SSC-LoRA: Semantic Spectral Control Low-Rank Adaptation for Generalizable 3D Gaze Estimation

Anonymous authors
Paper under double-blind review

Abstract

Accurately predicting 3D gaze directions from single images remains challenging due to the absence of unified, physically grounded representations and limited diverse annotated data. Although visual foundation models (VFMs) provide strong transferable priors, their inherent visual sensitivity misaligns substantially with downstream regression tasks. Specifically, gaze-irrelevant appearance attributes dominate feature learning and overshadow subtle gaze variations that are critical for precise prediction. Large-scale pretraining prioritizes generic visual capabilities without explicitly encoding fine-grained gaze sensitivity. When appearance patterns spuriously correlate with gaze signals during pretraining, such sensitivity imbalance induces shortcut learning in downstream adaptation and degrades cross-domain generalization. To address these issues, we propose a novel parameter-efficient fine-tuning framework, Semantic Spectral Control LoRA (SSC-LoRA). Built upon vanilla LoRA, SSC-LoRA integrates gradient-based semantic identification (SI) and spectral control (SC) for task-aligned adaptation. Specifically, SI first decomposes pretrained weights into singular components and ranks their responsiveness to gaze-relevant and gaze-irrelevant objectives. SC further suppresses weight updates from gaze-irrelevant spectral bases while reinforcing gaze-specific feature learning. Both modules operate inherently within LoRA's low-rank space without introducing additional trainable parameters. Evaluated on five gaze estimation benchmarks with CLIP-B/L and DINOv2-B/L backbones, SSC-LoRA consistently outperforms most SOTA methods in both in-domain fitting and cross-domain generalization. Extensive ablation analyses validate that the performance bottleneck stems from gaze-irrelevant-sensitive adaptation, and confirm that our spectral control mechanism effectively recalibrates update allocation across task-relevant and task-irrelevant feature subspaces.

1 Introduction

Gaze estimation aims to predict precise 3D gaze directions from single images under varying head poses, lighting conditions, and partial occlusions. Subtle eye orientation changes can significantly alter gaze labels, while illumination variations and facial accessories induce drastic appearance changes without affecting gaze direction. Learning to distinguish such task-relevant and task-irrelevant variations becomes particularly challenging given the limited scale and diversity of annotated data. Existing methods alleviate distracting appearance variations via task-specific feature learning (Cheng et al., 2022) or adopt text-guided CLIP adaptation for gaze modeling (Yin et al., 2024b; Wang et al., 2026). These approaches motivate the use of externally learned visual priors beyond limited gaze annotations to enhance generalization.

Visual foundation models (VFMs) provide transferable representations from large-scale pretraining. For instance, CLIP (Radford et al., 2021) learns via image-text alignment, and DINOv2 (Oquab et al., 2023) captures visual features through self-supervision. Adapting such models to gaze estimation enables leveraging this visual knowledge under limited task-specific annotations. Nevertheless, these broadly transferable representations lack sensitivity to the fine-grained variations required for gaze prediction. Pretraining objectives do not explicitly enforce discrimination of subtle gaze-

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.

direction shifts, and appearance cues useful during pretraining remain prominent even when they are gaze-irrelevant.

We call this phenomenon *factor-sensitivity mismatch*: pretrained features respond more strongly to gaze-irrelevant appearance cues than to the gaze variations that determine predicted behavior. Figure 1 illustrates this mismatch for CLIP. On IVGaze (Cheng et al., 2024b), features cluster by accessory conditions; on MPIIFaceGaze (Zhang et al., 2017), by subject identity. Coloring MPIIFaceGaze features by gaze yaw reveals mixed gaze values within identity clusters, violating the smoothness assumption that nearby features correspond to similar gaze directions. The features are thus structured primarily by identity rather than the gaze variations required for regression.

For cross-task transfer, factor-sensitivity mismatch encourages models to exploit spurious correlations between appearance cues and gaze labels. Such correlations reduce empirical training loss but do not generalize to unseen subjects or domains, thereby inducing shortcut learning and degrading cross-domain performance (Geirhos et al., 2020). LoRA enables parameter-efficient adaptation of visual foundation models by freezing pretrained weights and learning compact low-rank updates (Hu et al., 2021). Despite its computational and memory efficiency for gaze estimation, LoRA's rank constraint only reduces parameter counts and cannot disentangle gaze-relevant and gaze-irrelevant update directions. Consequently, low-rank fine-tuning still reinforces spurious appearance correlations inherited from pretraining. Resolving this fundamental mismatch requires explicitly identifying pretrained spectral components sensitive to distinct visual factors and dynamically regulating their weight updates during adaptation.

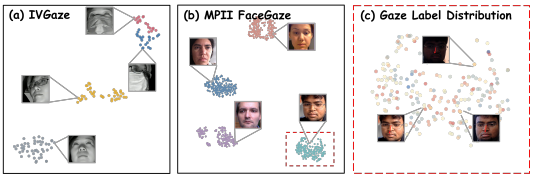


Figure 1: t-SNE (Van der Maaten & Hinton, 2008) visualizations for sample and label distribution. (a) Frozen CLIP on IVGaze, colored by accessory condition, including glasses only, mask only, neither, or both. (b) Frozen CLIP on MPIIFaceGaze, with colors denoting four different subjects. (c) Features of the highlighted subject in (b), colored by gaze yaw, show the same gaze directions within a single identity cluster but under different lighting conditions.

To address this issue, we propose *Semantic Spectral Control LoRA* (SSC-LoRA), a targeted fine-tuning framework built upon vanilla LoRA that integrates gradient-based semantic identification (SI) and spectral control (SC) to regulate parameter updates. Specifically, SI first decomposes pretrained weight matrices via singular value decomposition (SVD), factorizing each matrix into spectral components characterized by singular values and paired input–output basis directions. These components form the fundamental spectral subspace for our downstream control mechanism. SI further ranks individual spectral components according to their sensitivity toward gaze-relevant and gaze-irrelevant objectives. Concretely, a trainable gaze regressor provides gradient supervision from continuous gaze labels, while CLIP-derived pseudo-labels supervise the learning of gaze-irrelevant visual factors. Guided by the component ranking from SI, SC modulates LoRA weight updates through bilateral projection on the corresponding input and output subspaces, suppressing undesired updates within gaze-irrelevant subspaces while reinforcing task-beneficial updates in gaze-relevant subspaces. Both control branches reuse the inherent low-rank factors of LoRA, freeze the entire pretrained backbone, and introduce no additional trainable adapter parameters.

In short, our contributions are summarized as follows:

1) We identify *factor-sensitivity mismatch* as a core bottleneck for VFM-based gaze estimation, and analyze how gaze-irrelevant sensitivity induces spurious correlations and shortcut learning.

2) We propose SSC-LoRA, a parameter-efficient framework that distinguishes gaze-relevant and gaze-irrelevant spectral components via gradient scoring. It suppresses irrelevant weight updates and reinforces gaze-relevant learning without introducing extra trainable parameters.

3) Experiments across five mainstream benchmarks validate SSC-LoRA's effectiveness under both in-domain and cross-domain setups. Further analysis reveals that selecting targeted spectral components and modulating their updates greatly improves generalization compared to vanilla LoRA, without increasing the number of trainable parameters.

2 Related Work

Appearance-based Gaze Estimation directly predicts 3D gaze directions from facial or eye images (Zhang et al., 2015). Early deep models integrate cropped eye regions and head pose information for gaze estimation, while subsequent studies (Zhang et al., 2017) validate that full-face features deliver richer contextual cues. GazeTR (Cheng & Lu, 2022) further adopts pure and hybrid Transformer architectures to model spatial dependencies among facial regions. Despite strong in-domain performance, these methods are vulnerable to domain shifts caused by varying identities, illumination, accessories and camera parameters. Generalizable gaze estimation mitigates such issues without target-domain training data. PureGaze eliminates appearance distractions via self-adversarial feature purification (Cheng et al., 2022). Gaze-Consistent enforces gaze feature consistency across appearance perturbations of identity, expression and illumination (Xu et al., 2023). AGG (Bao & Lu, 2024) adopts an analytical feature-to-gaze mapping to replace data-driven regression heads, and FSCI (Liang et al., 2024) improves generalization via causal intervention on gaze-irrelevant features. Vision-language foundation models offer a promising alternative for robust gaze modeling. CLIP-Gaze disentangles gaze and irrelevant appearance features via personalized language prompt learning (Yin et al., 2024b). LG-Gaze aligns visual features with geometry-aware continuous gaze embeddings via multimodal contrastive regression (Yin et al., 2024a). GazeCLIP leverages linguistic gaze cues and cross-modal attention to enhance gaze feature learning (Wang et al., 2026). Similar to these methods, our SSC-LoRA exploits language priors to supervise gaze-irrelevant factor learning. Differently, it unifies language-derived irrelevant supervision and continuous gaze regression signals to identify task-sensitive spectral components, and precisely modulates low-rank weight updates for targeted, generalization-friendly fine-tuning.

Parameter-Efficient Fine-Tuning and Spectral Adaptation. PEFT fine-tunes large pretrained models while freezing most backbone parameters. Representative paradigms include input token learning (VPT) (Jia et al., 2022), lightweight transformer modules (AdaptFormer) (Chen et al., 2022), and low-rank weight updates (LoRA) (Hu et al., 2021). DoRA (Liu et al., 2024) further improves fine-tuning by disentangling weight magnitude and direction. A concurrent line of work adapts models based on downstream task sensitivity. SPT (He et al., 2023) allocates parameters via one-shot sensitivity estimation, and GPS (Zhang et al., 2024) selects weight connections according to gradient importance. Spectral adaptation explicitly leverages the singular structure of pretrained weights. PiSSA (Meng et al., 2024) initializes low-rank branches from principal singular components, while OPLoRA (Xiong & Xie, 2026) constrains updates within minor subspaces to preserve pretrained knowledge. Our SSC-LoRA also fine-tunes within singular subspaces, but differs fundamentally from existing spectral methods. Instead of relying solely on singular-value magnitude, it ranks spectral components by their gradient sensitivity to gaze-relevant and gaze-irrelevant factors, and adaptively suppresses irrelevant updates while reinforcing gaze-aligned learning.

3 Method

Figure 2 shows an overview of SSC-LoRA. The framework mitigates factor-sensitivity mismatch via two core components: gradient-based semantic identification (SI) and spectral control (SC).

3.1 Gradient-Driven Semantic Identification

Gradient-Based Spectral Scoring. To identify pretrained weight components sensitive to gaze prediction versus gaze-irrelevant factors for differentiated fine-tuning, we first analyze their responses in the spectral basis derived from singular value decomposition (SVD). Let the pretrained weight matrix be $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, where d_{in} and d_{out} denote its input and output dimensions,

$$W_0 = U\Sigma V^\top = \sum_{i=1}^{q} \sigma_i u_i v_i^\top, \tag{1}$$

where $q = \min(d_{\text{out}}, d_{\text{in}})$, σ_i denotes the i-th singular value, and u_i, v_i are the corresponding unit left and right singular vectors. Each rank-1 component $M_i = \sigma_i u_i v_i^\top$ projects the input direction v_i onto the output direction u_i. To distinguish spectral components driven by gaze-relevant and gaze-irrelevant factors, we score each component via task-specific gradient responses.

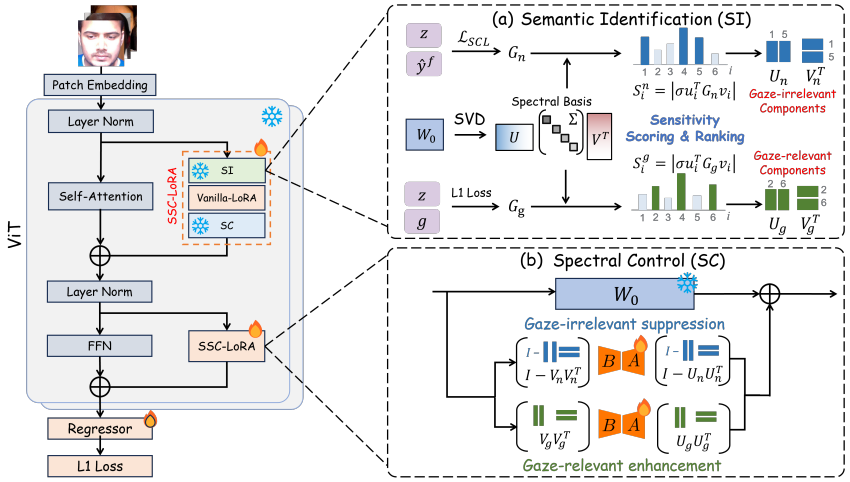


Figure 2: Overview of SSC-LoRA. Gradient-based semantic identification (SI) first leverages CLIP-generated pseudo-labels and gaze supervision to rank pretrained spectral components by gradient responses. Semantic spectral control (SC) then uses the selected input and output bases to suppress gaze-irrelevant updates and enhance gaze-relevant updates. The two branches share the same LoRA factors A and B and introduce no additional trainable parameters.

Definition 3.1 (Spectral Sensitivity)**.** For a scalar η near zero, we rescale M_i to $(1+\eta)M_i$ while keeping all other components unchanged, yielding the perturbed weight $W_0 + \eta M_i$. For a loss $\mathcal{L}$ differentiable at W_0, the sensitivity of M_i is defined as the magnitude of the loss derivative with respect to η evaluated at $\eta = 0$:

$$c_i(\mathcal{L}) = \left| \frac{\mathrm{d}}{\mathrm{d}\eta} \mathcal{L}(W_0 + \eta M_i) \Big|_{\eta=0} \right| = |\langle G, M_i \rangle_F| = \left| \sigma_i u_i^\top G v_i \right|. \tag{2}$$

Here, $G = \nabla_{W_0} \mathcal{L}$ denotes the gradient of $\mathcal{L}$ w.r.t. W_0, and $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. The equalities follow from the chain rule and the definition $M_i = \sigma_i u_i v_i^\top$.

A larger score implies the loss is more sensitive to small relative perturbations of that component. The absolute value captures the magnitude of the response, independent of its sign. This score is computed directly from gradients, without perturbing or updating the encoder weights. Appx. A.1 contains the full derivation.

To compute the scores in Eq. 2, we select losses tailored to the label type of each factor. We adopt the $L1$ regression loss $\mathcal{L}_g$ for continuous gaze angles. For each gaze-irrelevant factor f, we construct its loss $\mathcal{L}_f$ via supervised contrastive Loss (SCL) (Khosla et al., 2020) using the corresponding categorical pseudo-labels:

$$\mathcal{L}_g = \frac{1}{2|\mathcal{B}|} \sum_{j \in \mathcal{B}} \|h(z_j) - g_j\|_1, \quad \mathcal{L}_f = \mathcal{L}_{\mathrm{SCL}}\big(\{(z_j, \hat{y}_j^f)\}_{j \in \mathcal{B}}\big), \tag{3}$$

where $\mathcal{B}$ denotes a batch, z_j is the feature from the vision encoder, h is the gaze regressor, $g_j \in \mathbb{R}^2$ stores its ground-truth yaw and pitch, and $\hat{y}_j^f$ is its pseudo-label for factor f. Since existing VFMs do not natively include a gaze regression head (e.g., CLIP and DINOv2), we thus first train h using $\mathcal{L}_g$ with the encoder frozen. For factors lacking human annotations, we design class-specific text prompts and derive $\hat{y}_j^f$ via the pretrained vision-language model (e.g., CLIP). We then fix h and these pseudo-labels to compute scoring gradients without updating the encoder. Prompt and loss details are provided in Appx. B.3 and B.4, respectively.

Gradient Ranking and Component Selection. For each spectral component i, we compute gaze-relevant sensitivity s_i^g and gaze-irrelevant sensitivity s_i^f over the same T batches to obtain aggregate

responses over the sampled data.

$$s_i^g = \sum_{t=1}^{T} c_i(\mathcal{L}_g^{(t)}), \quad s_i^f = \sum_{t=1}^{T} c_i(\mathcal{L}_f^{(t)}). \tag{4}$$

We then combine normalized per-factor scores to form a unified score s_i^n. This per-factor normalization mitigates magnitude discrepancies across different gaze-irrelevant factors so that each factor contributes equally to the aggregated score:

$$s_i^n = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \frac{s_i^f}{\sum_{\ell=1}^{q} s_\ell^f}, \tag{5}$$

where $\mathcal{F}$ denotes the set of gaze-irrelevant factors. We compute s_i^g and s_i^n for every spectral component and sort the two score sets independently, producing rankings reflecting sensitivity to gaze-relevant and gaze-irrelevant factors, respectively. We select the Top-k spectral components from each ranking to form the gaze-relevant and gaze-irrelevant candidate index sets:

$$\mathcal{I}_g = \mathrm{TopK}(\{s_i^g\}_{i=1}^q, k), \quad \mathcal{I}_n = \mathrm{TopK}(\{s_i^n\}_{i=1}^q, k), \tag{6}$$

where TopK returns the indices of the k highest-scoring components. Note that the two index sets $\mathcal{I}_g$ and $\mathcal{I}_n$ are not mutually exclusive, so one spectral component may belong to both sets. We provide further discussion in §3.3.

3.2 Semantic Spectral Control

Spectral Subspaces Decomposition. Building on the identified spectral components, we suppress updates along gaze-irrelevant input or output directions and strengthen updates within the selected gaze-relevant subspaces. For a pretrained weight matrix W_0, vanilla LoRA (Hu et al., 2021) parameterizes the fine-tuned weight as

$$W_{\mathrm{ft}} = W_0 + \rho \Delta W, \quad \Delta W = BA, \quad \rho = \frac{\alpha}{r}, \tag{7}$$

where $A \in \mathbb{R}^{r \times d_{\mathrm{in}}}$ and $B \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ denote trainable low-rank matrices, r is the adapter rank, and α is a fixed scaling hyperparameter. The factor ρ rescales the low-rank update during forward propagation.

Definition 3.2 (Selected Spectral Subspaces and Projectors)**.** Consider a pretrained linear layer $y = W_0 x$, whose input and output feature spaces are $\mathbb{R}^{d_{\mathrm{in}}}$ and $\mathbb{R}^{d_{\mathrm{out}}}$, respectively. For a selected index set $\mathcal{I}_a$ with $a \in \{g, n\}$, let $U_a = [u_i]_{i \in \mathcal{I}_a}$ and $V_a = [v_i]_{i \in \mathcal{I}_a}$. The selected input and output subspaces are defined as

$$\begin{aligned} \mathcal{S}_{R,a} &= \mathrm{span}(V_a) = \{V_a c : c \in \mathbb{R}^{|\mathcal{I}_a|}\} \subseteq \mathbb{R}^{d_{\mathrm{in}}}, \\ \mathcal{S}_{L,a} &= \mathrm{span}(U_a) = \{U_a c : c \in \mathbb{R}^{|\mathcal{I}_a|}\} \subseteq \mathbb{R}^{d_{\mathrm{out}}}. \end{aligned} \tag{8}$$

The orthogonal projectors onto $\mathcal{S}_{L,a}$ and $\mathcal{S}_{R,a}$ are

$$P_a^L = U_a U_a^\top = \sum_{i \in \mathcal{I}_a} u_i u_i^\top, \quad P_a^R = V_a V_a^\top = \sum_{i \in \mathcal{I}_a} v_i v_i^\top. \tag{9}$$

For any matrix $W \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$ and input $x \in \mathbb{R}^{d_{\mathrm{in}}}$, right multiplication by P_a^R projects the input onto $\mathcal{S}_{R,a}$ prior to applying W, while left multiplication by P_a^L projects the resulting output onto $\mathcal{S}_{L,a}$:

$$(W P_a^R)x = W \underbrace{(P_a^R x)}_{\in \mathcal{S}_{R,a}}, \quad (P_a^L W)x = \underbrace{P_a^L (W x)}_{\in \mathcal{S}_{L,a}}. \tag{10}$$

Bilateral projection $P_a^L W P_a^R$ thus extracts the linear mapping restricted to the selected input and output subspaces. Appx. A.2 derives the projectors and their properties.

Definition 3.3 (Matrix Decomposition to Four Blocks)**.** For any spectral index set $\mathcal{I} \subseteq \{1, \ldots, q\}$ of W_0, we use the subspaces and projectors from Definition 3.2, omitting the factor index a to write $\mathcal{S}_L, \mathcal{S}_R$ and P^L, P^R. Any matrix $W \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$ admits the following decomposition:

$$\begin{aligned} W &= [(I - P^L) + P^L] W [(I - P^R) + P^R] \\ &= \underbrace{(I - P^L) W (I - P^R)}_{(1)\ \mathcal{S}_R^\perp \to \mathcal{S}_L^\perp} + \underbrace{(I - P^L) W P^R}_{(2)\ \mathcal{S}_R \to \mathcal{S}_L^\perp} + \underbrace{P^L W (I - P^R)}_{(3)\ \mathcal{S}_R^\perp \to \mathcal{S}_L} + \underbrace{P^L W P^R}_{(4)\ \mathcal{S}_R \to \mathcal{S}_L} \end{aligned} \tag{11}$$

where $\perp$ denotes the orthogonal complement of the corresponding input or output space. Each arrow indicates the input subspace and target output subspace for that block. For example, the second block $(I - P^L)WP^R$ maps inputs from $\mathcal{S}_R$ to outputs within $\mathcal{S}_L^\perp$, and annihilates inputs lying in $\mathcal{S}_R^\perp$.

Appx. A.3 derives this decomposition and characterizes the input-output restrictions of each block. The two cross-subspace blocks illustrate why suppressing updates along selected input or output directions requires bilateral projection.

Block-wise suppression and enhancement. We apply Definition 3.3 to the LoRA update separately with the gaze-irrelevant and gaze-relevant projector pairs. Each pair defines a four-block decomposition relative to its selected subspaces: the gaze-irrelevant projectors identify the blocks to suppress, while the gaze-relevant projectors identify the block to enhance.

For gaze-irrelevant control, we adopt $P^L = P_n^L$ and $P^R = P_n^R$. The second, third, and fourth blocks all involve selected gaze-irrelevant input or output directions. We discard these three blocks and keep only the first block: $(I-P_n^L)\Delta W(I-P_n^R)$. Output-only suppression preserves the second block, whereas input-only suppression retains the third. Bilateral suppression eliminates both cross-subspace mappings and satisfies $P_n^L(I-P_n^L)\Delta W(I-P_n^R) = 0$ and $(I-P_n^L)\Delta W(I-P_n^R)P_n^R = 0$.

For gaze-relevant enhancement, we adopt $P^L = P_g^L$ and $P^R = P_g^R$, the fourth block $P_g^L \Delta W P_g^R$ is the sole block whose input and output directions lie entirely within the selected gaze subspaces. We adopt this block to amplify the update over the mapping from $\mathcal{S}_{R,g}$ to $\mathcal{S}_{L,g}$.

To retain both effects within the combined update, gaze-irrelevant directions must remain excluded while the selected gaze block is amplified. We introduce the following assumption to avoid conflicting control over shared directions.

Assumption 3.4. The selected gaze-relevant and gaze-irrelevant spectral index sets from VFMs are disjoint, i.e., $\mathcal{I}_g \cap \mathcal{I}_n = \emptyset$.

Assumption 3.4 guarantees that the suppression operation leaves the gaze-relevant subspace untouched, and the enhancement operation does not reintroduce updates within the gaze-irrelevant subspaces (see Appx. A.4.1). The final SSC-LoRA update is written as:

$$\Delta W_{\mathrm{SSC}} = (I - P_n^L)\Delta W(I - P_n^R) + P_g^L \Delta W P_g^R. \tag{12}$$

Following vanilla LoRA notation, we combine this controlled low-rank update with a scaling factor ρ to obtain the adapted weight of the linear layer:

$$W_{\mathrm{ft}} = W_0 + \rho \Delta W_{\mathrm{SSC}}. \tag{13}$$

Appx. A.4.2 derives the preservation and enhancement properties of this update.

3.3 Discussion: Can the Disjointness Assumption Hold in Practice?

Our empirical study in §4.2 shows that gaze-relevant and gaze-irrelevant spectral selections within pretrained VFMs are not inherently disjoint. For CLIP-B/16 trained on Gaze360, Figure 3a reports a mean overlap of $\approx 40\%$ between gaze-relevant and factor-specific gaze-irrelevant top-32 selections. Layer-wise analysis in Appx. C.6 reveals larger overlap in earlier layers. This observation aligns with prior work showing that ViTs encode generic visual features in shallow layers and task-specific semantic concepts in deeper layers (Ghiasi et al., 2022; Dorszewski et al., 2025; Kim et al., 2026). Our results indicate that spectral sensitivities for gaze-relevant and gaze-irrelevant objectives become more separable with increasing network depth.

In practice, we enforce Assumption 3.4 by removing shared indices from the top-k candidate sets $\mathcal{I}_g$ and $\mathcal{I}_n$, constructing

$$\hat{\mathcal{I}}_g = \{i \in \mathcal{I}_g \mid i \notin \mathcal{I}_n\}, \quad \hat{\mathcal{I}}_n = \{i \in \mathcal{I}_n \mid i \notin \mathcal{I}_g\}, \tag{14}$$

which satisfy $\hat{\mathcal{I}}_g \cap \hat{\mathcal{I}}_n = \emptyset$. These disjoint sets replace $\mathcal{I}_g$ and $\mathcal{I}_n$ in our implementation. Spectral components falling into the shared intersection are excluded from both control branches and retain their original LoRA update magnitude, whereas updates associated with the remaining selected gaze-irrelevant input or output directions are suppressed.

Table 1: In-domain gaze estimation (MAE in degrees). † denotes using additional training data. Best and second-best results are in **bold** and underlined.

Method	MPIIFaceGaze	EyeDiap	Gaze360	IVGaze	GazeGene
FullFace [CVPRW17]	4.93	6.53	14.99	13.67	3.54
RT-Gene [ECCV18]	4.66	6.02	12.26	–	3.31
Dilated-Net [ACCV18]	4.42	6.19	13.73	–	**3.08**
Gaze360 [ICCV19]	4.06	5.36	11.04	8.15	3.53
CANet [AAAI20]	4.27	5.27	11.20	–	–
XGaze [ECCV20]	4.80	6.50	–	7.06	–
GazeTR [ICPR22]	4.00	5.17	10.62	7.33	3.37
AGE-Net [ICIP24]	3.61	4.78	–	–	–
3DGazeNet† [ECCV24]	4.00	–	9.60	–	–
OmniGaze† [NeurIPS25]	**2.97**	**4.07**	**9.12**	6.72	–
HARL† [AAAI26]	3.36	5.02	9.26	–	–
SSC-LoRA w/ CLIP-B/16	3.73	5.23	10.43	7.22	3.85
SSC-LoRA w/ CLIP-L/14	3.35	4.69	10.21	**6.61**	3.73
SSC-LoRA w/ DINOv2-B/14	3.99	5.28	10.64	7.22	3.63
SSC-LoRA w/ DINOv2-L/14	3.58	4.95	10.31	6.66	3.56

4 EXPERIMENTS

We evaluate SSC-LoRA on pretrained visual foundation models, including CLIP-B/16, CLIP-L/14, DINOv2-B/14, and DINOv2-L/14. Detailed experimental settings are provided in Appx. B. Following prior works, we conduct in-domain and cross-domain evaluations on widely used benchmarks: **Gaze360** (Kellnhofer et al., 2019), **MPIIFaceGaze** (Zhang et al., 2017), **EyeDiap** (Funes Mora et al., 2014), **IVGaze** (Cheng et al., 2024b), and **GazeGene** (Bao et al., 2025). Mean angular error (MAE) is adopted as the evaluation metric across all experiments. Dataset descriptions, train/validation/test splits, and supplementary metrics for IVGaze are given in Appx. B.1.

For comparison, we select SOTA gaze estimation methods: FullFace (Zhang et al., 2017), RT-Gene (Fischer et al., 2018), Dilated-Net (Chen & Shi, 2018), Gaze360 (Kellnhofer et al., 2019), CANet (Cheng et al., 2020), XGaze (Zhang et al., 2020), GazeTR (Cheng & Lu, 2022), AGE-Net (Shi et al., 2024), 3DGazeNet (Ververas et al., 2024), OmniGaze (Qu et al., 2026), and HARL (Tan et al., 2026). The cross-domain baselines includes PureGaze (Cheng et al., 2022), RAT (Bao et al., 2022), Gaze-Consistent (Xu et al., 2023), AGG (Bao & Lu, 2024), CLIP-Gaze (Yin et al., 2024b), FSCI (Liang et al., 2024), LG-Gaze (Yin et al., 2024a), and SeeTN (Peng et al., 2026).

4.1 COMPARISON WITH SOTA METHODS

In-domain Evaluation. Table 1 compares SSC-LoRA with SOTA gaze estimation methods on five benchmarks. With CLIP-L/14, SSC-LoRA gives the lowest estimation errors among methods trained on the benchmark data alone on MPIIFaceGaze, EyeDiap, Gaze360, and IVGaze. Relative to the strongest method under the same data condition, the errors on MPIIFaceGaze, EyeDiap, and Gaze360 decrease by $0.26°$, $0.09°$, and $0.41°$, respectively. These gains cover both controlled recordings and unconstrained scenes, showing that the adapted encoder generalizes across held-out subjects and recording conditions. On IVGaze, the error of $6.61°$ improves on GazeDPTR's $6.71°$ by $0.10°$, as reported in Table 7. SSC-LoRA also outperforms OmniGaze, which is trained with large-scale external auxiliary data. The model with DINOv2-L/14 improves over DINOv2-B/14 on all five benchmarks and achieves $6.66°$ on IVGaze, the second-lowest error among the compared methods. Appx. C.1 and C.2 further compare our method with SOTA PEFT baselines and report performance across error thresholds and accessory partitions on IVGaze.

Cross-domain Evaluation. We evaluate four transfer settings: from Gaze360 to MPIIFaceGaze and EyeDiap ($\mathcal{D}_{360} \rightarrow \mathcal{D}_{\mathrm{M}}$, $\mathcal{D}_{360} \rightarrow \mathcal{D}_{\mathrm{D}}$), and from GazeGene to the same two datasets ($\mathcal{D}_{\mathrm{Gene}} \rightarrow \mathcal{D}_{\mathrm{M}}$, $\mathcal{D}_{\mathrm{Gene}} \rightarrow \mathcal{D}_{\mathrm{D}}$). In particular, the two GazeGene settings evaluate transfer from synthetic to real faces. As shown in Table 2, when trained on Gaze360, SSC-LoRA with CLIP-L/14 achieve better performance compared with SeeTN and LG-Gaze on the respective targets by $0.65°$ and $0.09°$ improvement. With GazeGene as the source domain, increasing the model from CLIP-B/16 to

Table 2: Cross-domain gaze estimation (MAE in degrees). $\mathcal{D}_{360}$, $\mathcal{D}_{\text{Gene}}$, $\mathcal{D}_{\text{M}}$, and $\mathcal{D}_{\text{D}}$ denote Gaze360, GazeGene, MPIIFaceGaze, and EyeDiap. † denotes using additional training data. Best and second-best results are in **bold** and underlined.

Method	$\mathcal{D}_{360} \to \mathcal{D}_{\text{M}}$	$\mathcal{D}_{360} \to \mathcal{D}_{\text{D}}$	$\mathcal{D}_{\text{Gene}} \to \mathcal{D}_{\text{M}}$	$\mathcal{D}_{\text{Gene}} \to \mathcal{D}_{\text{D}}$
FullFace [CVPRW17]	12.35	30.15	9.89	23.13
CANet [AAAI20]	27.13	31.41	–	–
PureGaze [AAAI22]	9.28	9.32	7.75	18.24
GazeTR [ICPR22]	7.96	8.88	7.80	18.20
RAT [CVPR22]	7.69	7.08	–	–
Gaze-Consistent [AAAI23]	7.55	9.03	–	–
AGG [CVPR24]	7.87	7.90	**5.61**	10.34
CLIP-Gaze [AAAI24]	6.89	7.06	–	–
FSCI [ECCV24]	7.06	7.99	5.97	16.90
LG-Gaze [ECCV24]	6.83	6.86	–	–
OmniGaze† [NeurIPS25]	**4.95**	**5.75**	–	–
HARL† [AAAI26]	7.49	7.63	–	–
SeeTN [CVPR26]	6.57	7.57	–	–
SSC-LoRA w/ CLIP-B/16	6.83	7.15	6.21	11.74
SSC-LoRA w/ CLIP-L/14	<u>5.92</u>	6.77	6.01	11.02
SSC-LoRA w/ DINOv2-B/14	7.06	7.16	<u>5.63</u>	<u>9.50</u>
SSC-LoRA w/ DINOv2-L/14	6.65	<u>6.72</u>	5.79	**9.21**

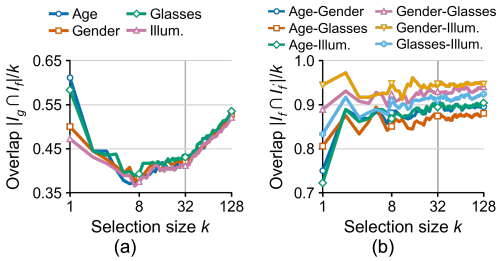


Figure 3: Top-k overlap on Gaze360 between (a) gaze and each gaze-irrelevant factor and (b) pairs of gaze-irrelevant factors

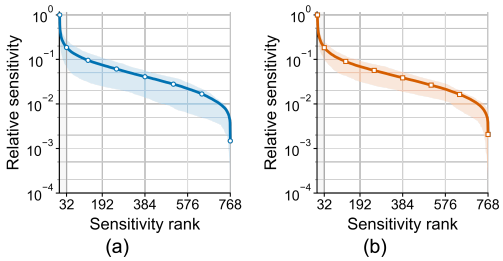


Figure 4: Normalized sensitivity scores for (a) gaze and (b) gaze-irrelevant objectives.

CLIP-L/14 reduces the MPIIFaceGaze error from $6.21°$ to $6.01°$ and the EyeDiap error from $11.74°$ to $11.02°$. Both DINOv2 backbones achieve lower errors than their corresponding CLIP model sizes in these synthetic-to-real cross-domain settings. Specifically, DINOv2-B/14 obtains $5.63°$ and $9.50°$ on the two targets, while DINOv2-L/14 obtains $5.79°$ and $9.21°$. These results demonstrate that SSC-LoRA improves cross-domain generalization and achieves strong performance even in challenging synthetic-to-real transfer.

4.2 Ablation Studies

All analyses and ablation studies use CLIP-B/16 that was trained on the Gaze360, and we only report in-domain and cross-domain results with it.

4.2.1 Ablation of Semantic Identification

Figure 3 compares the Top-k selections across objectives. At $k = 32$, gaze-relevant and gaze-irrelevant selections overlap by $\sim 40\%$, compared with $\sim 85\%$ overlap among gaze-irrelevant factors. This gap demonstrates that SI differentiates components responsive to the two objective categories, supporting separate enhancement and suppression even in the presence of shared indices.

We next examine sensitivity within each ranking. Figure 4 show high scores for the leading components, followed by a sharp drop and a low-scoring tail. This concentration motivates prioritizing these top components for our control mechanism.

Table 3: Effect of selecting spectral components from different gradient-ranked regions. Results are reported as mean±standard deviation over three random samplings of spectral components from the Middle or Low regions, with the training seed fixed.

Selected region		Gaze360	$\mathcal{D}_{360} \to \mathcal{D}_M$	$\mathcal{D}_{360} \to \mathcal{D}_D$
Gaze	**Gaze-irrelevant**			
Low	High	$11.02_{\pm 0.03}$	$7.23_{\pm 0.11}$	$\underline{7.71}_{\pm 0.33}$
Middle	High	$\underline{10.98}_{\pm 0.07}$	$7.42_{\pm 0.28}$	$7.89_{\pm 0.07}$
High	Low	$11.03_{\pm 0.20}$	$7.25_{\pm 0.11}$	$7.72_{\pm 0.37}$
High	Middle	$11.02_{\pm 0.15}$	$\underline{7.13}_{\pm 0.22}$	$7.87_{\pm 0.37}$
High	High	**10.43**	**6.83**	**7.15**

Table 4: Ablation of the semantic-control branches and their input/output application. MAE is reported in degrees, and lower is better.

Variant	**Update**	Gaze360	$\mathcal{D}_{360} \to \mathcal{D}_M$	$\mathcal{D}_{360} \to \mathcal{D}_D$
Vanilla LoRA	ΔW	11.02	7.17	8.39
w/ P_n only	$(I - P_n^L)\Delta W(I - P_n^R)$	<u>11.01</u>	6.88	7.91
w/ P_g only	$P_g^L \Delta W P_g^R$	11.34	<u>6.87</u>	<u>7.56</u>
Output-only	$(I - P_n^L)\Delta W + P_g^L \Delta W$	11.05	7.26	7.87
Input-only	$\Delta W(I - P_n^R) + \Delta W P_g^R$	11.15	7.47	7.61
Full (SSC-LoRA)	$(I - P_n^L)\Delta W(I - P_n^R) + P_g^L \Delta W P_g^R$	**10.43**	**6.83**	**7.15**

To test whether these components improve adaptation, we compare High (ranks 1–32), Middle (129–640), and Low (641–768) selections in Table 3. We sample 32 candidates for each objective: High corresponds to the Top-32 components, while Middle and Low groups are sampled without replacement. We fix one objective's selection to High and vary the other. The High–High configuration yields the lowest error across all three settings; replacing either selection with Middle or Low components increases the mean gaze error. These results validate the use of SI's top-ranked components for gaze-relevant enhancement and gaze-irrelevant suppression.

4.2.2 Ablation of Spectral Control

Table 4 evaluates the two control branches and their input/output projection settings. Both suppression-only (denoted by w/ P_n only) and enhancement-only (w/ P_g only) updates improve cross-domain performance relative to vanilla LoRA, and combining both branches yields the lowest errors across all three settings. The two branches are complementary: suppression removes updates along selected gaze-irrelevant input or output directions, while enhancement strengthens the selected gaze mapping without reintroducing such spurious directions. Full model also surpasses input-only and output-only control on both target datasets. As shown by the four-block decomposition in Eq. 11, single-sided suppression retains updates along gaze-irrelevant directions on the uncontrolled subspace, and single-sided enhancement reinforces cross-subspace updates outside the selected gaze block. Bilateral control mitigates both limitations: it suppresses updates involving any selected gaze-irrelevant input or output directions and confines enhancement strictly within the selected gaze subspaces. Further analysis of spectral update allocation is provided in Appx. C.7.

5 Conclusion

This paper presents SSC-LoRA, a novel and efficient spectral-controlled low-rank adaptation framework for 3D gaze estimation. Unlike vanilla LoRA, which jointly updates all spectral components, SSC-LoRA leverages objective-aware spectral sensitivity to separate gaze-relevant and gaze-irrelevant subspaces from pretrained visual foundation models. Guided by these semantic bases, bilateral control regularizes LoRA updates: it preserves gaze-irrelevant mappings while amplifying updates only within gaze-relevant subspaces. Evaluated on five gaze benchmarks with CLIP and DINOv2, SSC-LoRA achieves highly competitive performance against SOTA gaze estimation and PEFT baselines, with no increase in trainable adapter parameters.

AI USE STATEMENT

In the preparation of this paper, the authors used the AI tools to assist with language refinement, including improving and checking potential grammatical issues as well as enhancing clarity and readability. AI tools were not used to develop the core methodology, conduct experiments, analyze data, interpret the results. The authors take full responsibility for the accuracy and integrity of the paper's content.

REFERENCES

Yiwei Bao and Feng Lu. From feature to gaze: A generalizable replacement of linear layer for gaze estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1409–1418. IEEE, 2024.

Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4197–4206. IEEE, 2022.

Yiwei Bao, Zhiming Wang, and Feng Lu. Gazegene: Large-scale synthetic gaze dataset with 3d eyeball annotations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18749–18759. IEEE, 2025.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in neural information processing systems*, 35:16664–16678, 2022.

Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian conference on computer vision*, pp. 309–324. Springer, 2018.

Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 3341–3347. IEEE, 2022.

Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 10623–10630, 2020.

Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 436–443, 2022.

Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7509–7528, 2024a.

Yihua Cheng, Yaning Zhu, Zongji Wang, Hongquan Hao, Yongwei Liu, Shiqing Cheng, Xi Wang, and Hyung Jin Chang. What do you see in vehicle? comprehensive vision solution for in-vehicle gaze estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1556–1565. IEEE, 2024b.

Teresa Dorszewski, Lenka Tětková, Robert Jenssen, Lars Kai Hansen, and Kristoffer Knutsen Wickstrøm. From colors to classes: Emergence of concepts in vision transformers. In *World Conference on Explainable Artificial Intelligence*, pp. 28–47. Springer, 2025.

Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. RT-GENE: Real-time eye gaze estimation in natural environments. In *European conference on computer vision*, pp. 339–357. Springer, 2018.

Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, pp. 255–258, 2014.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022.

Shreya Ghosh, Abhinav Dhall, Garima Sharma, Sarthak Gupta, and Nicu Sebe. Speak2label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2896–2905. IEEE, 2021.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024.

Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. Sensitivity-aware visual parameter-efficient fine-tuning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11791–11801. IEEE, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

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.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6911–6920. IEEE, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Jinyeong Kim, Junhyeok Kim, Yumin Shim, Joohyeok Kim, Sunyoung Jung, and Seong Jae Hwang. Interpreting vision transformers via residual replacement model. *Advances in Neural Information Processing Systems*, 38:55760–55820, 2026.

Dawid Kopiczko, Tijmen Blankevoort, and Yuki Asano. Vera: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 6815–6835, 2024.

Ziyang Liang, Yiwei Bao, and Feng Lu. De-confounded gaze estimation. In *European Conference on Computer Vision*, pp. 219–235. Springer, 2024.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yanming Peng, Shijing Wang, Yaping Huang, and Yi Tian. See through the noise: Improving domain generalization in gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31346–31355, June 2026.

Hongyu Qu, Jianan Wei, Xiangbo Shu, Yazhou Yao, Wenguan Wang, and Jinhui Tang. Omnigaze: Reward-inspired generalizable gaze estimation in the wild. *Advances in Neural Information Processing Systems*, 38:105999–106026, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yichen Shi, Feifei Zhang, Wenming Yang, Guijin Wang, and Nan Su. Agent-guided gaze estimation network by two-eye asymmetry exploration. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 2320–2326. IEEE, 2024.

Qida Tan, Hongyu Yang, and Wenchao Du. Hybrid-domain adaptative representation learning for gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9323–9331, 2026.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Evangelos Ververas, Polydefkis Gkagkos, Jiankang Deng, Michail Christos Doukas, Jia Guo, and Stefanos Zafeiriou. 3dgazenet: Generalizing 3d gaze estimation with weak-supervision from synthetic views. In *European Conference on Computer Vision*, pp. 387–404. Springer, 2024.

Jun Wang, Hao Ruan, Liangjian Wen, Yong Dai, and Mingjie Wang. Gazeclip: Enhancing gaze estimation through text-guided multimodal learning. *Neurocomputing*, pp. 133584, 2026.

Yifeng Xiong and Xiaohui Xie. OPLoRA: Orthogonal projection LoRA prevents catastrophic forgetting during parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34088–34096, 2026.

Mingjie Xu, Haofei Wang, and Feng Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3027–3035, 2023.

Pengwei Yin, Jingjing Wang, Guanzhong Zeng, Di Xie, and Jiang Zhu. Lg-gaze: Learning geometry-aware continuous prompts for language-guided gaze estimation. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024a.

Pengwei Yin, Guanzhong Zeng, Jingjing Wang, and Di Xie. Clip-gaze: Towards general gaze estimation via visual-linguistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6729–6737, 2024b.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4511–4520, 2015.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 2299–2308. IEEE, 2017.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European conference on computer vision*, pp. 365–381. Springer, 2020.

Zhi Zhang, Qizhe Zhang, Zijun Gao, Renrui Zhang, Ekaterina Shutova, Shiji Zhou, and Shanghang Zhang. Gradient-based parameter selection for efficient fine-tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28566–28577. IEEE, 2024.

A THEORETICAL DERIVATIONS AND ANALYSIS

A.1 DERIVATION OF SPECTRAL SENSITIVITY SCORE

Consider one frozen pretrained matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ with singular value decomposition (SVD):

$$W_0 = \sum_{i=1}^{q} \sigma_i u_i v_i^\top, \quad q = \min(d_{\text{out}}, d_{\text{in}}).$$

We treat $M_i = \sigma_i u_i v_i^\top$ as the i-th spectral component. For a loss $\mathcal{L}$ evaluated over one batch, let $G = \nabla_{W_0} \mathcal{L}$. We perturb only the magnitude of this component via a scalar η:

$$W_0^{(i)}(\eta) = W_0 + \eta M_i.$$

The first-order Taylor expansion around W_0 reads

$$\begin{aligned} \mathcal{L}(W_0^{(i)}(\eta)) &= \mathcal{L}(W_0) + \eta \langle G, M_i \rangle_F + o(|\eta|) \\ &= \mathcal{L}(W_0) + \eta \sigma_i u_i^\top G v_i + o(|\eta|). \end{aligned} \tag{15}$$

Since $o(|\eta|)/\eta \to 0$, differentiating this expansion at $\eta = 0$ yields

$$\begin{aligned} \left| \frac{\mathrm{d}}{\mathrm{d}\eta} \mathcal{L}(W_0^{(i)}(\eta)) \Big|_{\eta=0} \right| &= |\langle G, M_i \rangle_F| = \left| \sum_{a=1}^{d_{\text{out}}} \sum_{b=1}^{d_{\text{in}}} G_{ab} (M_i)_{ab} \right| \\ &= \left| \sigma_i \sum_{a=1}^{d_{\text{out}}} \sum_{b=1}^{d_{\text{in}}} (u_i)_a G_{ab} (v_i)_b \right| \\ &= \left| \sigma_i u_i^\top G v_i \right| = c_i(\mathcal{L}). \end{aligned} \tag{16}$$

The first line expands the Frobenius inner product entrywise. The second substitutes $(M_i)_{ab} = \sigma_i (u_i)_a (v_i)_b$, and the final line rewrites the double summation as a matrix product.

A.2 SPECTRAL PROJECTORS

This subsection derives the orthogonal projectors introduced in Definition 3.2 and establishes their basic properties.

For an index set $\mathcal{I}_a$, where $a \in \{g, n\}$, define

$$U_a = [u_i]_{i \in \mathcal{I}_a}, \quad V_a = [v_i]_{i \in \mathcal{I}_a}.$$

Following Definition 3.2, the selected input and output subspaces are

$$\mathcal{S}_{R,a} = \operatorname{span}(V_a), \quad \mathcal{S}_{L,a} = \operatorname{span}(U_a),$$

with corresponding orthogonal projectors

$$P_a^L = U_a U_a^\top, \quad P_a^R = V_a V_a^\top.$$

The selected singular vectors form orthonormal bases, so $U_a^\top U_a = I$ and $V_a^\top V_a = I$. For any output vector $y \in \mathbb{R}^{d_{\text{out}}}$, its coordinates along the selected left singular vectors are $U_a^\top y$. Reconstructing from these coordinates yields

$$\operatorname{proj}_{\mathcal{S}_{L,a}}(y) = \sum_{i \in \mathcal{I}_a} (u_i^\top y) u_i = U_a U_a^\top y = P_a^L y. \tag{17}$$

The residual is orthogonal to the selected subspace since $U_a^\top (y - P_a^L y) = 0$. Moreover,

$$(P_a^L)^\top = P_a^L, \quad (P_a^L)^2 = U_a (U_a^\top U_a) U_a^\top = P_a^L. \tag{18}$$

Thus P_a^L is the orthogonal projector onto $\mathcal{S}_{L,a}$. The same reasoning gives $\operatorname{proj}_{\mathcal{S}_{R,a}}(x) = V_a V_a^\top x = P_a^R x$ for any input vector $x \in \mathbb{R}^{d_{\text{in}}}$, with $(P_a^R)^\top = P_a^R$ and $(P_a^R)^2 = P_a^R$. The complementary projectors $I - P_a^L$ and $I - P_a^R$ project onto $\mathcal{S}_{L,a}^\perp$ and $\mathcal{S}_{R,a}^\perp$, respectively. These properties directly follow from the orthonormality of the selected basis vectors.

A.3 FOUR-BLOCK DECOMPOSITION

Using the projectors from Definition 3.2, we derive the decomposition stated in Definition 3.3 and discuss its application to LoRA updates. This result holds for arbitrary selected index sets and does not require disjoint gaze-relevant and gaze-irrelevant selections.

For any weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, we insert identity decompositions on both sides:

$$\begin{aligned} W &= [(I - P^L) + P^L]W[(I - P^R) + P^R] \\ &= (I - P^L)W[(I - P^R) + P^R] + P^L W[(I - P^R) + P^R] \\ &= (I - P^L)W(I - P^R) + (I - P^L)WP^R \\ &\quad + P^L W(I - P^R) + P^L WP^R. \end{aligned}$$

Thus the four blocks sum exactly to W, as stated in Definition 3.3.

To derive their subspace mapping properties, choose $D^L \in \{P^L, I - P^L\}$ and $D^R \in \{P^R, I - P^R\}$, and define $W_{\text{blk}} = D^L W D^R$. Each D^L, D^R is an orthogonal projector; for any orthogonal projector P, $(I - P)^2 = I - 2P + P^2 = I - P$. Hence $(D^L)^2 = D^L$ and $(D^R)^2 = D^R$.

Input Restriction. Right multiplication by D^R preserves W_{blk}, while multiplication by $I - D^R$ yields zero:

$$\begin{aligned} W_{\text{blk}} D^R &= D^L W D^R D^R = D^L W (D^R)^2 = D^L W D^R = W_{\text{blk}}, \\ W_{\text{blk}}(I - D^R) &= D^L W D^R (I - D^R) = D^L W \left[D^R - (D^R)^2\right] = 0. \end{aligned}$$

Every input feature vector $x \in \mathbb{R}^{d_{\text{in}}}$ decomposes as

$$x = D^R x + (I - D^R)x,$$

where $D^R x$ is the orthogonal projection of x onto $\text{range}(D^R)$, and $(I - D^R)x$ lies in its orthogonal complement $\text{range}(D^R)^\perp$. Here, $\text{range}(D^R)$ equals $\mathcal{S}_R$ when $D^R = P^R$, and $\mathcal{S}_R^\perp$ when $D^R = I - P^R$. Substituting this decomposition and applying the identities above gives

$$W_{\text{blk}} x = W_{\text{blk}} D^R x + W_{\text{blk}}(I - D^R)x = W_{\text{blk}} D^R x.$$

Thus the block's output depends only on $D^R x$, and the complementary component $(I - D^R)x$ contributes nothing. Consequently, inputs sharing the same projection $D^R x$ yield identical outputs through this block, regardless of their components within $\text{range}(D^R)^\perp$. In particular, if $x \in \text{range}(D^R)^\perp$, then $D^R x = 0$ and $W_{\text{blk}} x = 0$.

Output Restriction. Similarly, left multiplication yields

$$\begin{aligned} D^L W_{\text{blk}} &= D^L D^L W D^R = (D^L)^2 W D^R = D^L W D^R = W_{\text{blk}}, \\ (I - D^L)W_{\text{blk}} &= (I - D^L) D^L W D^R = \left[D^L - (D^L)^2\right] W D^R = 0. \end{aligned}$$

For any input $x \in \mathbb{R}^{d_{\text{in}}}$, let $y_{\text{blk}} = W_{\text{blk}} x$ denote the block's output. The identities above imply

$$\begin{aligned} D^L y_{\text{blk}} &= D^L W_{\text{blk}} x = W_{\text{blk}} x = y_{\text{blk}}, \\ (I - D^L) y_{\text{blk}} &= (I - D^L) W_{\text{blk}} x = 0. \end{aligned}$$

Thus y_{blk} lies entirely in $\text{range}(D^L)$, with no component in its orthogonal complement. Since this holds for every input x, we obtain $\text{range}(W_{\text{blk}}) \subseteq \text{range}(D^L)$.

These subspace restrictions hold for all four combinations of (D^L, D^R). For example, $D^L = I - P^L$ and $D^R = P^R$ produce $W_{\text{blk}} = (I - P^L)WP^R$: this block maps $\mathcal{S}_R$ into $\mathcal{S}_L^\perp$ and vanishes on $\mathcal{S}_R^\perp$. Selecting the complements on both sides yields $(I - P^L)W(I - P^R)$, which discards contributions from both the selected input and output subspaces.

Application to Low-Rank Updates. The four-block decomposition applies to any matrix of the same dimension as W_0, regardless of its rank. Since the LoRA update $\Delta W = BA \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$

matches these dimensions, we may set $W = \Delta W$ and reuse the same projectors. Each resulting block satisfies

$$D^L \Delta W D^R = \underbrace{(D^L B)}_{d_{\text{out}} \times r} \underbrace{(A D^R)}_{r \times d_{\text{in}}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}, \quad \operatorname{rank}(D^L \Delta W D^R) \le r.$$

Therefore, LoRA updates admit the identical four-block decomposition. Each block has rank at most r, and the four blocks sum to the full update $\Delta W = BA$.

A.4 Compatibility and Effect of SC

We first analyze the compatibility of suppression and enhancement under Assumption 3.4, and further derive their combined effects on the adapted model weights. For any matrix W sharing the same dimension as W_0, we define the suppression and enhancement operators as

$$\Pi_{\bar{n}}(W) = (I - P_n^L) W (I - P_n^R), \quad \Pi_g(W) = P_g^L W P_g^R.$$

The overall transformation of the proposed semantic spectral control is formulated as

$$\mathcal{T}_{\text{SSC}}(W) = \Pi_{\bar{n}}(W) + \Pi_g(W).$$

A.4.1 Compatibility of Suppression and Enhancement

First, disjoint index sets induce orthogonal subspaces. For any $i \in \mathcal{I}_n$ and $j \in \mathcal{I}_g$, Assumption 3.4 guarantees $i \neq j$. Since all singular vectors from the SVD basis are orthonormal, it follows that

$$u_i^\top u_j = 0, \quad v_i^\top v_j = 0, \quad \text{which yields} \quad U_n^\top U_g = \mathbf{0}, \quad V_n^\top V_g = \mathbf{0}.$$

Consequently, $\mathcal{S}_{L,g} \subseteq \mathcal{S}_{L,n}^{\perp}$ and $\mathcal{S}_{R,g} \subseteq \mathcal{S}_{R,n}^{\perp}$.

Second, eliminating gaze-irrelevant directions preserves all gaze-relevant basis vectors. For each $j \in \mathcal{I}_g$, the gaze-relevant singular vectors are orthogonal to all gaze-irrelevant singular vectors. This gives

$$P_n^L u_j = \sum_{i \in \mathcal{I}_n} u_i (u_i^\top u_j) = 0, \qquad (I - P_n^L) u_j = u_j,$$
$$P_n^R v_j = \sum_{i \in \mathcal{I}_n} v_i (v_i^\top v_j) = 0, \qquad (I - P_n^R) v_j = v_j.$$

Thus, the suppression operation leaves every selected gaze basis vector unchanged. Substituting the projector definitions yields

$$(I - P_n^L) P_g^L = \sum_{j \in \mathcal{I}_g} \big[(I - P_n^L) u_j\big] u_j^\top = \sum_{j \in \mathcal{I}_g} u_j u_j^\top = P_g^L,$$
$$(I - P_n^R) P_g^R = \sum_{j \in \mathcal{I}_g} \big[(I - P_n^R) v_j\big] v_j^\top = \sum_{j \in \mathcal{I}_g} v_j v_j^\top = P_g^R.$$

Taking the transpose and exploiting projector symmetry leads to

$$\begin{aligned} P_g^L (I - P_n^L) &= (I - P_n^L) P_g^L = P_g^L, \\ P_g^R (I - P_n^R) &= (I - P_n^R) P_g^R = P_g^R. \end{aligned} \tag{19}$$

Third, the gaze-relevant block is invariant to suppression. Applying Eq. 19 to both sides of the suppressed update yields

$$\begin{aligned} & P_g^L \big[(I - P_n^L) \Delta W (I - P_n^R)\big] P_g^R \\ &= \big[P_g^L (I - P_n^L)\big] \Delta W \big[(I - P_n^R) P_g^R\big] \\ &= P_g^L \Delta W P_g^R. \end{aligned} \tag{20}$$

In terms of the previously defined operators, this result implies

$$\Pi_{\bar{n}}\big(\Pi_g(W)\big) = \Pi_g\big(\Pi_{\bar{n}}(W)\big) = \Pi_g(W).$$

These identities demonstrate that suppression leaves the gaze-relevant subspace block fully invariant, regardless of application order. Therefore, explicitly adding the enhanced gaze block does not reintroduce undesired gaze-irrelevant update components.

A.4.2 Effect of SC on the Adapted Weight

Preservation along the selected gaze-irrelevant directions. The projector properties yield $P_n^L(I - P_n^L) = 0$, and the orthogonality derived in Appx. A.4.1 gives $P_n^L P_g^L = 0$. Left-multiplying the SSC-LoRA update from Eq. 12 by P_n^L therefore yields zero. Analogous properties for the right-hand projector produce the same result upon right multiplication by P_n^R:

$$P_n^L \Delta W_{\text{SSC}} = \mathbf{0}, \quad \Delta W_{\text{SSC}} P_n^R = \mathbf{0}. \tag{21}$$

For each $i \in \mathcal{I}_n$, we have $P_n^R v_i = v_i$ and $P_n^L u_i = u_i$. Consequently,

$$\Delta W_{\text{SSC}} v_i = \Delta W_{\text{SSC}} P_n^R v_i = 0,$$
$$\Delta W_{\text{SSC}}^\top u_i = (P_n^L \Delta W_{\text{SSC}})^\top u_i = 0.$$

Substituting into the adapted weight matrix yields

$$W_{\text{ft}} v_i = W_0 v_i + \rho \Delta W_{\text{SSC}} v_i = W_0 v_i = \sigma_i u_i,$$
$$W_{\text{ft}}^\top u_i = W_0^\top u_i + \rho \Delta W_{\text{SSC}}^\top u_i = W_0^\top u_i = \sigma_i v_i.$$

Hence the adapted weight preserves the pretrained responses along these selected gaze-irrelevant directions.

Enhancement of the selected gaze update. The enhancement property applies exclusively to the gaze-relevant block of the update. We inspect this block via $P_g^L \Delta W_{\text{SSC}} P_g^R$: the right projector selects gaze-relevant input directions, while the left projector selects gaze-relevant output directions. Eq. 20 shows that the suppression term preserves the gaze-relevant block. The enhancement term adds an identical block, since $(P_g^L)^2 = P_g^L$ and $(P_g^R)^2 = P_g^R$. Accordingly,

$$\begin{aligned} P_g^L \Delta W_{\text{SSC}} P_g^R &= P_g^L \left[(I - P_n^L) \Delta W (I - P_n^R) + P_g^L \Delta W P_g^R \right] P_g^R \\ &= P_g^L (I - P_n^L) \Delta W (I - P_n^R) P_g^R + (P_g^L)^2 \Delta W (P_g^R)^2 \\ &= P_g^L \Delta W P_g^R + P_g^L \Delta W P_g^R \\ &= 2\, P_g^L \Delta W P_g^R. \end{aligned}$$

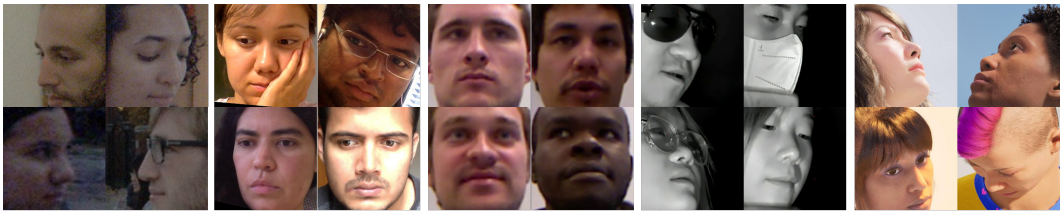
Substituting into the adapted weight gives

$$\begin{aligned} P_g^L W_{\text{ft}} P_g^R &= P_g^L W_0 P_g^R + \rho P_g^L \Delta W_{\text{SSC}} P_g^R \\ &= P_g^L W_0 P_g^R + 2\rho\, P_g^L \Delta W P_g^R. \end{aligned}$$

Outside the selected gaze-relevant block, the update components retained by suppression remain unaffected by the enhancement branch.

B Experimental Settings

B.1 Datasets and Evaluation Metrics



(a) Gaze360 (b) MPIIFaceGaze (c) EyeDiap (d) IVGaze (e) GazeGene

Figure 5: Representative samples from the datasets used in our experiments. Four samples are shown for each dataset to illustrate its visual conditions.

Datasets. We briefly describe each dataset below. For Gaze360, MPIIFaceGaze, and EyeDiap, we adopt the preprocessing pipeline from Cheng et al. (2024a) for their gaze estimation benchmark.

Figure 5 shows representative samples from the five datasets used in our experiments. These datasets span unconstrained indoor/outdoor scenes, everyday laptop usage, laboratory recordings, in-vehicle scenarios, and synthetic multi-camera captures.

- **Gaze360** (Kellnhofer et al., 2019). Gaze360 consists of 197,588 frames from 238 subjects captured across indoor and outdoor environments. It provides 3D gaze annotations covering a wide range of head poses, viewing directions, distances, lighting conditions, and backgrounds. We use the released face crops and adhere to the official train/validation/test split. Gaze360 is used for in-domain evaluation and acts as the primary source domain for cross-domain experiments.
- **MPIIFaceGaze** (Zhang et al., 2017). MPIIFaceGaze includes 45,000 normalized full-face images from 15 participants, with 3,000 samples per subject. Images are collected during routine laptop use and exhibit natural variations in illumination, appearance, and head pose. We adopt 15-fold leave-one-person-out cross-validation for in-domain evaluation. For cross-domain tests, MPIIFaceGaze serves as a target domain.
- **EyeDiap** (Funes Mora et al., 2014). EyeDiap contains 94 RGB and RGB-D video sequences recorded from 16 participants. Recordings cover diverse gaze targets under static and moving head motion within a laboratory setting. We perform four-fold cross-validation for in-domain evaluation following (Cheng et al., 2024a). EyeDiap is used as a target domain in cross-domain experiments.
- **IVGaze** (Cheng et al., 2024b). IVGaze comprises 44,705 in-vehicle images from 125 subjects. It covers diverse gaze and head poses under variable outdoor illumination, with appearance variations including glasses, masks, their combinations, and sunglasses. We follow its three predefined folds for in-domain cross-validation. The accessory annotations are reserved to validate the pseudo-labels adopted in semantic identification and to report gaze errors grouped by accessory types.
- **GazeGene** (Bao et al., 2025). GazeGene is a synthetic dataset with over one million images of 56 virtual subjects captured from nine viewpoints. Each subject is rendered with variations in head pose, gaze direction, facial expression, blinking, pupil size, and illumination. Fine-grained 3D eyeball annotations are also provided. Following the official split, we use Subjects 1–46 for training and Subjects 47–56 for testing. GazeGene is evaluated under the in-domain setting and also serves as an alternative source domain for cross-domain experiments.

Evaluation Metrics. For in-domain evaluation, each model is trained and tested using the dataset-specific splits defined above. For cross-domain evaluation, models are trained on Gaze360 or GazeGene and directly evaluated on MPIIFaceGaze and EyeDiap without target-domain fine-tuning. We report the Mean Angular Error (MAE), measured in degrees, between predicted and ground-truth gaze vectors. For IVGaze, we additionally report the fraction of samples below fixed angular-error thresholds and the mean error stratified by accessory conditions.

B.2 Implementation Details

We use CLIP-B/16 and CLIP-L/14 with pretrained weights released by OpenAI, and DINOv2-B/14 and DINOv2-L/14 with pretrained weights released by Meta. Face images are resized to 224×224; CLIP backbones use CLIP normalization, while DINOv2 uses ImageNet normalization. The image feature from each encoder is ℓ_2-normalized and fed into a two-layer gaze regressor to predict yaw and pitch. For DINOv2 experiments, a frozen CLIP model provides gaze-irrelevant pseudo-labels via zero-shot image-text matching. The gaze loss, gaze-irrelevant losses, gradient rankings, and spectral bases are all computed on top of DINOv2. During inference, only the adapted image encoder and gaze regressor are used.

SSC-LoRA applies low-rank adaptation to Q, V, attention output projection, and both FFN linear layers across all Transformer blocks. For CLIP, these correspond to Q and V within `attn.in_proj_weight`, `attn.out_proj`, `mlp.c_fc`, and `mlp.c_proj`. For DINOv2, they are Q and V in `attn.qkv`, `attn.proj`, `mlp.fc1`, and `mlp.fc2`. For all parameter-efficient fine-tuning (PEFT) baselines, the pre-trained encoder weights are kept frozen, and only the gaze regressor and method-specific parameters are optimized.

Table 5 summarizes the hyperparameters for CLIP and DINOv2. We adopt AdamW (Loshchilov & Hutter, 2017) with weight decay 10^{-4} as the optimizer for both backbones. For a given backbone, all PEFT baselines share identical gaze regressor architecture, training schedule, batch size, and optimizer configurations. Experiments use random seed 42 and report results from the final training epoch.

Table 5: SSC-LoRA hyperparameters for CLIP and DINOv2. Base and Large denote backbone sizes. The regressor learning rate applies to its pre-training stage before spectral scoring.

Hyperparameter	CLIP-B/16 & DINOv2-B/14	CLIP-L/14 & DINOv2-L/14
Regressor learning rate	10^{-3}	10^{-3}
Fine-tuning learning rate	10^{-4}	10^{-4}
Fine-tuning epochs	30	30
Fine-tuning batch size	128	32
Adapter rank r	32	32
LoRA scaling α	64	64
Selection size k	32	32
Adapter dropout	0	0
SCL temperature τ	0.07	0.07
Spectral scoring batches	200	200

Fine-tuning starts with a newly initialized gaze regressor. We apply Kaiming-uniform initialization (He et al., 2015) to matrix A and zero initialization to matrix B. This setup ensures the initial adapted weight matrix equals W_0. All experiments are run on a single RTX 4090 GPU.

B.3 Text Prompt Design and Pseudo-Label Generation

We use prompt groups for glasses, illumination, gender, and age across all datasets except IVGaze. For IVGaze, we substitute the illumination prompt group with mask prompts to align with its accessory variations. All prompts are fixed prior to training and are listed below.

- **Glasses.** *a close-up face of a person* ***wearing glasses*** and *a close-up face of a person* ***without glasses***.
- **Mask.** *a close-up face of a person* ***wearing a mask*** and *a close-up face of a person* ***without a mask***.
- **Illumination.** *a close-up face photo under* ***normal illumination***, *a close-up face photo under* ***dark illumination***, and *a close-up face photo under* ***overexposed illumination***.
- **Gender.** *a photo of a* ***female*** and *a photo of a* ***male***.
- **Age.** *a close-up face of a* ***young adult***, *a close-up face of a* ***middle-aged adult***, and *a close-up face of an* ***elderly person***.

Figure 6 visualizes the pseudo-label matching pipeline. Let E_v and E_t denote the frozen CLIP vision and text encoders, respectively. For each gaze-irrelevant factor f, let t_c^f represent the prompt corresponding to class $c \in \{1, \ldots, C_f\}$, where C_f is the total number of classes for factor f. Given a face image x_j, we assign the class with maximum cosine similarity between the image feature and text feature (Radford et al., 2021):

$$\hat{y}_j^f = \underset{c \in \{1,\ldots,C_f\}}{\arg\max} \frac{E_v(x_j)^\top E_t(t_c^f)}{\|E_v(x_j)\|_2 \, \|E_t(t_c^f)\|_2}. \tag{22}$$

This assignment runs independently for every factor, yielding one categorical pseudo-label per factor for each input image.

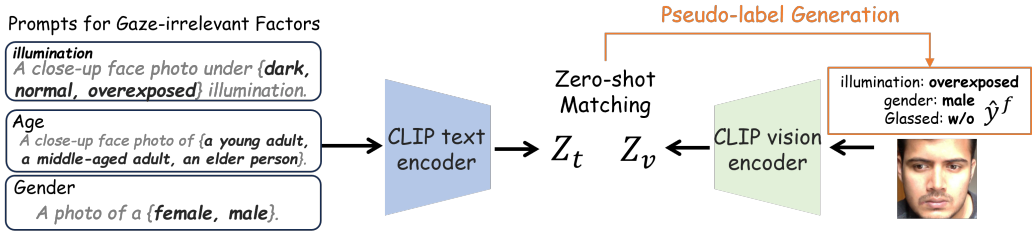


Figure 6: Pseudo-label generation using CLIP. Each image is matched against the class prompts within each gaze-irrelevant factor, and the class with the highest cosine similarity is selected.

B.4 Supervised Contrastive Loss for Spectral Scoring

For each gaze-irrelevant factor f, samples sharing the identical pseudo-label constitute positive pairs. We adopt these pairs to compute the supervised contrastive loss (Khosla et al., 2020):

$$\mathcal{L}_f = \mathcal{L}_{\text{SCL}} = -\frac{1}{|\mathcal{V}_f|} \sum_{j \in \mathcal{V}_f} \frac{1}{|\mathcal{P}_f(j)|} \sum_{p \in \mathcal{P}_f(j)} \log \frac{\exp(z_j^\top z_p/\tau)}{\sum_{j' \in \mathcal{B} \setminus \{j\}} \exp(z_j^\top z_{j'}/\tau)}, \tag{23}$$

where $\mathcal{B}$ denotes the batch. $\mathcal{P}_f(j)$ collects other batch samples that share the same pseudo-label as j for factor f, and $\mathcal{V}_f$ only includes samples with at least one positive counterpart. The vector z_j represents the normalized encoder feature, and τ is the temperature hyperparameter.

C More Ablation Analysis

This section presents additional results on IVGaze, evaluates supervision and component-selection designs for semantic identification, and investigates the range of adapted blocks and feature distributions. The supervision and selection ablations are conducted with CLIP, whereas IVGaze results and feature visualizations are reported for both CLIP and DINOv2.

C.1 Comparison with Representative PEFT Methods

Baselines. We further compare against widely used parameter-efficient fine-tuning (PEFT) baselines: LoRA (Hu et al., 2021), LoRA+ (Hayou et al., 2024), DoRA (Liu et al., 2024), PiSSA (Meng et al., 2024), VeRA (Kopiczko et al., 2024), AdaptFormer (Chen et al., 2022), VPT-Deep (Jia et al., 2022), and linear probing with a frozen encoder.

Quantitative Results. Table 6 compares SSC-LoRA against representative PEFT baselines across both in-domain and cross-domain settings. SSC-LoRA achieves the lowest error across all three evaluation settings for both CLIP backbones. For CLIP-L/14, it reduces LoRA errors from $10.38°$ to $10.21°$ on Gaze360, from $6.92°$ to $5.92°$ on MPIIFaceGaze, and from $7.26°$ to $6.77°$ on EyeDiap. Using DINOv2-L/14, SSC-LoRA also outperforms vanilla LoRA in all three settings, yielding error reductions of $0.05°$, $0.37°$, and $0.45°$ respectively, and attains the second-best performance among all competitors. For DINOv2-B/14, SSC-LoRA obtains the lowest error on EyeDiap. Furthermore, SSC-LoRA delivers stronger generalization with identical parameter budgets relative to most competing approaches. These comparisons demonstrate that SSC-LoRA adapts VFMs for gaze estimation more effectively by suppressing gaze-irrelevant weight updates and amplifying gaze-relevant updates.

C.2 Fine-Grained Evaluation on IVGaze

Table 7 reports the mean angular error, the fraction of samples below $2°$, $4°$, $6°$, and $8°$, and per-accessory errors across IVGaze partitions. All results are obtained under the three-fold evaluation protocol.

Baselines. We compare SSC-LoRA against FullFace (Zhang et al., 2017), DWG (Ghosh et al., 2021), Gaze360 (Kellnhofer et al., 2019), FullFace$^+$, GazeTR (Cheng & Lu, 2022), XGaze (Zhang

Table 6: PEFT comparison with CLIP and DINOv2 backbones. Models are trained on Gaze360 and evaluated in-domain and on MPIIFaceGaze ($\mathcal{D}_{360} \rightarrow \mathcal{D}_M$) and EyeDiap ($\mathcal{D}_{360} \rightarrow \mathcal{D}_D$). We report MAE in degrees and total trainable parameters, including the gaze regressor, in millions (M). The best and second-best errors within each backbone are marked in **bold** and underlined.

Setting	Linear probing	LoRA [ICLR22]	LoRA+ [ICML24]	DoRA [ICML24]	PiSSA [NeurIPS24]	VeRA [ICLR24]	AdaptFormer [NeurIPS22]	VPT-Deep [ECCV22]	SSC-LoRA (Ours)
CLIP-B/16									
Gaze360	30.40	11.02	11.04	11.09	11.07	13.76	12.23	13.03	**10.43**
$\mathcal{D}_{360} \rightarrow \mathcal{D}_M$	40.20	7.17	7.88	7.08	7.49	13.38	8.40	10.84	**6.83**
$\mathcal{D}_{360} \rightarrow \mathcal{D}_D$	27.04	8.39	8.71	7.84	7.90	13.98	7.80	10.17	**7.15**
params. (M)	0.26	4.98	4.98	5.06	4.98	0.34	1.45	0.36	4.98
CLIP-L/14									
Gaze360	28.60	10.38	22.47	10.40	10.60	12.47	11.39	12.31	**10.21**
$\mathcal{D}_{360} \rightarrow \mathcal{D}_M$	27.30	6.92	28.02	7.65	7.71	9.99	8.56	8.84	**5.92**
$\mathcal{D}_{360} \rightarrow \mathcal{D}_D$	26.94	7.26	25.95	6.84	7.38	9.54	8.18	8.14	**6.77**
params. (M)	0.39	12.98	12.98	13.17	12.98	0.60	3.57	0.64	12.98
DINOv2-B/14									
Gaze360	20.05	10.70	10.88	10.55	**10.35**	13.10	11.84	12.78	10.64
$\mathcal{D}_{360} \rightarrow \mathcal{D}_M$	16.09	7.04	8.57	7.12	**6.67**	13.94	8.82	11.05	7.06
$\mathcal{D}_{360} \rightarrow \mathcal{D}_D$	15.41	7.20	8.34	7.82	7.72	10.85	8.86	9.79	**7.16**
params. (M)	0.40	5.11	5.11	5.19	5.11	0.47	1.58	0.49	5.11
DINOv2-L/14									
Gaze360	23.72	10.36	11.69	**10.05**	10.45	12.32	11.24	11.43	10.31
$\mathcal{D}_{360} \rightarrow \mathcal{D}_M$	16.73	7.02	9.36	**6.59**	6.85	12.88	8.45	8.22	6.65
$\mathcal{D}_{360} \rightarrow \mathcal{D}_D$	21.00	7.17	8.65	**6.68**	7.41	9.36	7.40	7.38	6.72
params. (M)	0.53	13.11	13.11	13.31	13.11	0.73	3.70	0.77	13.11

et al., 2020), GazePTR and GazeDPTR (Cheng et al., 2024b). FullFace$^+$ substitutes the original AlexNet backbone of FullFace with ResNet-18. Baseline numbers in Table 7 are directly quoted from Tables 2 and 3 in the original IVGaze paper (Cheng et al., 2024b).

Table 7: Gaze estimation on IVGaze under overall and accessory-specific metrics. Mean and accessory-specific errors are reported in degrees. AP denotes the percentage of samples below each error threshold. The best and second-best results are marked in **bold** and underlined.

Method	Overall performance					Accessory-specific MAE				
	Mean	AP (%) $< 2^\circ$	AP (%) $< 4^\circ$	AP (%) $< 6^\circ$	AP (%) $< 8^\circ$	w/ Glasses	w/o Glasses	w/ Mask	w/o Mask	w/ Sunglasses
FullFace [CVPRW17]	13.67	2.3	8.8	17.8	28.0	14.43	12.40	15.20	13.35	21.39
DWG [ICCVW21]	8.82	6.6	21.7	38.1	53.2	9.20	8.19	9.43	8.69	17.43
Gaze360 [ICCV19]	8.15	9.2	27.3	44.6	58.9	8.30	7.91	8.95	7.99	17.99
FullFace$^+$ [CVPRW17]	7.48	14.2	31.1	46.7	63.1	7.59	7.30	8.37	7.30	16.50
GazeTR [ICPR22]	7.33	17.0	32.8	47.5	64.7	7.40	7.22	8.12	7.17	17.49
XGaze [ECCV20]	7.06	11.7	32.7	51.5	66.7	7.07	7.03	7.80	6.90	15.15
GazePTR [CVPR24]	7.04	17.6	34.0	49.3	66.7	7.13	6.90	7.78	6.89	16.54
GazeDPTR [CVPR24]	6.71	**22.1**	36.0	50.3	68.4	6.77	6.63	7.44	6.57	16.41
SSC-LoRA w/ CLIP-B/16	7.22	12.1	32.8	51.1	65.6	7.45	6.84	7.10	7.23	15.06
SSC-LoRA w/ CLIP-L/14	**6.61**	17.3	**38.4**	55.5	69.6	6.72	**6.44**	**6.31**	6.64	**14.70**
SSC-LoRA w/ DINOv2-B/14	7.22	11.3	31.9	50.9	65.5	7.29	7.10	7.68	7.10	16.04
SSC-LoRA w/ DINOv2-L/14	6.66	15.9	37.9	**55.7**	**70.1**	**6.67**	6.64	7.06	**6.56**	14.71

Overall Performance across Error Thresholds. Table 7 demonstrates that SSC-LoRA built upon CLIP-L/14 achieves the lowest mean angular error among all listed methods and surpasses GazeDPTR at the 4°, 6°, and 8° thresholds. GazeDPTR maintains the highest fraction of samples below 2°, whereas SSC-LoRA with DINOv2-L/14 yields the highest fraction at 6° and 8°. These

observations reveal performance disparities across precision thresholds: CLIP-L/14 delivers the best overall mean error, while DINOv2-L/14 places more samples under the two larger error thresholds.

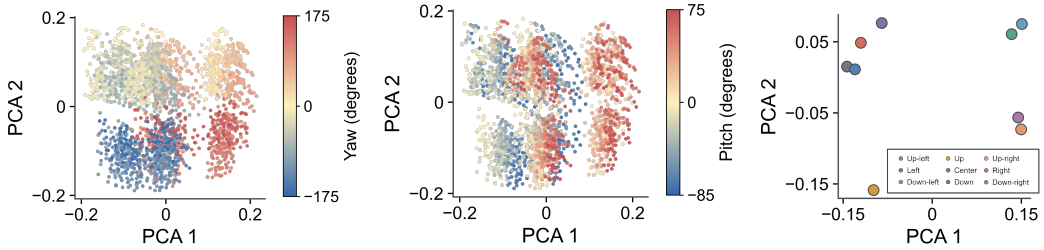
Accessory-Specific Performance. SSC-LoRA using CLIP-L/14 outperforms gaze estimation baselines on four out of five accessory partitions in Table 7. Among all compared approaches, CLIP-L/14 achieves the lowest error for the no-glasses, mask, and sunglasses subsets, while DINOv2-L/14 obtains the best results for glasses-wearing and mask-free partitions. Hence, the optimal choice of encoder depends on the accessory condition. Sunglasses constitute the most challenging setting across all four SSC-LoRA backbones, indicating that eye occlusion remains a persistent bottleneck for gaze estimation.

C.3 Gaze Prompt for Semantic Identification

Gaze-Prompt Alternatives. We investigate how the gaze prompt adopted for spectral scoring influences downstream adaptation. With CLIP-B/16 as the backbone, we compare gradients derived from the trained gaze regressor against gradients obtained from frozen CLIP image–text similarity using two families of gaze prompts:

- **Numerical Prompts.** Each prompt encodes a yaw–pitch pair, e.g., *a face looking at yaw **10** degrees and pitch **-5** degrees*. We construct 2,343 prompts spanning the gaze range of Gaze360.
- **Directional Prompts.** Each prompt describes one of nine coarse gaze directions, such as *a close-up face with eyes looking upward* or *a close-up face with eyes looking down and to the right*.

Numerical Prompts explicitly encode yaw–pitch values, while Directional Prompts define coarse gaze categories. Following the identical pipeline adopted for gaze-irrelevant factors, we compute the supervised contrastive loss and leverage its gradients to score gaze-relevant spectral components.



(a) Numerical prompts colored by yaw. (b) Numerical prompts colored by pitch. (c) Directional prompts.

Figure 7: PCA of CLIP text features used as alternative gaze supervision. Numerical prompts are colored by yaw in (a) and pitch in (b). (c) Text features of the nine directional prompts.

Prompt Text-Feature Visualization. Figure 7 visualizes features for all variants. Numerical Prompts show local color variations with yaw and pitch, yet their projected features form disjoint clusters lacking a clear global ordering with respect to gaze angles. Directional Prompts occupy separate regions in feature space, and their arrangement does not faithfully reflect relative gaze directions. We evaluate the suitability of each supervision strategy via the downstream comparisons presented below.

Table 8: Effect of the gaze supervision used for gradient-based selection. The gaze regressor row is the CLIP-B/16 main configuration from Table 6. All errors are reported in degrees.

Gaze supervision	Gaze360	$\mathcal{D}_{360} \rightarrow \mathcal{D}_M$	$\mathcal{D}_{360} \rightarrow \mathcal{D}_D$	Average
Numerical Prompts	10.71	6.94	7.86	8.50
Directional Prompts	10.92	7.38	8.08	8.79
Gaze Regressor	**10.43**	**6.83**	**7.15**	**8.14**

Downstream Performance Favors the Gaze Regressor. As shown in Table 8, the gaze regressor yields the lowest error under both in-domain and cross-domain settings. Furthermore, Numerical Prompts consistently outperform Directional Prompts across all benchmarks. This indicates that direct regression supervision from continuous yaw and pitch labels provides more informative gradients for spectral component selection compared to discrete text prompt supervision. These downstream results validate that regression-based gradients better identify reliable gaze-relevant spectral components for adaptation.

C.4 Accuracy of Gaze-Irrelevant Pseudo-Label Generation

To verify whether CLIP zero-shot image–text matching yields reliable accessory pseudo-labels for semantic identification, we evaluate CLIP-B/16 and CLIP-L/14 on IVGaze, where ground-truth accessory annotations are available. As shown in Table 9, CLIP-L/14 achieves higher pseudo-label accuracy under most conditions, especially for samples with both glasses and masks. In contrast, CLIP-B/16 performs better on samples with glasses only. These findings validate that CLIP-derived accessory pseudo-labels can serve as supervision for semantic identification, eliminating the need for manually annotated accessory labels.

Table 9: Accuracy (%) of CLIP zero-shot accessory pseudo-labels on IVGaze.

Backbone	Sunglasses	Mask	Glasses	Glasses & Mask	Without Accessory
CLIP-B/16	100.00%	94.73%	97.70%	79.04%	81.74%
CLIP-L/14	100.00%	95.33%	92.13%	94.03%	83.06%

C.5 Sensitivity to Selection Size

The selection size k defines the number of candidate spectral components retrieved from both gaze-relevant and gaze-irrelevant rankings prior to removing overlapping indices. Table 10 compares different k values using CLIP-B/16 trained on Gaze360; our main experiments adopt $k = 32$. Increasing k from 16 to 32 reduces error across all three evaluation settings. However, a larger selection size does not yield consistent performance gains. $k = 128$ achieves the lowest error on MPIIFaceGaze, while $k = 32$ delivers the best results on Gaze360 and EyeDiap. $k = 256$ underperforms $k = 128$ in all three settings. Combined with the concentrated sensitivity profiles shown in Figures 4a and 4b, these findings indicate that effective subspace control hinges on the quality of selected spectral components, instead of merely increasing the quantity of selected components.

Table 10: Sensitivity to the number k of candidate spectral components selected from each ranking before overlap removal.

Top-k	Gaze360	$\mathcal{D}_{360} \to \mathcal{D}_M$	$\mathcal{D}_{360} \to \mathcal{D}_D$
16	11.02	7.03	7.83
32	**10.43**	<u>6.83</u>	**7.15**
128	<u>10.99</u>	**6.58**	<u>7.16</u>
256	11.34	6.86	7.30

C.6 Layer-Wise Analysis

Motivated by the discussion in Section 3.3, we investigate how the separation between gaze-relevant and gaze-irrelevant spectral selections changes across network depth. We further study how the range of adapted Transformer layers impacts in-domain accuracy and cross-domain generalization. Both analyses are conducted with CLIP-B/16 trained on Gaze360.

Spectral Selections Become More Distinct with Depth. Figure 8 demonstrates that the Top-32 gaze-relevant and gaze-irrelevant spectral selections exhibit larger overlap in shallow layers and gradually become more distinct in deeper layers. This trend aligns with the interpretation in Section 3.3: shallow layers encode generic features shared across different objectives, whereas deeper

layers learn more task-specialized representations. The magnitude of this trend differs across projection modules: Q and K maintain relatively high overlap, while V and `attn.out_proj` yield a more substantial overlap reduction. K is included for comparison but is not adapted in our method. Since SSC-LoRA removes overlapping indices from both control sets, reduced overlap preserves more selected components for independent suppression and enhancement. This observation motivates us to investigate whether limiting adaptation to later layers improves cross-domain generalization.

Effect of the Adapted Layer Range. Table 11 compares three layer adaptation strategies. Adapting layers 3–12 reduces the MPIIFaceGaze error from $6.83°$ to $6.41°$ and slightly decreases the EyeDiap error from $7.15°$ to $7.13°$, at the cost of increased Gaze360 error from $10.43°$ to $11.28°$. Further restricting adaptation to layers 6–12 raises errors across all three evaluation settings. These findings indicate that discarding the earliest layers can benefit cross-domain transfer, yet a broader adaptation range remains critical for preserving in-domain accuracy. The selection of adapted layers thus balances two factors: the separation degree of spectral selections and the available adaptation capacity. Our main configuration uses all Transformer layers; it attains the best in-domain performance while maintaining competitive cross-domain generalization.

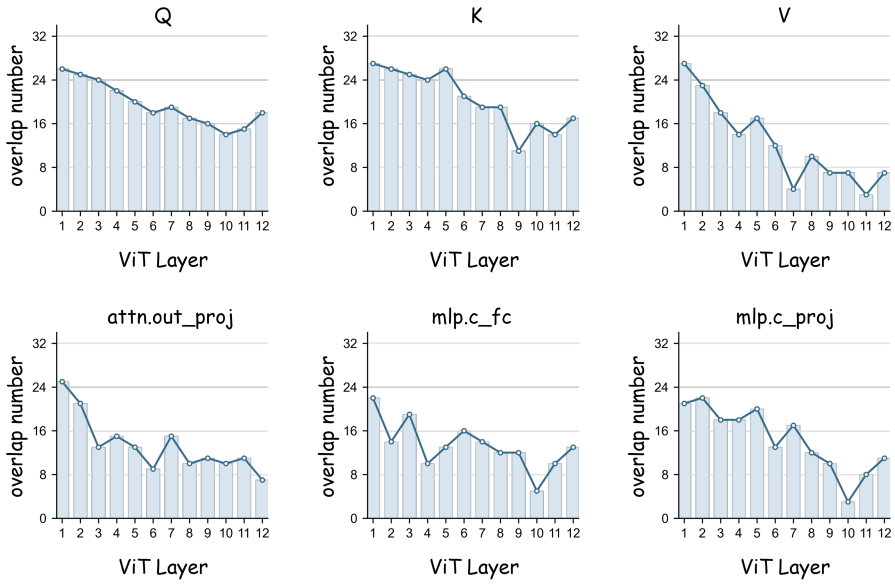


Figure 8: Layer-wise overlap between the gaze-relevant and -irrelevant Top-32 rankings. Bars show shared spectral components and lines trace their depth-wise trend. Q, K, and V denote the attention projections, and `out_proj` denotes the attention output projection. `c_fc` and `c_proj` denote the two FFN projections.

Table 11: Effect of the adapted layer range using CLIP-B/16. Gaze360 denotes in-domain evaluation, while $\mathcal{D}_{360} \rightarrow \mathcal{D}_M$ and $\mathcal{D}_{360} \rightarrow \mathcal{D}_D$ denote cross-domain evaluation on MPIIFaceGaze and EyeDiap. Errors are reported in degrees.

Adapted layers	Gaze360	$\mathcal{D}_{360} \rightarrow \mathcal{D}_M$	$\mathcal{D}_{360} \rightarrow \mathcal{D}_D$
1–12	**10.43**	6.83	7.15
3–12	11.28	**6.41**	**7.13**
6–12	12.63	7.29	7.36

C.7 Analysis of Spectral Update Control

We analyze how spectral control reshapes the allocation of weight updates within the first training epoch for CLIP-B/16. We compare vanilla LoRA, suppression-only updates (Projection in Figure 9), and full SSC-LoRA under identical fixed spectral selections.

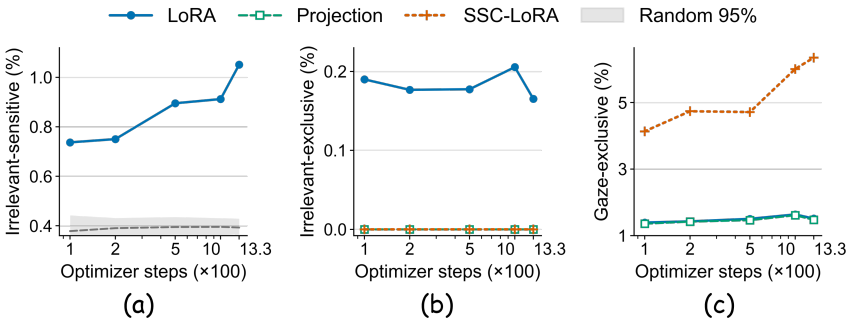


Figure 9: Spectral update allocation during the first training epoch. Vertical axes show projected update-norm ratios (%): (a) irrelevant-sensitive components before overlap removal, (b) irrelevant-exclusive components, and (c) gaze-exclusive components. (a) Vanilla LoRA allocates a larger relative update norm to gaze-irrelevant components than to random spectral selections of the same size. (b) Both suppression-only control and SSC-LoRA reduce updates in the selected gaze-irrelevant block to nearly zero. (c) SSC-LoRA increases the relative update norm in the selected gaze-relevant block compared with vanilla LoRA and suppression-only control. The dashed line and shaded band in (a) show the median and central 95% range over 200 random selections.

For two adjacent checkpoints t^- and t, let $\delta W_m^{(t)} = W_{\mathrm{ft},m}^{(t)} - W_{\mathrm{ft},m}^{(t^-)}$ denote the actual weight update of module m. Let $P_{\mathcal{I},m}^L$ and $P_{\mathcal{I},m}^R$ be the projectors associated with spectral selection $\mathcal{I}$ for module m, following Definition 3.2. We define the projected update-norm ratio as

$$R_{\mathcal{I}}^{(t)} = \left(\frac{\sum_m \left\| P_{\mathcal{I},m}^L \, \delta W_m^{(t)} \, P_{\mathcal{I},m}^R \right\|_F^2}{\sum_m \left\| \delta W_m^{(t)} \right\|_F^2} \right)^{1/2}, \tag{24}$$

which quantifies the relative magnitude of the fourth block from Definition 3.3, aggregated across all analyzed modules. Figure 9(a), (b), and (c) show $R_{\mathcal{I}_n}^{(t)}$, $R_{\hat{\mathcal{I}}_n}^{(t)}$, and $R_{\hat{\mathcal{I}}_g}^{(t)}$ as percentages. The first is the gaze-irrelevant selection before overlap removal. The latter two are the gaze-irrelevant and gaze-relevant selections after shared indices are removed. The random controls in Figure 9(a) sample spectral components outside $\mathcal{I}_n$ and match its size within each module.

LoRA Update Preference. In Figure 9(a), vanilla LoRA yields a persistently higher update-norm ratio for gaze-irrelevant-sensitive spectral components compared with an equal number of randomly sampled spectral components. This preference emerges even when training is supervised solely by gaze labels. This finding demonstrates that optimizing the gaze objective inevitably modifies mappings associated with gaze-irrelevant components, which motivates explicit control over such weight updates.

Effect of Gaze-Irrelevant Suppression. Figure 9(b) shows that both suppression-only updates and full SSC-LoRA reduce the update-norm ratio within the selected gaze-irrelevant block to nearly zero, while vanilla LoRA maintains a non-negligible residual component. This observation aligns with the exclusion property described in Eq. 21. Incorporating gaze enhancement preserves this suppression effect, validating the compatibility between the two control branches.

Effect of Gaze-Relevant Enhancement. In Figure 9(c), suppression-only updates keep the update-norm ratio of selected gaze-relevant components close to that of vanilla LoRA. Full SSC-LoRA sustains a higher ratio that gradually rises throughout training, indicating an increasing relative contribution from the selected gaze block. Collectively, Figure 9(b) and Figure 9(c) illustrate that SSC-LoRA suppresses updates within the gaze-irrelevant block while assigning a larger fraction of the total update norm to the gaze-relevant block. This mechanism accounts for the cross-domain gains reported in Table 4.

C.8 Feature Visualization of CLIP and DINOv2

Feature Organization by Gaze-Irrelevant Factors. Figure 10 compares CLIP and DINOv2 features computed on identical samples, colored by identity, illumination, accessories, age, and gender. Both encoders form feature clusters correlated with these factors. The illumination and accessory subplots use a single subject, demonstrating that such clustering emerges even within one identity. These visualizations illustrate the inherent coupling between feature organization and gaze-irrelevant variations in both VFMs.

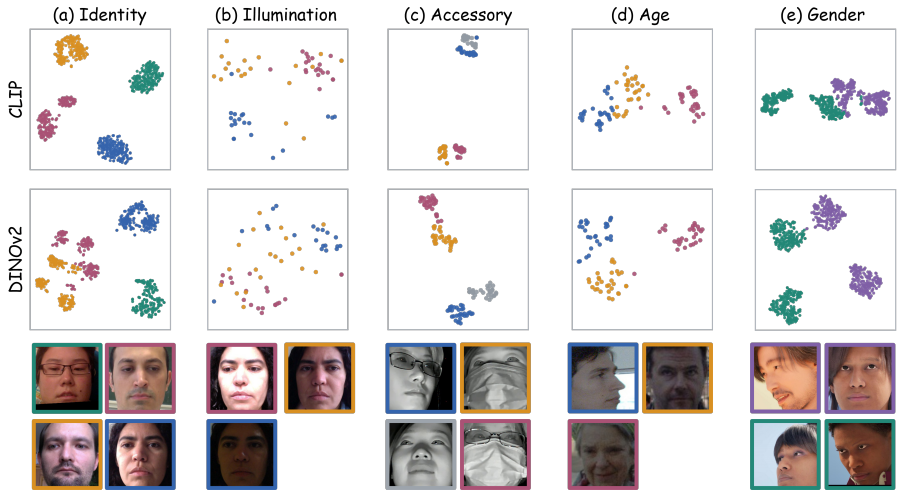


Figure 10: Gaze-irrelevant factors in CLIP and DINOv2 features. (a) Identity and (b) illumination on MPIIFaceGaze. (c) Accessories on IVGaze. (d) Age on Gaze360. (e) Gender on GazeGene. Example images are boxed in the corresponding category colors.

Visualization Analysis. Figure 11 contrasts features learned by vanilla LoRA and SSC-LoRA on MPIIFaceGaze for CLIP-B/16 and DINOv2-B/14. The comparison uses identical samples, highlighted pairs share similar gaze labels but differ substantially in appearance. For both backbones, many highlighted pairs lie closer in feature space under SSC-LoRA than vanilla LoRA. These examples indicate that SSC-LoRA learns more robust representations for samples with comparable gaze directions despite divergent appearance.

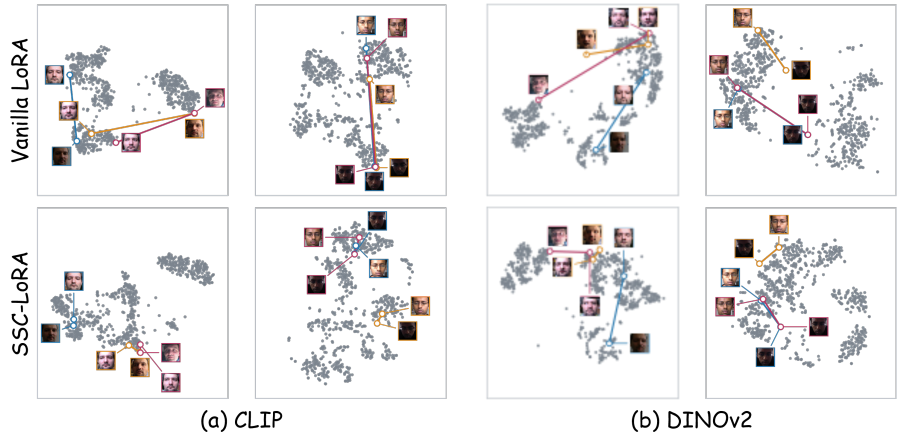


Figure 11: t-SNE feature visualization under $\mathcal{D}_{360} \rightarrow \mathcal{D}_M$ using (a) CLIP and (b) DINOv2. Top and bottom rows show vanilla LoRA and SSC-LoRA, respectively. Colored boxes and connecting lines mark samples with similar gaze labels.