Alfa: Attentive Low-Rank Filter Adaptation for Structure-Aware Cross-Domain Personalized Gaze Estimation

He-Yen Hsieh, Wei-Te Mark Ting, H.T. Kung

Harvard University

Abstract

Pre-trained gaze models learn to identify useful patterns commonly found across users, but subtle user-specific variations (*i.e.*, eyelid shape or facial structure) can degrade model performance. Test-time personalization (TTP) adapts pre-trained models to these user-specific domain shifts using only a few unlabeled samples. Efficient fine-tuning is critical in performing this domain adaptation: data and computation resources can be limited—especially for on-device customization. While popular parameter-efficient fine-tuning (PEFT) methods address adaptation costs by updating only a small set of weights, they may not be taking full advantage of structures encoded in pre-trained filters. To more effectively leverage existing structures learned during pre-training, we reframe personalization as a process to reweight existing features rather than learning entirely new ones.

We present Attentive Low-Rank Filter Adaptation (Alfa) to adapt gaze models by reweighting semantic patterns in pre-trained filters. With Alfa, singular value decomposition (SVD) extracts dominant spatial components that capture eye and facial characteristics across users. Via an attention mechanism, we need only a few unlabeled samples to adjust and reweight pre-trained structures, selectively amplifying those relevant to a target user. Alfa achieves the lowest average gaze errors across four cross-dataset gaze benchmarks, outperforming existing TTP methods and low-rank adaptation (LoRA)-based variants. We also show that Alfa's attentive low-rank methods can be applied to applications beyond vision, such as diffusion-based language models.

1 Introduction

Gaze estimation can infer the direction a person is looking from facial or eye-region images. This capability is central to many applications in augmented reality, human-computer interaction, and assistive technologies. For example, in gaze-assisted communication systems (Lee et al. 2024; Khan et al. 2022), detecting the user's eye focus is essential for delivering intuitive and effective responses.

Gaze estimation models perform well under controlled conditions, but may struggle in real-world settings. Differences across users, camera configurations, and environments, such as changes in lighting or head pose, create discrepancies between training and deployment conditions. These domain shifts decrease accuracy when models lack robustness outside their initial training conditions.

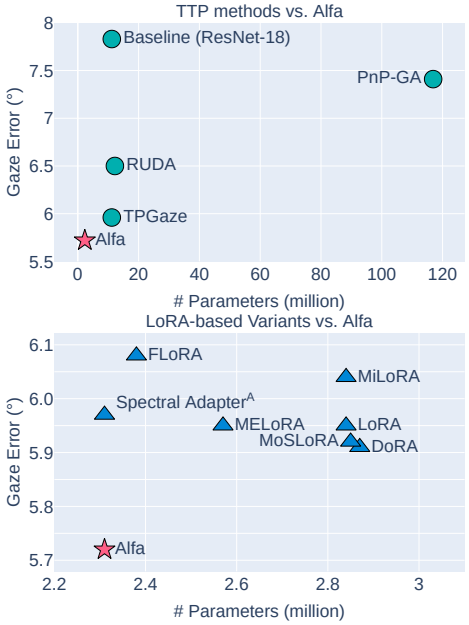

Figure 1: Alfa achieves the lowest average gaze error, with the smallest model size, across four cross-dataset benchmarks: from ETH-XGaze to MPIIGaze, ETH-XGaze to EyeDiap, Gaze360 to MPIIGaze, and Gaze360 to EyeDiap. Top: Comparison with other test-time personalization (TTP) methods. Baseline refers to a ResNet-18 without fine-tuning. Bottom: Comparison with low-rank adaptation (LoRA)-based variants.

Test-time personalization (TTP) (Liu et al. 2024a; Bao et al. 2022; Liu et al. 2021) is a variant of unsupervised domain adaptation (UDA) (Wang et al. 2022; Bao et al. 2022; Liu et al. 2021) in which the model adapts to a new user during deployment, relying only on unlabeled samples collected at test time. TTP provides privacy-preserving, on-device adaptation in a few-shot setting and offers a practical solution for real-world deployment without access to

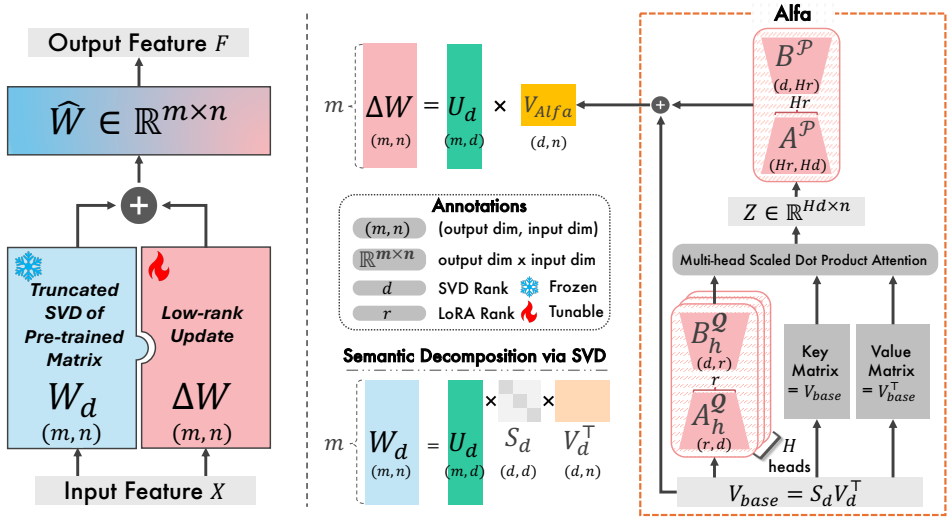

Figure 2: Overview of Attentive Low-Rank Filter Adaptation (Alfa). (a) The pre-trained weight matrix is approximated using truncated SVD: $W_d = U_d S_d V_d^\top$. Then, a tunable low-rank update ΔW is added for adaptation. (b) Alfa adapts gaze models by reweighting spatial structures encoded in pre-trained filters. Alfa extracts dominant spatial patterns ($V_{\text{base}} = S_d V_d^\top$) via singular value decomposition (SVD). For personalization, multi-head low-rank modules $A^{\mathcal{Q}}$ and $B^{\mathcal{Q}}$ generate query weights, and V_{base} and $V_{\text{base}}^\top$ are reused as key and value matrices. Using multi-head scaled dot-product attention, Alfa identifies the spatial structures most relevant to a target user. Alfa aggregates this into a personalized update using additional low-rank modules $A^{\mathcal{P}}$ and $B^{\mathcal{P}}$, forming V_{Alfa}, which encodes gaze-specific adaptations informed by the pre-trained spatial structure.

the original training data or ground-truth gaze labels. TTP also addresses some key challenges in gaze estimation: differences in users' eye shapes, appearances, or camera placement may degrade model performance.

Intuitively, human faces may exhibit spatially coherent patterns (*i.e.*, eye and facial geometry) (Tian et al. 2023; Hong 2022). With sufficient pre-training, models learn to capture useful general features that correspond to these patterns, encoding them into spatial structures (*e.g.* filters) within the weights. Yet, small changes to these patterns can significantly reduce prediction accuracy. While fine-tuning can help mitigate this, leveraging more information from the pre-trained model may improve performance (Wang et al. 2025; Meng, Wang, and Zhang 2024). We find that leveraging existing features rather than learning entirely *new* features is effective for data-scarce adaptation in gaze estimation (*e.g.*, with only five new images).

In this paper, we propose Attentive Low-Rank Filter Adaptation (Alfa), which adapts gaze models by reweighting the *spatial structure* encoded in pre-trained filters. As depicted in Figure 2, Alfa first performs SVD on pre-trained weights to obtain dominant spatial structures that correspond to common patterns (*e.g.*, geometric features in the eye and face). Instead of learning new filters or initializing new weights, Alfa adapts models by modulating the influence of these dominant patterns by using a few unlabeled samples from a target user. This enables personalized predictions from minimal additional data while maintaining alignment with the original learned representation. Visualization results demonstrate that Alfa attends to common and intuitive localized regions, such as the eyelids, which vary across users and should be effectively detected during adaptation (See Figure 5).

In summary, our contributions are as follows:

1. Attentive Low-Rank Filter Adaptation (Alfa) adapts gaze models by attending over structured spatial patterns derived from SVD, instead of treating weights as unstructured tensors.
2. Within Alfa, a multi-head low-rank adaptation module accepts scalable personalization capacity during fine-tuning. Storing the pre-trained weight in truncated SVD form reduces model size, and makes Alfa's updates fully mergeable without increasing model size at deployment (See Section 3.4).
3. Empirical demonstration of Alfa outperforming prior methods on four cross-domain gaze benchmarks using only a few unlabeled test-time samples.
4. Extension of Alfa's structured adaptation to diffusion-based large language models (LLMs), showing improved zero-shot reasoning across multiple benchmarks.

2 Related Work

2.1 Test-Time Personalization (TTP)

To improve generalization across domains, many gaze estimation methods adopt unsupervised domain adaptation (UDA). In the standard UDA setting (Liu et al. 2021; Guo et al. 2020; Kellnhofer et al. 2019), models are trained with labeled source-domain data and unlabeled target-domain data. PnP-GA (Liu et al. 2021) uses an outlier-guided loss

to improve adaptation with a few source samples and unlabeled target data. To relax the need for source data, source-free UDA methods (Bao et al. 2022; Wang et al. 2022) adapt models using only unlabeled target data, though they still require a relatively large amount of it. RUDA (Bao et al. 2022), for example, leverages rotation-based gaze synthesis and geometric constraints to improve pseudo-label quality. Personalization provides a complementary way by adapting models to individual users. While few-shot methods (Ghosh et al. 2022; Chen and Shi 2020; Park et al. 2019; Yu, Liu, and Odobez 2019) fine-tune models with a small number of labeled personal samples, obtaining gaze labels is often expensive or impractical. TTP (Liu et al. 2024a) addresses this challenge by adapting the model to each user during inference using a few unlabeled personal samples. TTP is well-suited for on-device scenarios, where user appearance varies and updating the full model is often too resource-intensive. These constraints motivate efficient, lightweight adaptation methods that support per-user personalization without requiring labeled data or access to the source domain. Alfa supports test-time personalization with no additional inference cost, is scalable through multi-head adaptation during fine-tuning, and uses SVD to select critical semantic components, resulting in a compact and efficient model.

2.2 Low-Rank Adaptation

Low-rank adaptation injects small trainable matrices into pre-trained models to fine-tune them efficiently without altering the original weights. LoRA (Hu et al. 2022) is a widely adopted method that adds rank-constrained updates to existing layers and has shown strong performance across various tasks. Building on this idea, MiLoRA (Wang et al. 2025) initializes adaptation weights using principal components from SVD, while PiSSA (Meng, Wang, and Zhang 2024) directly updates the top spectral components. Spectral Adapter^A (Zhang and Pilanci 2024) introduces learnable weights applied to the spectral bases, and MoSLoRA (Wu et al. 2024) places a mixer matrix between LoRA modules to improve representation capacity. DoRA (Liu et al. 2024b) decomposes pre-trained weights into independent direction and magnitude components. MELoRA (Ren et al. 2024) ensembles multiple mini-expert LoRA branches, and FLoRA (Si et al. 2025) applies Tucker decomposition to capture multi-dimensional parameter changes through a shared low-rank core. Despite their varied adaptation strategies, they often treat model weights as unstructured tensors and overlook the spatial structure encoded in pre-trained filters. In contrast, Alfa attends to these spatial structures and reweights them during personalization, enabling structure-aware adaptation guided by meaningful semantics.

3 Alfa

3.1 Problem Definition and Preliminary

We consider the TTP task for gaze estimation, formulated as a variant of UDA. The goal is to adapt a gaze model trained on a general source domain $\mathcal{D}_S$ to a new, unseen user in a target domain $\mathcal{D}_T$, with only a few unlabeled samples. Let each RGB input image be denoted as I, and its ground-truth gaze direction represented by a 2D vector $g \in \mathbb{R}^2$ (yaw and pitch). The source domain provides labeled data $\mathcal{D}_S = \{(I_i^S, g_i^S)\}_{i=1}^{N_S}$, while the target domain provides a small unlabeled set $\mathcal{D}_T = \{I_k^T\}_{k=1}^{N_T}$, typically with $N_T = 5$. In a gaze model, we denote the input and output features of a convolutional or linear layer as $X \in \mathbb{R}^{n\times h\times w}$ and $F \in \mathbb{R}^{m\times h'\times w'}$, respectively, where n and m are the input and output channel dimensions, and h, w and h', w' are the height and width of the input and output spatial resolutions. These features are connected through a pre-trained weight matrix $W \in \mathbb{R}^{m\times n}$.

To enable data- and parameter-efficient adaptation from a few unlabeled samples, we build on LoRA, which injects a trainable low-rank update into a pre-trained weight matrix: $\Delta W = AB$, where $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times n}$, and $r \ll \min(m, n)$, with r as the LoRA rank. Typical initializations set $A \sim \mathcal{N}(0, \sigma^2)$ and $B = 0$. However, this formulation overlooks the spatial and geometric structure embedded in the pre-trained weights: structure that captures critical visual patterns relevant to gaze across users. Alfa addresses this limitation by personalizing gaze models through structured reweighting of semantic filters extracted from pre-trained weights. Section 3.2 introduces SVD-based decomposition of pre-trained weights to extract a semantic basis dictionary. Section 3.3 describes the structured reweighting mechanism that selectively adapts semantic components for personalization.

3.2 Structured Decomposition of Gaze Filters

Gaze estimation relies on structured visual patterns shaped by the anatomy of the eyes and surrounding facial regions. While these patterns vary across individuals, the variations tend to follow a few set of consistent spatial changes, such as shifts in iris position or subtle deformations in surrounding facial muscles. These consistencies suggest that pre-trained weights encode recurring spatial structures that are broadly shared across users (see Section 4.6).

When trained on a diverse population, the model learns to encode this structure in its weights, forming a strong foundation for general gaze prediction. To make this structure explicit, we apply SVD to the pre-trained weight matrix. Let $W \in \mathbb{R}^{m\times n}$ denote the weight matrix of a convolutional or linear layer. We compute a truncated SVD:

$$W \approx W_d = U_d S_d V_d^\top \tag{1}$$

where $d \ll \min(m, n)$ is the target rank, yielding:

- $U_d \in \mathbb{R}^{m\times d}$: output projection matrix (left singular vectors),
- $S_d \in \mathbb{R}^{d\times d}$: singular values representing the importance of each direction,
- $V_d^\top \in \mathbb{R}^{d\times n}$: dominant spatial directions in input space.

Thus, we obtain the **semantic basis dictionary**, defined as:

$$V_{\text{base}} = S_d V_d^\top \in \mathbb{R}^{d\times n}$$

V_{base} reflects the highest-energy components learned during gaze pre-training. These components capture key spatial patterns relevant to gaze behavior (*e.g.*, iris and peri-ocular cues

co-activating). Components reused most frequently during pre-training to reduce gaze loss have the highest energy. Since SVD ranks weight-space patterns by energy, truncating to these leading components preserves the dominant gaze-relevant structure, yielding a compact, structure-aware basis for personalization. Retaining only the top d components ensures we focus on the most expressive, information-rich parts of the filter space while providing a compact basis for downstream adaptation. In Alfa, this semantic structure facilitates learning efficient, personalized updates in later stages.

3.3 Personalizing the Semantic Basis Dictionary

The semantic basis dictionary extracted via SVD provides a compact set of spatial patterns that generalize well across individuals in gaze estimation. However, these patterns may not fully capture the unique appearance characteristics of each user. Alfa introduces a low-rank update that reweights the components of the semantic basis dictionary without discarding the shared structure learned during pre-training. This approach allows the model to adapt to individual differences while preserving gaze-relevant information encoded in the pre-trained filters. To personalize the model, we add a low-rank update ΔW on top of the pre-trained weight W_d. The adapted weight is defined as:

$$\hat{W} = W_d + \Delta W \tag{2}$$

where $\Delta W \in \mathbb{R}^{m\times n}$ is a low-rank personalization term computed as $\Delta W = U_d V_{\text{alfa}}$, where, V_{alfa} is a learnable update produced by the Alfa module.

Attending to Semantic Basis Dictionary Alfa computes a personalized adaptation by applying a multi-head attention mechanism over the semantic basis dictionary $V_{\text{base}} \in \mathbb{R}^{d\times n}$. The dictionary captures shared spatial patterns from pre-training, and Alfa uses attention to reweight the slices most relevant to the target subject. Let H be the number of attention heads. Different heads attend to different rank slices, combining complementary cues and reducing drift when a few personal samples are available (see supplementary material Section F). For each attention head indexed by $h \in \{1,\dots,H\}$, we define a pair of low-rank projection matrices: $A_h^{\mathcal{Q}} \in \mathbb{R}^{r\times d}$ for the query projection, and $B_h^{\mathcal{Q}} \in \mathbb{R}^{d\times r}$ for the query back-projection. The query projections are initialized as

$$A_h^{\mathcal{Q}} \sim \mathcal{N}(0,\sigma^2), \quad B_h^{\mathcal{Q}} = 0, \tag{3}$$

where σ denotes the standard deviation of the initialization distribution. Each head computes a query matrix as

$$\mathcal{Q}_h = B_h^{\mathcal{Q}} A_h^{\mathcal{Q}} V_{\text{base}} \in \mathbb{R}^{d\times n} \tag{4}$$

Key and value matrices are directly derived from V_{base} and shared across all attention heads. The key matrix is defined as $\mathcal{K} = V_{\text{base}} \in \mathbb{R}^{d\times n}$. The value matrix is its transpose, $\mathcal{V} = V_{\text{base}}^\top \in \mathbb{R}^{n\times d}$. For each head, scaled dot-product attention is computed using its query $\mathcal{Q}_h$ as:

$$\text{Attn}_h = \text{softmax}\left(\frac{\mathcal{Q}_h \mathcal{K}^\top}{\sqrt{n}}\right) \in \mathbb{R}^{d\times d} \tag{5}$$

$$Z_h = \mathcal{V}\,\text{Attn}_h^\top \in \mathbb{R}^{n\times d} \tag{6}$$

Each output Z_h is transposed and then stacked across heads:

$$Z = [Z_1^\top, \dots, Z_H^\top] \in \mathbb{R}^{Hd\times n} \tag{7}$$

Integrating Multi-Head Adaptation After aggregating the multihead outputs into $Z \in \mathbb{R}^{Hd\times n}$, we project them back into the semantic space using two low-rank matrices, $A^{\mathcal{P}} \in \mathbb{R}^{rH\times Hd}$ and $B^{\mathcal{P}} \in \mathbb{R}^{d\times rH}$. These are initialized as:

$$A^{\mathcal{P}} \sim \mathcal{N}(0,\sigma^2), \quad B^{\mathcal{P}} = 0 \tag{8}$$

We compute the personalized update as:

$$V_{\text{Alfa}} = B^{\mathcal{P}} A^{\mathcal{P}} Z + V_{\text{base}} \in \mathbb{R}^{d\times n} \tag{9}$$

This completes the adaptation process, resulting in a low-rank update $\Delta W = U_d V_{\text{Alfa}}$.

3.4 Fine-tuning and Inference with Alfa

We describe the fine-tuning procedure for Alfa and discuss how it maintains computational efficiency during inference, despite the inclusion of multi-head LoRA modules.

Fine-Tuning Human faces typically exhibit left-right symmetry, and gaze behavior should remain consistent under horizontal flips. To exploit this property, we apply a symmetry loss following Kellnhofer et al. (2019) using the five unlabeled personal samples $\{I_k^T\}_{k=1}^{N_T}$, where $N_T = 5$. Let $f_\theta(\cdot)$ denote the gaze model. For each image, we generate a horizontally flipped version $I_k^{T,\text{flip}}$ and obtain predictions:

$$\hat{g}_k^T = f_\theta(I_k^T), \quad \hat{g}_k^{T,\text{flip}} = f_\theta(I_k^{T,\text{flip}}), \tag{10}$$

where $\hat{g} = [\hat{g}_{\text{yaw}}, \hat{g}_{\text{pitch}}] \in \mathbb{R}^2$. The symmetry loss is computed as:

$$\mathcal{L}_{\text{fine-tune}} = \frac{1}{N_T}\sum_{k=1}^{N_T}\left|\hat{g}_k^T - \text{FlipYaw}(\hat{g}_k^{T,\text{flip}})\right|_1 \tag{11}$$

Inference Adaptation reuses the left basis U_d from the pre-trained weight W_d, preserving the original model structure. This enables efficient update merging at inference time. Specifically, the full adapted weight becomes:

$$\begin{aligned}\hat{W} = W_d + \Delta W &= U_d V_{\text{base}} + U_d V_{\text{Alfa}} \\ &= U_d\,(V_{\text{base}} + V_{\text{Alfa}})\end{aligned} \tag{12}$$

We denote the sum in parentheses as a factor V_{adapt}, yielding:

$$\hat{W} = U_d V_{\text{adapt}}, \quad \text{where} \quad V_{\text{adapt}} = V_{\text{base}} + V_{\text{Alfa}} \tag{13}$$

Adapted weights stay in low-rank form, keeping the same structure as the original compressed model. We can simply update the right-hand side to get $\hat{W} = U_d V_{\text{adapt}}$ without first reconstructing the full weight matrix. In contrast, standard LoRA variants add the low-rank term AB to the full matrix:

$$\hat{W} = W_d + AB = U_d S_d V_d^\top + AB. \tag{14}$$

Merging AB necessitates expanding W_d to its full size in $\mathbb{R}^{m\times n}$ and negates the benefit of parameter compression (*e.g.*, model increases from around 2M to 11M parameters). Storing W_d in its truncated SVD form $U_d S_d V_d^\top$ would preclude directly merging the low-rank term AB. Alfa updates only the SVD right factor $S_d V_d^\top$ and keeps U_d fixed, ensuring the adaptation remains low-rank and directly mergeable.

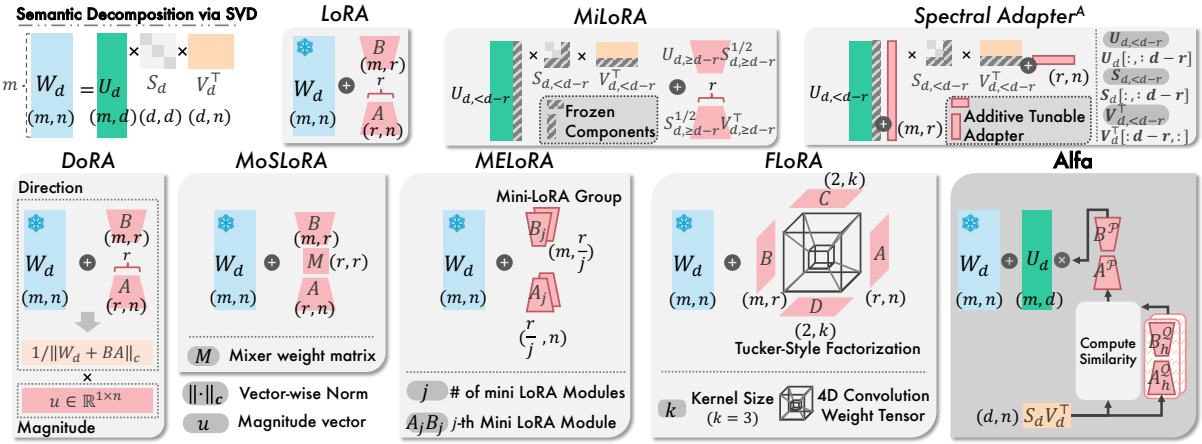

Figure 3: Comparison with other LoRA-based variants, including LoRA, MiLoRA, Spectral Adapter^A, DoRA, MoSLoRA, MELoRA, and FLoRA. Alfa selectively reuses semantic patterns encoded in pre-trained weights and activates the most relevant ones during adaptation. Only blocks with red backgrounds are tunable. Best viewed in color.

Method	# Parameters (M)			Source → Target Domain (5-shot)				Avg
	Train	Tuned	Test	$D_E \to D_M$	$D_E \to D_D$	$D_G \to D_M$	$D_G \to D_D$	
Baseline (ResNet-18)	11.18	0	11.18	8.02	7.30	7.79	8.19	7.83
PnP-GA† (Liu et al. 2021)	116.9	116.9	116.9	6.91	7.18	7.36	8.17	7.41
RUDA† (Bao et al. 2022)	12.20	12.20	12.20	6.86	6.84	6.96	5.32	6.50
TPGaze (Liu et al. 2024a)	11.18	**0.13**	11.18	6.30	5.89	6.62	**5.04**	5.96
Alfa	5.26	2.98	**2.31**	**5.30**	**5.82**	**5.91**	5.86	**5.72**

Table 1: Comparison with state-of-the-art TTP methods across four cross-domain benchmarks. Datasets are denoted as follows: D_E = ETH-XGaze, D_M = MPIIGaze, D_G = Gaze360, and D_D = EyeDiap. Baseline is a ResNet-18 model without any fine-tuning. Results are reported in angular gaze error (°). **Bold** indicates the best result. The symbol † indicates results obtained from the re-implementation of TPGaze (Liu et al. 2024a).

4 Experiments

4.1 Datasets

For gaze estimation, we use a cross-domain setup (Liu et al. 2024a; Bao et al. 2022; Liu et al. 2021) with models trained on ETH-XGaze (D_E) or Gaze360 (D_G) as source domains, and evaluated on MPIIGaze (D_M) and EyeDiap (D_D) as target domains. All preprocessing follows TPGaze (Liu et al. 2024a). For LLM experiments, we fine-tune on the s1K reasoning dataset (Muennighoff et al. 2025) and evaluate zero-shot performance on GSM8K (Cobbe et al. 2021), MATH500 (Lightman et al. 2024), Countdown (Pan et al. 2025), and Sudoku (Arel 2025). Full dataset details are in the supplementary material.

4.2 Experimental Setup

We use one NVIDIA RTX 4090 for gaze experiments, and two NVIDIA A40 GPUs for LLM experiments. For TTP pre-training and personalization, learning rate is set to 10^{-4}. For LLM adaptation, we follow the fine-tuning from d1 (Zhao et al. 2025) with LLaDA-8B-Instruct (Nie et al. 2025) as the base model. Below, we describe the details for TTP:

Pre-training We use Adam with a batch size of 120 for 50 epochs on source domains D_E and D_G. After applying SVD, we fine-tune all parameters on the same data for 25 additional epochs to mitigate information loss, resulting in the truncated pre-trained weights W_d.

Personalization We use the first 5 images per subject as in TPGaze (Liu et al. 2024a). Each image is repeated once and augmented with ColorJitter, GaussianBlur, and RandomAffine (see supplementary material for details). We use AdamW with a $10\times$ learning rate for layer3 and layer4 of ResNet-18, and apply the same fine-tuning scheme across all LoRA variants for fair comparison.

Evaluation Metric We report angular gaze error (in degrees), measuring the angle between predicted and ground-truth gaze directions. We convert model outputs, 2D yaw and pitch, into 3D vectors to compute the error, following prior works (Liu et al. 2021; Bao et al. 2022; Liu et al. 2024a).

4.3 Comparison with SOTA Methods

Table 1 compares Alfa with state-of-the-art TTP methods across four cross-domain gaze estimation benchmarks. We evaluate model adaptation from ETH-XGaze or Gaze360 to

Method	Rank		# Parameters (M)			Source → Target Domain (5-shot)				Avg
	SVD	LoRA	Train	Tuned	Test	$D_E \rightarrow D_M$	$D_E \rightarrow D_D$	$D_G \rightarrow D_M$	$D_G \rightarrow D_D$	
Baseline (No Adaptation)	64	-	2.31	0	2.31	6.60	8.84	6.86	6.83	7.29
LoRA (Hu et al. 2022)	64	8	2.84	0.53	2.84	5.66	6.17	6.23	5.72	5.95
MiLoRA (Wang et al. 2025)	64	8	2.84	0.29	2.84	5.67	6.25	6.23	6.00	6.04
DoRA (Liu et al. 2024b)	64	8	2.87	0.56	2.87	5.51	5.83	6.30	5.98	5.91
MoSLoRA (Wu et al. 2024)	64	8	2.85	0.54	2.85	5.55	6.13	6.31	**5.70**	5.92
MELoRA (Ren et al. 2024)	64	8	2.57	0.27	2.57	5.56	6.12	6.29	5.84	5.95
Spectral Adapter^A (Zhang and Pilanci 2024)	64	8	2.59	0.29	**2.31**	5.50	6.15	6.23	6.00	5.97
FLoRA (Si et al. 2025)	64	8	2.38	**0.07**	2.38	5.85	6.36	6.40	5.71	6.08
Alfa	64	8	5.26	2.98	**2.31**	**5.30**	**5.82**	**5.91**	5.86	**5.72**

Table 2: Comparison with other LoRA-based variant methods. The baseline is *not* fine-tuned on the target domain (*i.e.* no adaptation). All methods use the same truncated pre-trained weight matrix $W_d = U_d S_d V_d^\top$. While some variants introduce additional components that cannot be merged into W_d, Alfa supports full mergeability, enabling efficient inference without extra computational overhead. **Bold** indicates the best result.

Backbone	Method	LoRA Rank	Tuned Params (Usage %)	GSM8K (0-shot)		MATH500 (0-shot)		Countdown (0-shot)		Sudoku (0-shot)	
				128	256	128	256	128	256	128	256
LLaDA-8B-Instruct	LoRA	128	100.7M (1.24 %)	66.5	**78.8**	*26.2*	32.6	20.3	14.5	*16.5*	**8.5**
	DoRA†	128	101.1M (1.25%)	*68.1*	76.8	*26.2*	*33.4*	*21.5*	*16.0*	**17.2**	8.0
	LoRA†	64	**50.3M (0.62%)**	67.9	*77.9*	25.0	33.0	16.4	12.5	14.5	7.9
	Alfa	64	69.2M (0.85%)	**68.4**	77.1	**26.6**	**33.8**	**27.3**	**17.2**	9.7	*8.3*

Table 3: Comparison of Alfa, LoRA, and DoRA on LLaDA-8B-Instruct across four zero-shot reasoning tasks. Alfa achieves competitive or better performance while using only 0.85% of tunable parameters. **Bold** indicates the best result, and *italics* indicate the second best. The symbol † indicates re-implementation.

MPIIGaze and EyeDiap. The baseline is a ResNet-18 (He et al. 2016) model without fine-tuning. For fair comparison, we also adopt ResNet-18 as the backbone. Alfa achieves the lowest average angular gaze error across all benchmarks while remaining around $5\times$ smaller than other methods.

4.4 Comparison with LoRA-based Methods

We compare Alfa with other LoRA-based methods (see Section 2.2) in Table 2, and illustrate their architectural differences in Figure 3. All methods operate on the same truncated pre-trained weight $W_d = U_d S_d V_d^\top$. However, not all LoRA-based variants support merging updates into this decomposed form of W_d during inference. In contrast, Alfa exploits the spatial structure of pre-trained weights and reweights semantically meaningful patterns. Alfa's personalized updates are fully compatible with the truncated SVD form (see Section 3.4), and incur no additional inference-time cost over other methods. This gaze-specific structural prior guides adaptation toward meaningful filter reweighting without disrupting pre-trained semantics, leading to the lowest average gaze error across four cross-domain benchmarks.

4.5 Ablation Studies

We conduct ablation experiments to evaluate the effect of attention head count and the LoRA rank on Alfa's performance. For all settings, we fix SVD rank to 64. As shown in Table 4, increasing the number of attention heads generally improves personalization performance, dropping gaze error from 6.20 (1 head) to 5.72 (16 heads). Since Alfa reuses the same left basis U_d from the pre-trained W_d, all adaptations are merged into the base weight at inference time, incurring no additional computational cost regardless of the number of heads (see Section 3.4). Additional ablation studies are included in the supplementary materials.

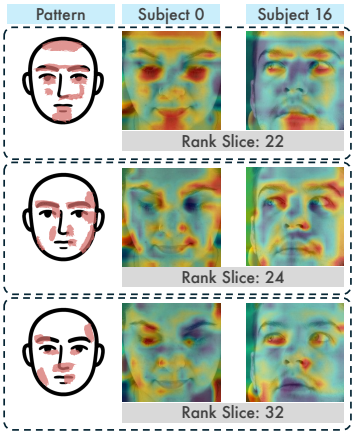

Figure 4: Spatial patterns captured during pre-training. Visualizations use rank slices from SVD-decomposed weights of ResNet-18 (pre-trained on ETH-XGaze) from conv1 and the first block of layer3. Left column: visualization of encoded pattern. Middle and right columns: activations for Subject 0 and 16 from ETH-XGaze using $U_d[:, s] S_d[s] V_d^\top[s]$ for slice s. Red regions indicate higher activations.

Heads	# Parameters (M)			Source → Target Domain (5-shot)				Avg
	Train	Tuned	Test	$D_E \rightarrow D_M$	$D_E \rightarrow D_D$	$D_G \rightarrow D_M$	$D_G \rightarrow D_D$	
1	2.35	0.04	2.31	5.74	6.84	6.36	5.87	6.20
2	2.40	0.10	2.31	5.91	6.65	6.32	**5.75**	6.16
4	2.58	0.27	2.31	5.47	6.12	6.32	5.82	5.93
8	3.16	0.86	2.31	5.67	6.19	6.24	5.76	5.97
16	5.26	2.98	2.31	5.30	**5.82**	**5.91**	5.86	**5.72**
32	13.20	10.90	2.31	**5.20**	6.17	6.10	6.17	5.91

Table 4: Ablation study on attention head count in Alfa. Personalization performance generally improves with head count, with the lowest average gaze error at 16 heads. Heads share the left basis U_d, so adapted weights can be merged into the base model without extra computation at inference.

4.6 Visualization of Pre-trained Spatial Patterns

Figure 4 illustrates some spatial patterns encoded during pre-training by visualizing individual rank slices from the SVD-decomposed weights of a ResNet-18 model trained on ETH-XGaze. The s-th SVD slice is computed as:

$$U_d[:, s]S_d[s]V_d^\top[s] \tag{15}$$

where s indexes the rank component of the decomposition. The left column sketches the spatial structure encoded by that slice, while the middle and right columns display activations for two different subjects (Subject 0 and Subject 16). For example, the 22nd slice emphasizes the eyebrows, lower eyelids, and lower mouth, while the 24th slice activates around the nose sides, regions beside the eyes, and facial muscles near the mouth. Consistent patterns across individuals demonstrate that pre-training captures reusable spatial structures aligned with facial geometry relevant to gaze.

4.7 Visualization of Adaptation Behavior

We visualize low-rank updates $\Delta\mathcal{W}$ for three users from the MPIIGaze dataset (subjects p02, p04, and p13) in Figure 5. Alfa's updates consistently focus on gaze-relevant facial regions, such as the eyes and surrounding muscles. The patterns vary slightly across users, reflecting personalized adjustments while still maintaining alignment with the model's original spatial semantics. In contrast, LoRA yields dispersed and inconsistent updates even when using the same backbone layers with personalization. However, we note that LoRA does not explicitly *avoid* key regions: it simply lacks targeted semantic guidance and can rely on gaze-relevant signals (*e.g.*, pose cues) from other regions. This visualization highlights Alfa's ability to identify useful components from the semantic basis dictionary. Since the components reflect domain-specific discrepancies between the pre-trained source model and the target user, we need only reweight them for effective user-specific adaptation.

4.8 Applying Alfa to Diffusion-Based LLMs

Table 3 compares Alfa to LoRA and DoRA when applied to the diffusion-based LLaDA-8B-Instruct model across four zero-shot reasoning benchmarks: GSM8K, MATH500, Countdown, and Sudoku. We adapt with 1,000 samples.

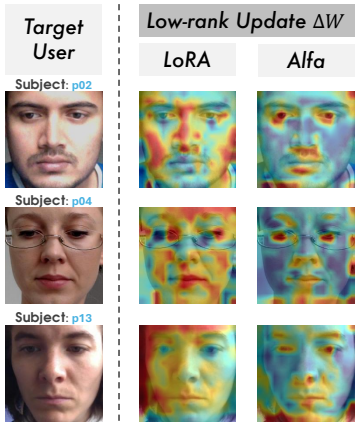

Figure 5: Visualization of low-rank updates ΔW on the MPIIGaze test set for LoRA and Alfa using filters from conv1 and the first block of layer3 in the ResNet-18 model (pre-trained on ETH-XGaze). Red regions indicate higher activation values. When using LoRA updates, model is highly inconsistent with respect to the significant regions of focus across users. In contrast, Alfa captures localized regions consistently across users. This shows Alfa identifies useful components that translate well between source and target domains from the semantic base dictionary. Reweighting these components allows for effective adaptation.

We include LLM experiments as reasoning tasks reuse token-interaction patterns (*e.g.*, formats, step markers), and reweighting these patterns is shown to be beneficial when data is limited. LoRA results are from d1 (Zhao et al. 2025). We reimplement DoRA using HuggingFace PEFT library[1]. While LoRA and DoRA use a LoRA rank of 128, Alfa uses a lower rank of 64 and tunes only 0.85% of the model's parameters. For all experiments in Table 3, we retain the full pre-trained weight matrix W without SVD truncation. Alfa uses a SVD rank of 128 and 8 heads for computing the low-rank update ΔW. Despite this smaller footprint, Alfa achieves comparable or superior accuracy across benchmarks. This suggests that reasoning patterns in language models may also be representable by generalizable components encoded during pre-training and of interest for future work.

5 Conclusion

We present Alfa, a structure-aware method for test-time personalization of gaze estimation models. By attending over spatial patterns extracted via SVD, Alfa reuses meaningful components from pre-trained filters, enabling efficient domain adaptation through a multi-head low-rank design. This approach allows scalable personalization during fine-tuning and maintains a compact model without increasing inference cost. Experiments on four cross-domain gaze benchmarks demonstrate state-of-the-art performance with only a few unlabeled samples. Furthermore, Alfa's structured adaptation shows promise for other applications, such as zero-shot reasoning tasks with diffusion-based language models.

[1]https://github.com/huggingface/peft

References

Arel. 2025. Arel's sudoku generator. https://www.ocf.berkeley.edu/~arel/sudoku/main.html. Accessed: 2025-04-08.

Bao, Y.; Liu, Y.; Wang, H.; and Lu, F. 2022. Generalizing Gaze Estimation with Rotation Consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 4197–4206.

Chen, Z.; and Shi, B. E. 2020. Offset Calibration for Appearance-Based Gaze Estimation via Gaze Decomposition. In *IEEE Winter Conference on Applications of Computer Vision, WACV*, 259–268.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Ghosh, S.; Hayat, M.; Dhall, A.; and Knibbe, J. 2022. MTGLS: Multi-Task Gaze Estimation with Limited Supervision. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV*, 1161–1172.

Guo, Z.; Yuan, Z.; Zhang, C.; Chi, W.; Ling, Y.; and Zhang, S. 2020. Domain Adaptation Gaze Estimation by Embedding with Prediction Consistency. In *Proceedings of the Asian Conference on Computer, ACCV*, volume 12626, 292–307.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 770–778.

Hong, Y.-J. 2022. Facial identity verification robust to pose variations and low image resolution: Image comparison based on anatomical facial landmarks. *Electronics*, 11(7): 1067.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR*.

Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 6911–6920.

Khan, A. A.; Newn, J.; Bailey, J.; and Velloso, E. 2022. Integrating Gaze and Speech for Enabling Implicit Interactions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI*, 349:1–349:14.

Lee, J.; Wang, J.; Brown, E.; Chu, L.; Rodriguez, S. S.; and Froehlich, J. E. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI*, 408:1–408:20.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2024. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations, ICLR*.

Liu, H.; Qi, J.; Li, Z.; Hassanpour, M.; Wang, Y.; Plataniotis, K. N.; and Yu, Y. 2024a. Test-Time Personalization with Meta Prompt for Gaze Estimation. In *AAAI*, 3621–3629.

Liu, S.; Wang, C.; Yin, H.; Molchanov, P.; Wang, Y. F.; Cheng, K.; and Chen, M. 2024b. DoRA: Weight-Decomposed Low-Rank Adaptation. In *Proceedings of the 36th International Conference on Machine Learning, ICML*.

Liu, Y.; Liu, R.; Wang, H.; and Lu, F. 2021. Generalizing Gaze Estimation with Outlier-guided Collaborative Adaptation. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 3815–3824.

Meng, F.; Wang, Z.; and Zhang, M. 2024. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems, NeurIPS*.

Muennighoff, N.; Yang, Z.; Shi, W.; Li, X. L.; Fei-Fei, L.; Hajishirzi, H.; Zettlemoyer, L.; Liang, P.; Candès, E. J.; and Hashimoto, T. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Nie, S.; Zhu, F.; You, Z.; Zhang, X.; Ou, J.; Hu, J.; Zhou, J.; Lin, Y.; Wen, J.; and Li, C. 2025. Large Language Diffusion Models. *arXiv preprint arXiv:2502.09992*.

Pan, J.; Zhang, J.; Wang, X.; Yuan, L.; Peng, H.; and Suhr, A. 2025. TinyZero. https://github.com/Jiayi-Pan/TinyZero. Accessed: 2025-01-24.

Park, S.; Mello, S. D.; Molchanov, P.; Iqbal, U.; Hilliges, O.; and Kautz, J. 2019. Few-Shot Adaptive Gaze Estimation. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 9367–9376.

Ren, P.; Shi, C.; Wu, S.; Zhang, M.; Ren, Z.; de Rijke, M.; Chen, Z.; and Pei, J. 2024. MELoRA: Mini-Ensemble Low-Rank Adapters for Parameter-Efficient Fine-Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL*, 3052–3064.

Si, C.; Wang, X.; Yang, X.; Xu, Z.; Li, Q.; Dai, J.; Qiao, Y.; Yang, X.; and Shen, W. 2025. Maintaining Structural Integrity in Parameter Spaces for Parameter Efficient Fine-tuning. In *The Twelfth International Conference on Learning Representations, ICLR*.

Tian, S.; Tu, H.; He, L.; Wu, Y. I.; and Zheng, X. 2023. FreeGaze: A Framework for 3D Gaze Estimation Using Appearance Cues from a Facial Video. *Sensors*, 23(23): 9604.

Wang, H.; Li, Y.; Wang, S.; Chen, G.; and Chen, Y. 2025. MiLoRA: Harnessing Minor Singular Components for Parameter-Efficient LLM Finetuning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, 4823–4836.

Wang, Y.; Jiang, Y.; Li, J.; Ni, B.; Dai, W.; Li, C.; Xiong, H.; and Li, T. 2022. Contrastive Regression for Domain Adaptation on Gaze Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 19354–19363.

Wu, T.; Wang, J.; Zhao, Z.; and Wong, N. 2024. Mixture-of-Subspaces in Low-Rank Adaptation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 7880–7899.

Yu, Y.; Liu, G.; and Odobez, J. 2019. Improving Few-Shot User-Specific Gaze Adaptation via Gaze Redirection Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 11937–11946.

Zhang, F.; and Pilanci, M. 2024. Spectral Adapter: Fine-Tuning in Spectral Space. In *Advances in Neural Information Processing Systems 38, NeurIPS*.

Zhao, S.; Gupta, D.; Zheng, Q.; and Grover, A. 2025. d1: Scaling Reasoning in Diffusion Large Language Models via Reinforcement Learning. *arXiv preprint arXiv:2504.12216*.