± 0.35	6.07 ± 0.24
	8	0.2440 ± 0.0138	0.1397 ± 0.0101	0.2284 ± 0.0142	6.50 ± 0.33	6.47 ± 0.26
	16	0.2517 ± 0.0125	0.1596 ± 0.0081	0.2401 ± 0.0122	7.24 ± 0.31	7.13 ± 0.26
LoRA	1	0.3166 ± 0.0199	0.2114 ± 0.0160	0.2611 ± 0.0205	6.03 ± 0.40	6.20 ± 0.29
	2	0.3276 ± 0.0195	0.2154 ± 0.0163	0.2652 ± 0.0194	6.19 ± 0.41	6.31 ± 0.29
	4	0.3399 ± 0.0198	0.2271 ± 0.0167	0.2747 ± 0.0212	6.47 ± 0.38	6.72 ± 0.28
	8	0.3417 ± 0.0169	0.2250 ± 0.0147	0.2740 ± 0.0187	6.65 ± 0.40	6.73 ± 0.31
	16	0.3555 ± 0.0185	0.2349 ± 0.0150	0.2834 ± 0.0202	7.32 ± 0.32	7.37 ± 0.26
PPlug	1	0.2947 ± 0.0167	0.1800 ± 0.0127	0.2332 ± 0.0171	5.16 ± 0.37	5.45 ± 0.29
	2	0.2950 ± 0.0167	0.1791 ± 0.0125	0.2282 ± 0.0170	5.32 ± 0.36	5.54 ± 0.29
	4	0.2984 ± 0.0162	0.1768 ± 0.0118	0.2295 ± 0.0165	5.46 ± 0.36	5.74 ± 0.29
	8	0.3039 ± 0.0167	0.1824 ± 0.0123	0.2277 ± 0.0163	5.68 ± 0.34	5.89 ± 0.27
	16	0.2990 ± 0.0150	0.1810 ± 0.0099	0.2300 ± 0.0138	6.32 ± 0.29	6.21 ± 0.28
Profile-to-PEFT	1	0.3230 ± 0.0197	0.2159 ± 0.0169	0.2594 ± 0.0195	6.14 ± 0.40	6.26 ± 0.29
	2	0.3253 ± 0.0201	0.2172 ± 0.0163	0.2538 ± 0.0193	6.31 ± 0.37	6.40 ± 0.26
	4	0.3190 ± 0.0205	0.2129 ± 0.0161	0.2592 ± 0.0193	6.60 ± 0.37	6.75 ± 0.26
	8	0.3223 ± 0.0204	0.2164 ± 0.0163	0.2639 ± 0.0189	6.72 ± 0.35	6.90 ± 0.25
	16	0.3469 ± 0.0172	0.2288 ± 0.0136	0.2738 ± 0.0186	7.33 ± 0.30	7.43 ± 0.24
LatentPersonal	1	0.3447 ± 0.0198	0.2268 ± 0.0162	0.2800 ± 0.0201	6.41 ± 0.37	6.39 ± 0.28
	2	0.3455 ± 0.0197	0.2289 ± 0.0178	0.2781 ± 0.0210	6.38 ± 0.36	6.48 ± 0.29
	4	0.3486 ± 0.0192	0.2259 ± 0.0171	0.2758 ± 0.0196	6.74 ± 0.34	6.70 ± 0.25
	8	0.3416 ± 0.0193	0.2240 ± 0.0164	0.2742 ± 0.0196	6.97 ± 0.36	6.96 ± 0.29
	16	0.3518 ± 0.0181	0.2278 ± 0.0128	0.2796 ± 0.0180	7.49 ± 0.33	7.47 ± 0.25

Table 10: Full personalization results on PIR under different few-shot settings.

Method	Shot	R-1	R-L	MET.	CP	PP
Llama-3.1-8B-Instruct						
Base	/	0.2954 ± 0.0142	0.1568 ± 0.0079	0.1966 ± 0.0087	4.08 ± 0.32	4.12 ± 0.25
ICL	1	0.3028 ± 0.0116	0.1643 ± 0.0067	0.2066 ± 0.0085	4.19 ± 0.30	4.90 ± 0.24
	2	0.3090 ± 0.0114	0.1691 ± 0.0065	0.2114 ± 0.0089	5.16 ± 0.31	5.80 ± 0.24
	4	0.3125 ± 0.0113	0.1707 ± 0.0065	0.2129 ± 0.0083	5.89 ± 0.33	6.24 ± 0.26
	8	0.3082 ± 0.0116	0.1677 ± 0.0066	0.2082 ± 0.0080	6.51 ± 0.32	6.71 ± 0.26
LoRA	1	0.3805 ± 0.0157	0.2070 ± 0.0113	0.2962 ± 0.0140	4.28 ± 0.35	5.35 ± 0.29
	2	0.3795 ± 0.0152	0.2079 ± 0.0104	0.2931 ± 0.0138	5.25 ± 0.33	6.00 ± 0.28
	4	0.3836 ± 0.0148	0.2059 ± 0.0086	0.2983 ± 0.0128	5.72 ± 0.33	6.47 ± 0.26
	8	0.3846 ± 0.0141	0.2063 ± 0.0083	0.2977 ± 0.0126	6.67 ± 0.26	7.31 ± 0.20
PPlug	1	0.3711 ± 0.0126	0.1997 ± 0.0084	0.2495 ± 0.0107	3.96 ± 0.30	4.88 ± 0.25
	2	0.3633 ± 0.0128	0.1950 ± 0.0080	0.2401 ± 0.0096	5.16 ± 0.33	5.71 ± 0.26
	4	0.3645 ± 0.0128	0.1929 ± 0.0078	0.2436 ± 0.0103	5.86 ± 0.32	6.02 ± 0.26
	8	0.3646 ± 0.0130	0.1943 ± 0.0078	0.2409 ± 0.0096	6.56 ± 0.28	6.79 ± 0.23
Profile-to-PEFT	1	0.3891 ± 0.0155	0.2099 ± 0.0103	0.3066 ± 0.0137	4.08 ± 0.33	5.45 ± 0.31
	2	0.3952 ± 0.0153	0.2161 ± 0.0110	0.3111 ± 0.0140	5.13 ± 0.31	6.15 ± 0.25
	4	0.3857 ± 0.0153	0.2168 ± 0.0121	0.3024 ± 0.0138	5.75 ± 0.33	6.77 ± 0.27
	8	0.3880 ± 0.0152	0.2144 ± 0.0112	0.3045 ± 0.0133	6.63 ± 0.32	7.35 ± 0.23
LatentPersonal	1	0.3932 ± 0.0146	0.2092 ± 0.0113	0.3087 ± 0.0136	4.38 ± 0.31	5.62 ± 0.27
	2	0.3926 ± 0.0144	0.2089 ± 0.0111	0.3094 ± 0.0128	5.30 ± 0.34	6.25 ± 0.27
	4	0.3912 ± 0.0150	0.2101 ± 0.0117	0.3084 ± 0.0139	5.90 ± 0.33	6.57 ± 0.27
	8	0.3852 ± 0.0149	0.2166 ± 0.0098	0.3163 ± 0.0128	6.58 ± 0.31	7.34 ± 0.24
Qwen2.5-7B-Instruct						
Base	/	0.2871 ± 0.0123	0.1581 ± 0.0069	0.1919 ± 0.0093	3.86 ± 0.31	4.28 ± 0.25
ICL	1	0.2981 ± 0.0129	0.1483 ± 0.0073	0.2012 ± 0.0093	4.08 ± 0.32	4.98 ± 0.28
	2	0.3014 ± 0.0135	0.1583 ± 0.0071	0.2194 ± 0.0099	5.06 ± 0.31	5.66 ± 0.24
	4	0.3123 ± 0.0143	0.1688 ± 0.0070	0.2123 ± 0.0101	5.81 ± 0.30	6.16 ± 0.25
	8	0.3103 ± 0.0139	0.1691 ± 0.0071	0.2217 ± 0.0101	6.50 ± 0.31	6.78 ± 0.29
LoRA	1	0.3756 ± 0.0163	0.1995 ± 0.0108	0.2902 ± 0.0150	4.17 ± 0.33	5.45 ± 0.28
	2	0.3827 ± 0.0158	0.2033 ± 0.0101	0.2945 ± 0.0143	5.13 ± 0.33	6.12 ± 0.28
	4	0.3868 ± 0.0167	0.2058 ± 0.0109	0.2954 ± 0.0157	6.12 ± 0.30	6.71 ± 0.27
	8	0.3806 ± 0.0157	0.2020 ± 0.0099	0.2933 ± 0.0140	6.51 ± 0.30	7.30 ± 0.23
PPlug	1	0.3404 ± 0.0100	0.1754 ± 0.0064	0.2272 ± 0.0082	4.00 ± 0.29	4.87 ± 0.23
	2	0.3351 ± 0.0105	0.1740 ± 0.0056	0.2255 ± 0.0086	4.88 ± 0.33	5.45 ± 0.27
	4	0.3397 ± 0.0113	0.1744 ± 0.0063	0.2288 ± 0.0096	5.79 ± 0.29	6.13 ± 0.25
	8	0.3483 ± 0.0107	0.1811 ± 0.0064	0.2324 ± 0.0089	6.35 ± 0.30	6.56 ± 0.26
Profile-to-PEFT	1	0.3958 ± 0.0160	0.2096 ± 0.0107	0.3073 ± 0.0151	4.18 ± 0.32	5.31 ± 0.26
	2	0.3891 ± 0.0142	0.2049 ± 0.0096	0.2993 ± 0.0121	5.20 ± 0.29	6.08 ± 0.22
	4	0.3871 ± 0.0152	0.2056 ± 0.0095	0.2988 ± 0.0134	5.85 ± 0.32	6.62 ± 0.27
	8	0.3903 ± 0.0159	0.2066 ± 0.0097	0.3015 ± 0.0131	6.47 ± 0.29	7.16 ± 0.22
LatentPersonal	1	0.3975 ± 0.0156	0.2074 ± 0.0097	0.3099 ± 0.0126	4.42 ± 0.32	5.75 ± 0.27
	2	0.3940 ± 0.0163	0.2100 ± 0.0100	0.3094 ± 0.0141	5.23 ± 0.32	6.20 ± 0.25
	4	0.3935 ± 0.0162	0.2099 ± 0.0100	0.3031 ± 0.0141	5.94 ± 0.33	6.73 ± 0.28
	8	0.3946 ± 0.0156	0.2079 ± 0.0096	0.2990 ± 0.0133	6.48 ± 0.29	7.48 ± 0.23

D Additional Experiments

D.1 Extended Personalization Results

The main text presents a subset of representative evaluation metrics for clarity. Here, we provide the full results across all metrics for Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and further report results for Qwen2.5-7B-Instruct (Yang et al., 2025) to assess whether the observed trends generalize across backbone models.

As shown in Tables 7-10, the full evaluation results are consistent with the observations in the main text. *LatentPersonal* performs particularly well in the low-shot regime on Llama-3.1-8B-Instruct, with strong performance across both lexical and preference-oriented metrics, especially on HiCUPID and PersonaMem. Similar behavior is observed with Qwen2.5-7B-Instruct, although the differences among adaptation-based methods become smaller on several datasets. These results indicate that the observed benefits of latent-guided personalization extend across evaluation metrics and are not specific to a single backbone model. We additionally report 95% confidence intervals, computed as $\mathrm{mean} \pm 1.96\,\mathrm{SE}$ over held-out test users.

D.2 Extended Ablation Results

Table 11 provides the full ablation results across all evaluation metrics and shot settings. Removing the latent representation (*w/o Latent*) leads to a substantial performance degradation across all settings, highlighting the importance of user-specific latent conditioning. Disabling latent navigation (*w/o Navigation*) results in a consistent drop in lexical metrics, particularly in the low-shot regime, while the gap narrows as more observations become available. Removing the information bottleneck (*w/o IB*) causes the largest overall degradation, suggesting that regularizing the latent representation plays an important role in learning effective user-specific adaptation from limited observations. Overall, these results support the complementary roles of latent representation, latent navigation, and information bottleneck regularization in *LatentPersonal*.

Table 11: Full ablation results of *LatentPersonal* on HiCUPID with Llama-3.1-8B-Instruct across different numbers of observed user interactions.

Variant	Shot	R-1	R-L	MET.	CP	PP
LatentPersonal	1	0.4162 ± 0.0408	0.3637 ± 0.0400	0.4133 ± 0.0404	5.95 ± 0.41	6.33 ± 0.31
	2	0.4160 ± 0.0415	0.3656 ± 0.0404	0.4142 ± 0.0412	5.92 ± 0.40	6.28 ± 0.30
	4	0.4156 ± 0.0412	0.3639 ± 0.0403	0.4137 ± 0.0412	5.94 ± 0.43	6.27 ± 0.33
	8	0.4162 ± 0.0412	0.3652 ± 0.0405	0.4135 ± 0.0408	6.24 ± 0.43	6.61 ± 0.32
	16	0.4323 ± 0.0428	0.3851 ± 0.0428	0.4280 ± 0.0424	6.44 ± 0.43	6.72 ± 0.36
w/o Latent	1	0.3550 ± 0.0260	0.3011 ± 0.0226	0.3667 ± 0.0231	4.72 ± 0.40	5.85 ± 0.34
	2	0.3499 ± 0.0230	0.2925 ± 0.0199	0.3841 ± 0.0231	4.86 ± 0.38	5.92 ± 0.33
	4	0.3308 ± 0.0163	0.2757 ± 0.0112	0.3423 ± 0.0141	5.29 ± 0.36	6.16 ± 0.35
	8	0.3479 ± 0.0187	0.2742 ± 0.0142	0.3475 ± 0.0184	5.51 ± 0.35	6.21 ± 0.32
	16	0.3478 ± 0.0158	0.2807 ± 0.0112	0.3531 ± 0.0150	5.36 ± 0.36	6.02 ± 0.36
w/o Navigation	1	0.3812 ± 0.0418	0.3298 ± 0.0411	0.3777 ± 0.0404	5.66 ± 0.45	6.12 ± 0.27
	2	0.3910 ± 0.0423	0.3452 ± 0.0420	0.3929 ± 0.0402	5.71 ± 0.44	6.26 ± 0.31
	4	0.3909 ± 0.0420	0.3415 ± 0.0401	0.3992 ± 0.0407	5.92 ± 0.45	6.07 ± 0.36
	8	0.3838 ± 0.0389	0.3347 ± 0.0377	0.3870 ± 0.0402	6.12 ± 0.49	6.52 ± 0.36
	16	0.4149 ± 0.0410	0.3643 ± 0.0418	0.4112 ± 0.0415	6.49 ± 0.39	6.88 ± 0.34
w/o IB	1	0.2684 ± 0.0079	0.2510 ± 0.0051	0.3423 ± 0.0114	4.57 ± 0.31	4.72 ± 0.27
	2	0.2703 ± 0.0083	0.2511 ± 0.0052	0.3422 ± 0.0118	4.65 ± 0.30	4.64 ± 0.28
	4	0.2695 ± 0.0079	0.2515 ± 0.0050	0.3444 ± 0.0112	4.74 ± 0.33	4.55 ± 0.29
	8	0.2674 ± 0.0077	0.2502 ± 0.0049	0.3413 ± 0.0112	4.64 ± 0.30	4.33 ± 0.27
	16	0.2740 ± 0.0072	0.2541 ± 0.0043	0.3504 ± 0.0101	4.45 ± 0.30	4.12 ± 0.27

D.3 SENSITIVITY ANALYSIS

We examine the sensitivity of *LatentPersonal* to two key hyperparameters: the information bottleneck coefficient β and the latent dimension d_z.

Effect of the Information Bottleneck Coefficient. Table 12 examines the effect of the information bottleneck coefficient β across different numbers of observed interactions. Setting $\beta = 0$ leads to a substantial degradation across all metrics and shot settings, highlighting the importance of regularizing the latent representation. With non-zero regularization, performance improves substantially and remains relatively stable across different values of β, although the optimal value varies across metrics and shot settings. We use $\beta = 0.01$ as the default, which provides consistently strong performance, particularly on the lexical metrics.

Table 12: Sensitivity analysis of the information bottleneck coefficient β on HiCUPID with Llama-3.1-8B-Instruct across different numbers of observed user interactions.

β	Shot	R-1	R-L	MET.	CP	PP
0	1	0.2684	0.2510	0.3423	4.57	4.72
	2	0.2703	0.2511	0.3422	4.65	4.64
	4	0.2695	0.2515	0.3444	4.74	4.55
	8	0.2674	0.2502	0.3413	4.64	4.33
	16	0.2740	0.2541	0.3504	4.45	4.12
0.1	1	0.3848	0.3386	0.3957	6.28	6.40
	2	0.3858	0.3379	0.3918	6.30	6.47
	4	0.3899	0.3428	0.3983	5.94	6.34
	8	0.3835	0.3400	0.3945	6.51	6.74
	16	0.4349	0.3774	0.4318	6.60	6.83
0.01	1	0.4162	0.3637	0.4133	5.95	6.33
	2	0.4160	0.3656	0.4142	5.92	6.28
	4	0.4156	0.3639	0.4137	5.94	6.27
	8	0.4162	0.3652	0.4135	6.24	6.61
	16	0.4323	0.3851	0.4280	6.44	6.72
0.001	1	0.3901	0.3381	0.4017	5.64	6.13
	2	0.3881	0.3360	0.3953	6.05	6.31
	4	0.3887	0.3369	0.3978	5.87	6.36
	8	0.3898	0.3371	0.3980	6.20	6.62
	16	0.4265	0.3702	0.4300	6.15	6.61

Effect of the Latent Dimension. Table 13 studies the effect of the latent dimension d_z across different numbers of observed interactions. In our LoRA-based instantiation, we set $d_z = r$, such that each latent dimension directly modulates a low-rank adaptation direction. We observe no consistent benefit from increasing the dimensionality: compact latent spaces already achieve strong performance, with $d_z = 4$ performing particularly well on several lexical metrics, while larger dimensions provide occasional gains on preference-oriented metrics. Overall, these results suggest that effective personalization does not require a high-dimensional adaptation space, supporting our motivation of navigating a compact latent space.

D.4 TEST-TIME PERSONALIZATION EFFICIENCY

We further examine the test-time efficiency of different personalization methods on HiCUPID using Llama-3.1-8B-Instruct. We separately measure *personalization latency*, defined as the method-specific test-time adaptation cost before response generation, and *generation latency*, defined as the subsequent response generation time after personalization. For ICL, personalization consists of constructing the user-specific in-context prompt; for LoRA, it involves gradient-based adaptation on the observed interactions; and for PPlug, Profile-to-PEFT, and *LatentPersonal*, it corresponds to their

Table 13: Sensitivity analysis of the latent adaptation dimension ($d_z = r$) on HiCUPID with Llama-3.1-8B-Instruct across different numbers of observed user interactions.

d_z	Shot	R-1	R-L	MET.	CP	PP
4	1	0.4169	0.3681	0.4217	6.12	6.52
	2	0.4176	0.3702	0.4214	5.87	6.17
	4	0.4195	0.3719	0.4268	5.91	6.37
	8	0.4153	0.3677	0.4204	6.32	6.64
	16	0.4345	0.3837	0.4330	6.27	6.68
8	1	0.3981	0.3483	0.4191	6.02	6.47
	2	0.4002	0.3497	0.4193	6.25	6.54
	4	0.4027	0.3532	0.4219	6.21	6.58
	8	0.3986	0.3476	0.4172	6.55	6.88
	16	0.4240	0.3775	0.4282	6.53	6.83
16	1	0.4162	0.3637	0.4133	5.95	6.33
	2	0.4160	0.3656	0.4142	5.92	6.28
	4	0.4156	0.3639	0.4137	5.94	6.27
	8	0.4162	0.3652	0.4135	6.24	6.61
	16	0.4323	0.3851	0.4280	6.44	6.72
32	1	0.3938	0.3455	0.4003	6.16	6.42
	2	0.3919	0.3411	0.3988	6.17	6.37
	4	0.3889	0.3436	0.4006	6.34	6.56
	8	0.3903	0.3423	0.3990	6.37	6.67
	16	0.4316	0.3789	0.4286	6.57	6.74
64	1	0.3962	0.3483	0.4120	6.10	6.51
	2	0.3975	0.3514	0.4173	6.20	6.37
	4	0.3989	0.3498	0.4151	6.19	6.55
	8	0.4010	0.3550	0.4203	6.54	6.96
	16	0.4318	0.3784	0.4390	6.55	6.75

respective forward personalization procedures. In particular, for *LatentPersonal*, personalization latency measures the forward inference required to obtain the user-specific latent representation from the observed interactions.

Table 14: Extended test-time personalization efficiency results on HiCUPID with Llama-3.1-8B-Instruct. Personalization latency measures the cost required to incorporate observed user interactions before response generation, while generation latency measures the subsequent response generation cost. Lower is better.

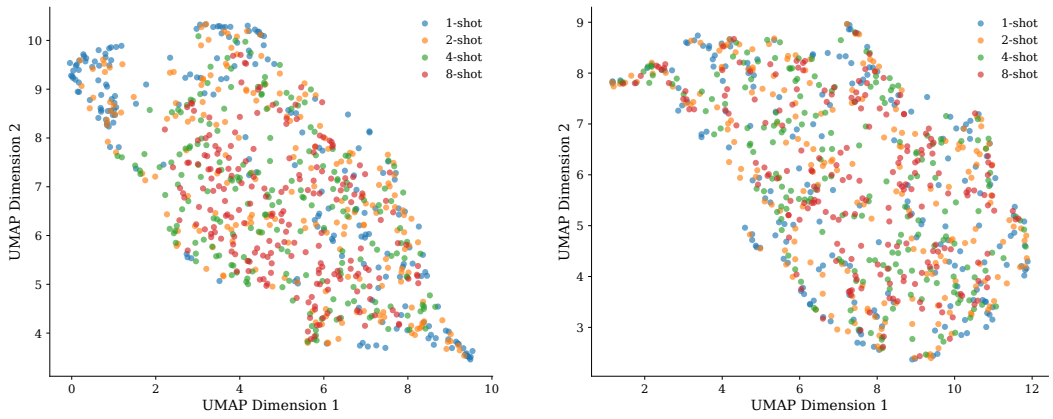
Method	Personalization Latency (s)				Generation Latency (s)			
	1-shot	2-shot	4-shot	8-shot	1-shot	2-shot	4-shot	8-shot
ICL	/	/	/	/	2.0896	2.1081	2.1571	2.2302
LoRA	0.3718	0.3622	0.3692	0.3640	0.6647	0.6993	0.7357	0.7550
PPlug	0.0004	0.0006	0.0009	0.0014	0.4597	0.4283	0.4861	0.4818
Profile-to-PEFT	0.0539	0.0538	0.0539	0.0544	0.8758	0.9091	0.8278	0.9599
LatentPersonal	0.0219	0.0304	0.0404	0.0676	0.8996	0.9145	0.8907	0.8838

As shown in Table 14, *LatentPersonal* introduces only modest personalization overhead, requiring 0.0219-0.0676 seconds across the 1-8-shot settings. Although its personalization latency increases with the number of observed interactions, it remains substantially lower than the gradient-based LoRA adaptation cost across all settings and is lower than Profile-to-PEFT in the 1-4-shot regime. PPlug incurs the lowest personalization latency, reflecting its lightweight forward personalization procedure. Importantly, the generation latency of *LatentPersonal* remains relatively stable as the number of observations increases, indicating that incorporating additional user evidence primarily

affects the personalization stage rather than the subsequent generation process. Overall, these results highlight the efficiency advantage of performing user adaptation through forward latent inference without requiring per-user gradient-based optimization.

D.5 Latent Space Analysis

Visualization of the Learned Latent Space Figure 5 visualizes the learned user representations under different numbers of observed interactions. Across all four datasets, representations inferred from different shot settings largely occupy a shared latent structure rather than forming isolated shot-specific clusters. At the same time, their distributions gradually shift as additional interactions are incorporated. This pattern is consistent with our view of personalization as navigation within a shared latent space, where additional user evidence refines the inferred position rather than inducing a separate representation space for each interaction regime.

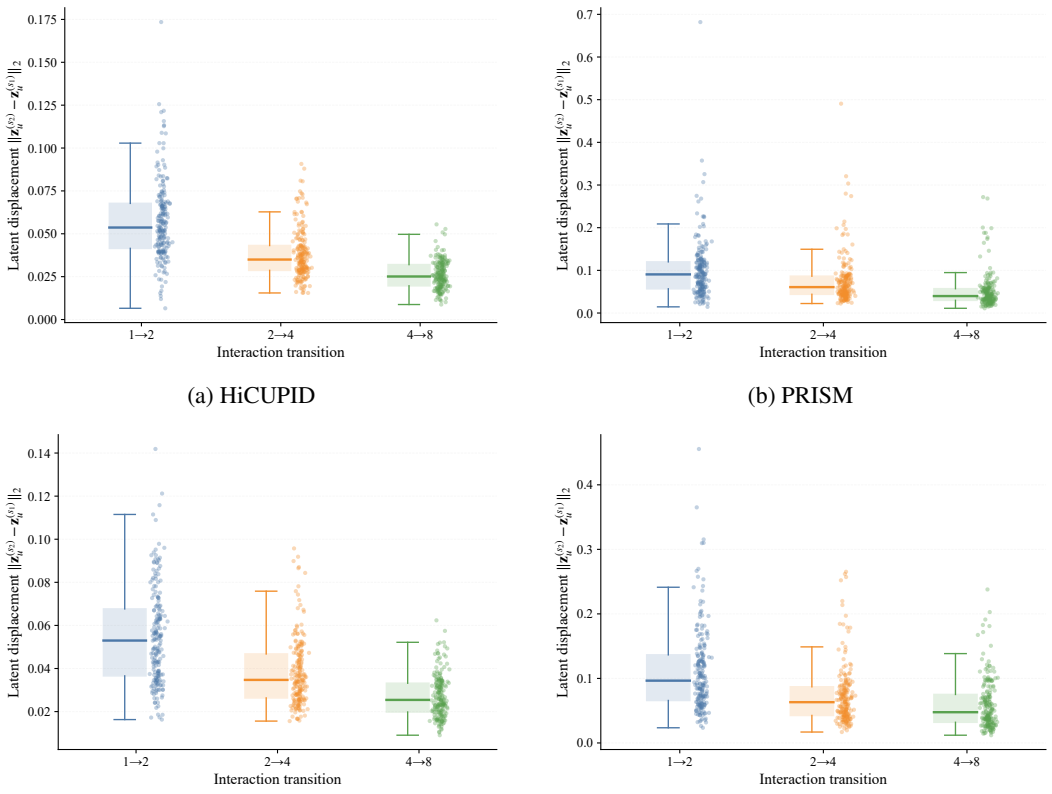


(a) HiCUPID (b) PRISM

(c) PersonaMem (d) PIR

Figure 5: UMAP visualization of learned user latent representations across four personalization datasets. Colors indicate the number of observed user interactions (1, 2, 4, and 8 shots).

Latent Stabilization across Interactions. To quantify how the inferred user representation evolves as additional interactions are observed, we measure the latent displacement $\|\mathbf{z}_u^{(s_2)} - \mathbf{z}_u^{(s_1)}\|_2$ for each user between consecutive shot settings, i.e., $1 \to 2$, $2 \to 4$, and $4 \to 8$. As shown in Figure 6, the displacement progressively decreases across interaction transitions on all four datasets, with the largest changes generally occurring in the earliest stage. This suggests that the first few interactions substantially update the inferred user position, whereas subsequent observations increasingly refine an already established latent representation.



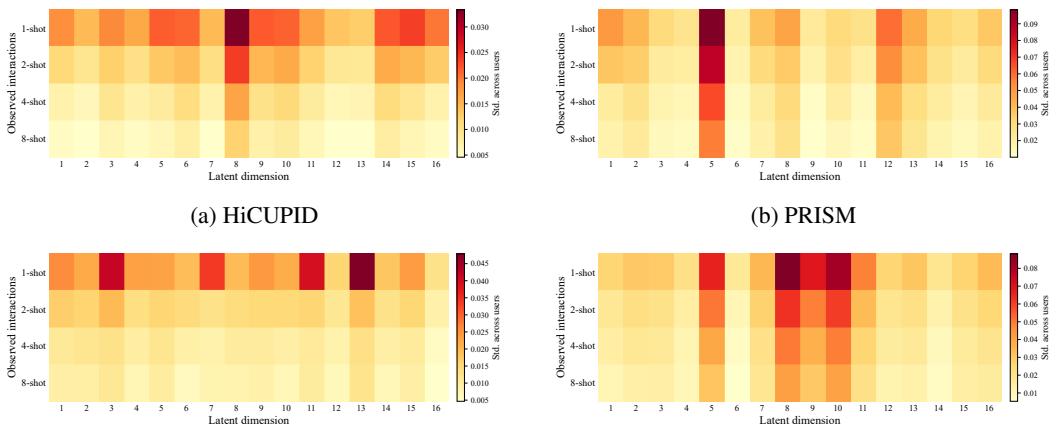
(a) HiCUPID (b) PRISM

(c) PersonaMem (d) PIR

Figure 6: Latent displacement across successive interaction transitions on four personalization datasets. Lower displacement at later transitions indicates that the inferred user representation gradually stabilizes as more interactions are observed.

Dimension-wise Latent Variability. We further examine how user-specific variation is distributed across the latent dimensions. For each shot setting and latent dimension, we compute the standard deviation of the corresponding latent coordinate across users. As shown in Figure 7, the variability is unevenly distributed across dimensions, with a subset of dimensions exhibiting substantially larger cross-user variation. Moreover, the overall variability tends to decrease as more interactions are observed across all four datasets. These results suggest that user-specific variation is concentrated along particular latent coordinates, while the inferred representations become progressively more stable as additional user evidence is incorporated.

Cross-shot Latent Alignment. We next examine whether the organization of the latent space is preserved as additional user interactions are incorporated. For each pair of consecutive shot settings, we compute the Pearson correlation across users between every pair of latent dimensions, producing a $d_z \times d_z$ cross-shot correlation matrix. A strong diagonal entry therefore indicates that the same latent coordinate preserves similar user-wise variation across two consecutive interaction regimes, whereas off-diagonal correlations capture dependencies between different coordinates.



(a) HiCUPID (b) PRISM

(c) PersonaMem (d) PIR

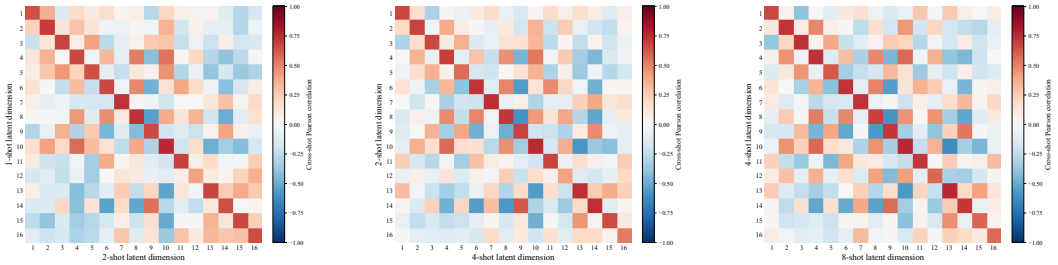
Figure 7: Dimension-wise variability of learned user latent representations across different numbers of observed interactions on four personalization datasets. Each cell reports the standard deviation of a latent dimension across users.

As shown in Figure 8, a clear diagonal structure persists from $1 \rightarrow 2$ through $4 \rightarrow 8$ across the datasets, indicating that corresponding latent coordinates remain aligned rather than being arbitrarily reorganized as more observations are introduced. The strength and surrounding correlation structure nevertheless differ across datasets. HiCUPID exhibits particularly clear diagonal correspondence with comparatively moderate off-diagonal correlations, while PersonaMem also maintains a stable coordinate-wise pattern across successive transitions. PIR exhibits substantially stronger positive and negative off-diagonal structure, suggesting greater coupling among latent coordinates, while still retaining correspondence along the diagonal. These dataset-specific patterns indicate that the latent representation need not consist of independent dimensions; instead, different coordinates can evolve jointly while preserving a recognizable cross-shot organization.

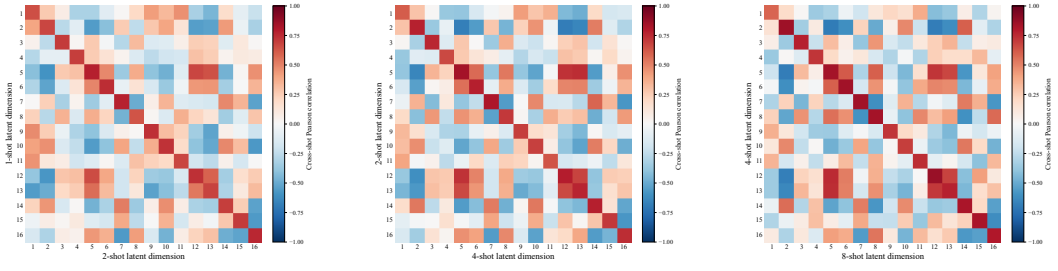
Together with the decreasing latent displacement in Figure 6, these results provide complementary evidence for a progressively refined representation: additional interactions move the inferred user representation while largely preserving the underlying latent coordinate structure. This behavior is consistent with the view of personalization as navigation within a shared latent adaptation space rather than repeatedly constructing unrelated representations as new user evidence becomes available.

Latent Scaling Analysis. We further examine the navigation interpretation underlying *LatentPersonal*. In our formulation, navigation refers to controlling user-specific adaptation through movement in the learned latent space, where the inferred $\mathbf{z}_u$ specifies a user-dependent direction in the shared adaptation space. To directly probe this behavior, we scale the inferred user latent as $\tilde{\mathbf{z}}_u = \alpha \mathbf{z}_u$, where $\alpha \in \{0, 0.25, 0.5, 0.75, 1, 1.5, 2\}$. This intervention preserves the inferred direction while varying how far the model moves along it: $\alpha = 0$ removes the user-specific latent signal, $\alpha = 1$ recovers the original representation, and $\alpha > 1$ moves beyond the inferred operating point along the same direction. It therefore provides a direct way to examine whether controlled movement in the latent space induces systematic changes in personalized behavior.

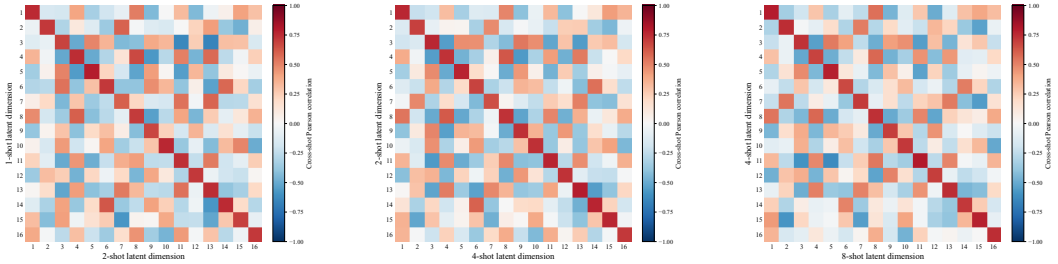
As shown in Table 15, removing the latent signal ($\alpha = 0$) causes a substantial performance degradation across all shot settings and evaluation metrics. Performance recovers rapidly as the model moves along the inferred direction, with $\alpha = 0.5$ already approaching the original representation ($\alpha = 1$) on several metrics. The inferred operating point at $\alpha = 1$ provides consistently strong performance across shot settings, whereas moving further along the same direction ($\alpha > 1$) yields no systematic improvement and generally degrades performance. This non-monotonic behavior indicates that personalization depends not only on identifying a user-specific direction but also on the position along that direction: moving away from the origin rapidly recovers personalized behavior, while excessive movement can move the model beyond a favorable personalization region. Taken together, these controlled interventions show that movement along the inferred latent direction pro-



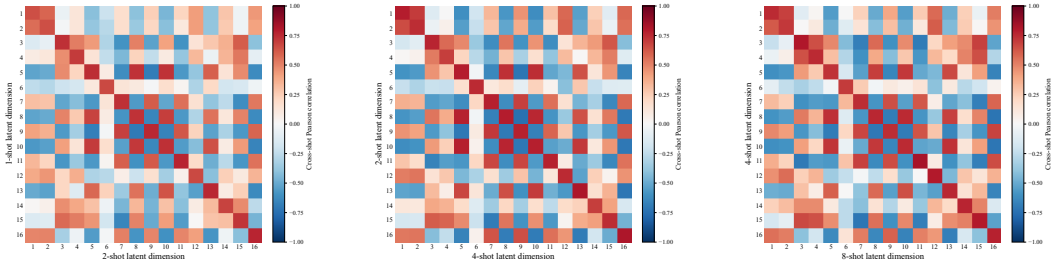
(a) HiCUPID: $1 \rightarrow 2$ (b) HiCUPID: $2 \rightarrow 4$ (c) HiCUPID: $4 \rightarrow 8$



(d) PRISM: $1 \rightarrow 2$ (e) PRISM: $2 \rightarrow 4$ (f) PRISM: $4 \rightarrow 8$



(g) PersonaMem: $1 \rightarrow 2$ (h) PersonaMem: $2 \rightarrow 4$ (i) PersonaMem: $4 \rightarrow 8$



(j) PIR: $1 \rightarrow 2$ (k) PIR: $2 \rightarrow 4$ (l) PIR: $4 \rightarrow 8$

Figure 8: Cross-shot alignment of learned latent dimensions across personalization datasets. Each matrix reports the Pearson correlation across users between latent dimensions inferred under two consecutive shot settings. The diagonal structure indicates consistent alignment of corresponding latent coordinates as additional user interactions are observed.

duces systematic changes in personalization behavior, providing empirical support for interpreting *LatentPersonal* as navigation in a shared adaptation space.

Table 15: Effect of latent scaling factor α across different numbers of observed user interactions on HiCUPID.

α	Shot	R-1	R-L	MET.	CP	PP
0	/	0.1614	0.1097	0.2166	3.57	4.29
0.25	1	0.3491	0.2958	0.3513	5.12	5.60
	2	0.3596	0.3109	0.3648	5.42	5.92
	4	0.3575	0.3088	0.3633	5.28	5.63
	8	0.3602	0.3077	0.3644	5.67	6.05
0.5	1	0.3941	0.3487	0.3972	5.75	6.42
	2	0.3896	0.3470	0.3945	5.75	6.22
	4	0.3989	0.3533	0.4002	6.07	6.33
	8	0.3922	0.3458	0.3922	6.08	6.67
0.75	1	0.3832	0.3407	0.3891	5.72	6.28
	2	0.3875	0.3466	0.3893	5.82	6.28
	4	0.3847	0.3437	0.3897	5.82	6.23
	8	0.3871	0.3462	0.3907	6.13	6.72
1	1	0.4162	0.3637	0.4133	5.95	6.33
	2	0.4160	0.3656	0.4142	5.92	6.28
	4	0.4156	0.3639	0.4137	5.94	6.27
	8	0.4162	0.3652	0.4135	6.24	6.61
1.5	1	0.3810	0.3413	0.3885	5.63	6.07
	2	0.3838	0.3420	0.3918	5.45	6.03
	4	0.3823	0.3403	0.3899	5.78	6.23
	8	0.3855	0.3446	0.3929	5.93	6.58
2	1	0.3879	0.3408	0.3869	5.63	6.20
	2	0.3739	0.3304	0.3774	5.72	6.30
	4	0.3735	0.3293	0.3704	5.77	6.07
	8	0.3698	0.3247	0.3760	5.97	6.50

User-Specificity Analysis. The latent-space visualization in Figure 5 suggests that different users occupy distinct regions within a shared latent structure. We therefore further examine whether these observed differences correspond to genuinely user-specific adaptation, rather than merely reflecting population-level adaptation learned from the training users. For each target test user u, we keep the trained model and corresponding held-out evaluation queries fixed and compare four different adaptation conditions: (i) *Matched*, using the latent representation $\mathbf{z}_u$ inferred from the target user's support interactions; (ii) *Swapped*, replacing $\mathbf{z}_u$ with a latent representation $\mathbf{z}_v$ inferred from another test user $v \neq u$; (iii) *Mean*, using a population-level latent representation $\bar{\mathbf{z}}$ as generic conditioning; and (iv) *Shared LoRA*, a separately trained LoRA model without any user-specific latent conditioning. The first three conditions share exactly the same trained *LatentPersonal* model and differ only in the latent representation supplied during inference. This controlled comparison isolates whether matching the inferred latent representation to the target user provides additional personalization benefits beyond mismatched or population-level conditioning, while the shared LoRA control measures the contribution of user-specific latent modulation beyond shared parameter adaptation alone.

For the *Swapped* condition, we replace the target user's representation $\mathbf{z}_u$ with $\mathbf{z}_v$ inferred from the support interactions of another test user $v \neq u$, while keeping the target user's original held-out evaluation queries unchanged. For each target user, we randomly sample five other users as independent swap sources and average the resulting scores to reduce sensitivity to any particular user pairing. The same target–source user mappings are maintained across all shot settings, so that only the number of support interactions used to infer each representation changes. For the *Mean* condition, we use the average latent representation computed over the training-user population, providing a generic population-level latent-conditioning control without access to the target

user's inferred representation. For the *Shared LoRA* condition, we train a separate shared LoRA model on the same training-user data and under the same training budget as *LatentPersonal*, while removing user-specific latent conditioning throughout the entire training process. Equivalently, the latent modulation is fixed to $\mathbf{z} = \mathbf{1}$ for every user, yielding $B\,\mathrm{diag}(\mathbf{1})A = BA$. This provides a population-level adaptation control that retains the same low-rank adaptation structure but contains no user-specific latent information during personalization.

Table 16: User-specificity analysis across different numbers of observed user interactions on HiCUPID.

Latent	Shot	R-1	R-L	MET.	CP	PP
Matched $\mathbf{z}_u$	1	0.4162	0.3637	0.4133	5.95	6.33
	2	0.4160	0.3656	0.4142	5.92	6.28
	4	0.4156	0.3639	0.4137	5.94	6.27
	8	0.4162	0.3652	0.4135	6.24	6.61
Swapped $\mathbf{z}_v$	1	0.3373	0.2965	0.3501	5.48	5.77
	2	0.3330	0.2841	0.3478	5.54	5.84
	4	0.3114	0.2818	0.3441	5.45	5.72
	8	0.3084	0.2765	0.3329	5.42	5.79
Mean $\bar{\mathbf{z}}$	1	0.3792	0.3245	0.3781	5.64	5.89
	2	0.3841	0.3285	0.3804	5.62	5.94
	4	0.3828	0.3398	0.3881	5.69	5.91
	8	0.3853	0.3357	0.3956	5.63	5.87
Shared LoRA	/	0.3712	0.3039	0.3425	5.38	5.34

The matched representation $\mathbf{z}_u$ consistently achieves the strongest performance across all shot settings, while replacing it with either a mismatched representation $\mathbf{z}_v$ or the population mean $\bar{\mathbf{z}}$ leads to clear degradation. Notably, the swapped condition exhibits a substantially larger drop in R-1, R-L, and METEOR than in CP and PP. This discrepancy is consistent with the different sensitivities of these metrics: reference-based metrics directly penalize deviations from the target users response, whereas CP and PP can remain relatively high when the generated response still follows broadly plausible personalization patterns shared across users. In contrast, the population-mean representation avoids navigating toward an incorrect individual user and therefore yields more stable performance than the swapped representation on most reference-based metrics. These results support the navigation view of *LatentPersonal*: effective personalization depends not only on entering the learned adaptation space, but on navigating toward the region associated with the target user.

D.6 Evaluation across PEFT Architectures

To examine whether the effectiveness of *LatentPersonal* depends on the specific PEFT architecture, we further evaluate the framework with Adapter and Prompt Tuning as alternative parameterizations. Their corresponding instantiations are described in Appendix E.

As shown in Table 17, incorporating *LatentPersonal* into Adapter consistently improves the lexical metrics across all shot settings, with particularly pronounced gains in the low-shot regime. For Prompt Tuning, latent conditioning similarly provides substantial improvements in the lexical metrics when only a few user interactions are available, while the advantage generally narrows as more observations are introduced. Preference-oriented metrics exhibit a more mixed pattern, indicating that the benefit of latent conditioning can depend on both the underlying PEFT parameterization and the evaluation criterion. Overall, the improvements observed with both Adapter and Prompt Tuning, particularly under limited user supervision, suggest that the latent personalization framework is not tied to the LoRA-based instantiation used in our main experiments and can be applied to different PEFT adaptation spaces.

E *LatentPersonal* ACROSS PEFT ARCHITECTURES

While our main implementation adopts LoRA as the primary PEFT parameterization, the latent navigation principle of *LatentPersonal* is not tied to a specific PEFT parameterization. In this section, we briefly illustrate how the same formulation can be instantiated with other representative PEFT architectures, including Adapter and Prompt Tuning modules.

E.1 ADAPTER INSTANTIATION

For a standard Adapter, the hidden representation at layer l is updated as

$$\tilde{\mathbf{h}}^{(l)} = \mathbf{h}^{(l)} + W_{\text{up}}^{(l)} \sigma\left(W_{\text{down}}^{(l)} \mathbf{h}^{(l)}\right), \tag{33}$$

where $W_{\text{down}}^{(l)}$ and $W_{\text{up}}^{(l)}$ denote the down- and up-projection matrices, respectively, and $\sigma(\cdot)$ denotes a nonlinear activation function.

The Adapter architecture naturally induces a low-dimensional intermediate space through $W_{\text{down}}^{(l)}$, which is directly compatible with the latent representation in *LatentPersonal*. We therefore instantiate *LatentPersonal* with Adapter by introducing the user latent z_u into this intermediate space:

$$\tilde{\mathbf{h}}_u^{(l)} = \mathbf{h}^{(l)} + W_{\text{up}}^{(l)} \operatorname{diag}(z_u) \sigma\left(W_{\text{down}}^{(l)} \mathbf{h}^{(l)}\right). \tag{34}$$

The dimensionality of z_u is matched to the Adapter bottleneck dimension, such that each latent component modulates the corresponding dimension of the intermediate representation. The preference encoder for obtaining z_u, the training objective, and the optimization procedure remain identical to those used in the LoRA instantiation.

E.2 PROMPT TUNING INSTANTIATION

For standard Prompt Tuning, a set of learnable continuous prompt embeddings is introduced at the input embedding level:

$$P = [p_1; p_2; \ldots; p_m] \in \mathbb{R}^{m \times d}, \tag{35}$$

where m denotes the number of soft prompt tokens and d is the hidden dimension of the pretrained model. Given an input sequence x with token embeddings $E(x)$, the model input is $[P; E(x)]$. During adaptation, only P is optimized while the model parameters remain frozen.

The soft prompt provides a natural component space for incorporating the latent representation in *LatentPersonal*. We therefore introduce the user latent z_u along the prompt-token dimension:

$$P_u = \operatorname{diag}(z_u) P, \tag{36}$$

or equivalently,

$$P_u = [z_{u,1} p_1; z_{u,2} p_2; \ldots; z_{u,m} p_m]. \tag{37}$$

The dimensionality of z_u is matched to the number of soft prompt tokens, such that each latent component modulates the corresponding prompt embedding. The resulting P_u replaces P in the input representation, while the preference encoder for obtaining z_u, the training objective, and the optimization procedure remain identical to those used in the LoRA instantiation.

F SCOPE AND LIMITATIONS

Our results support latent navigation as an effective formulation for few-shot personalization, but they do not distinguish whether the underlying personalization-relevant structure is inherited from pretraining or emerges during population-level personalization training. In addition, our evaluation focuses on benchmark-based personalization settings with limited observed interactions; understanding how the learned latent representations behave under longer-term and evolving user preferences remains an important direction for future work.

Table 17: Evaluation with alternative PEFT architectures on HiCUPID with Llama-3.1-8B-Instruct across different numbers of observed user interactions.

Method	**Shot**	**R-1**	**R-L**	**MET.**	**CP**	**PP**
Adapter	1	0.3627	0.3063	0.3591	6.04	6.39
	2	0.3848	0.3272	0.3749	6.34	6.48
	4	0.3917	0.3373	0.3876	6.35	6.48
	8	0.3944	0.3346	0.3748	6.23	6.59
Adapter + *LatentPersonal*	1	0.3980	0.3468	0.4139	6.19	6.41
	2	0.4074	0.3536	0.4191	6.23	6.43
	4	0.4018	0.3521	0.4127	6.29	6.39
	8	0.3978	0.3483	0.4112	6.29	6.50
Prompt Tuning	1	0.3215	0.2940	0.2978	5.54	5.47
	2	0.3534	0.3323	0.3231	5.76	5.59
	4	0.3362	0.3114	0.3064	6.00	5.66
	8	0.3757	0.3455	0.3364	6.05	5.64
Prompt Tuning + *LatentPersonal*	1	0.3712	0.3469	0.3399	5.76	5.76
	2	0.3691	0.3476	0.3359	5.85	5.66
	4	0.3686	0.3417	0.3327	5.86	5.58
	8	0.3729	0.3524	0.3393	5.94	5.42