

SPOT: Spectral Preconditioning of Text Embeddings

**On-Device, Privacy-Preserving Calibration of
Vision-Language Models for Fair Human Sensing**

Runkai Zheng

CMU-RI-TR-26-41

May 2026

School of Computer Science
The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Thesis Committee

Fernando De la Torre, Chair
Artur Dubrawski
Yinong Wang

Submitted in partial fulfillment of the requirements for the Degree of
Master of Science in Robotics

Abstract

Vision-language models such as CLIP enable zero-shot recognition by aligning images and natural-language prompts in a shared embedding space. Their appeal is strongest in settings where task-specific training data are scarce: a practitioner can describe classes with prompts, encode them as text axes, and classify images by image-text similarity. However, the same zero-shot pipeline can fail unevenly across subpopulations. Pretraining spans broad and heterogeneous data, and the resulting text axes may entangle intended class semantics with dataset-specific nuisances such as demographic attributes, background context, collection source, or other spurious factors. When these factors correlate with labels, worst-group accuracy can degrade even when average accuracy appears acceptable. This thesis presents SPOT (Spectral Preconditioning of Text Embeddings), a closed-form calibration method for CLIP-like zero-shot classifiers. SPOT estimates the covariance structure of target-domain image embeddings, decomposes the prompt-derived text axes in the corresponding eigenspace, and applies a smooth spectral response that preserves class-relevant bands while suppressing nuisance bands. The resulting calibrated axes replace the original text axes in the standard cosine-scoring pipeline; the image encoder, text encoder, and prompting interface remain frozen. The method is lightweight, data-efficient, label-efficient, and deterministic, reducing adaptation to a small validation search over spectral parameters and class margins. The thesis also introduces a Personal Calibration Protocol (PCP) for on-device, privacy-aware adaptation. PCP models a realistic human-sensing scenario in which a user has an unlabeled image album, task retrievals generated by a fixed reference model, and sparse oracle corrections of that model’s own errors. Unlike standard group-robustness evaluations, PCP uses error corrections as the supervision signal and does not require explicit group labels during adaptation. The PCP-Face-v1 suite instantiates this protocol on face-attribute recognition tasks using CelebA, FairFace, UTKFace, and a shared unlabeled pool. Experiments on standard group-robustness benchmarks show that SPOT improves worst-group accuracy and is competitive with, or better than, existing debiasing methods, particularly in multi-class settings. Under the PCP-Face-v1 protocol, common few-shot and adaptation baselines are often unstable or can fall below the zero-shot reference model, while SPOT improves worst-group accuracy across CelebA, FairFace, and UTKFace and scales with the number of oracle corrections. Finally, the thesis develops a differentially private

version, DP-SPOT, by releasing a noisy calibrated weight matrix with sensitivity bounds derived from spectral Lipschitzness. Together, these results support SPOT as a practical calibration layer for fair, private, and resource-constrained zero-shot recognition.

Keywords: vision-language models, zero-shot learning, fairness, group robustness, differential privacy, personal calibration

Acknowledgments

I am deeply grateful to my advisor, Fernando De la Torre, for guidance, encouragement, and intellectual generosity throughout this work. I thank Artur Dubrawski and Yinong Wang for serving on the thesis committee and for constructive feedback. I also thank Oliver, Sorgan, and the members of the Human Sensing Lab for discussions and support during the development of this thesis. Finally, I thank everyone who supported me over the past two years at Carnegie Mellon University.

Contents

Abstract	ii
Acknowledgments	iv
List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 Motivation: Zero-Shot Recognition and Subpopulation Failure	2
1.2 Why Existing Debiasing Assumptions Are Limiting	3
1.3 Thesis Statement and Contributions	3
1.4 Organization	4
2 Background and Related Work	5
2.1 CLIP Zero-Shot Classification	5
2.2 Bias, Groups, and Worst-Group Accuracy	6
2.3 Debiasing with Group Annotation	6
2.4 Debiasing with Known Bias Directions	7
2.5 Weakly Supervised and Unsupervised Debiasing	7
2.6 Few-Shot Adaptation of CLIP	7
2.7 Positioning of SPOT	8
3 Personal Calibration Protocol	9
3.1 Protocol Motivation	9
3.2 Dataset Abstraction	10
3.3 PCP Rules	12
3.4 PCP-Face-v1 Suite	12
3.5 Why PCP Is Hard for Standard Baselines	12
3.6 Role of SPOT in PCP	13

4	Spectral Preconditioning of Text Embeddings	14
4.1	Preliminaries	14
4.2	Motivation for Data-Aware Calibration	14
4.3	Objective	15
4.4	Closed-Form Solution	16
4.5	Designing the Spectral Response	16
4.6	Scoring and Margins	17
4.7	Why Text-Axis Calibration	17
4.8	Spectral Interpretation	18
4.9	Practical Properties	19
5	Differentially Private SPOT	20
5.1	Differential Privacy Background	20
5.2	Why Use the Non-Centered Second Moment	21
5.3	Sensitivity of the Spectral Map	21
5.4	Analytic Gaussian Mechanism	22
5.5	DP-SPOT Algorithm	23
5.6	Privacy-Utility Observations	23
5.7	Scope of the Privacy Claim	23
6	Experiments	24
6.1	Standard Group-Robustness Setup	24
6.2	Main Results on Standard Benchmarks	25
6.3	Validation Without Group Labels	27
6.4	Data Efficiency and Hyperparameter Smoothness on Standard Benchmarks	27
6.5	PCP-Face-v1: Baseline Failure Under Sparse Corrections	28
6.6	PCP-Face-v1 Main Results	28
6.7	Scaling with Oracle Budget	29
6.8	Per-Attribute Diagnostic on CelebA	29
6.9	Experimental Takeaways	30
7	Discussion and Limitations	31
7.1	Why Spectral Calibration Works	31
7.2	Why PCP Changes the Evaluation Problem	31
7.3	Limitations	32
7.4	Design Implications	32

8 Conclusion	34
A Implementation Details	35
A.1 Prompt Design	35
A.2 Hyperparameter Search	36
A.3 Shared Versus Classwise Spectral Parameters	36
B Additional Theory and Proofs	37
B.1 Convexity and Closed Form	37
B.2 Bijection Between Weight Profile and Shrinkage Profile	37
B.3 Spectral Lipschitz Bound	38
B.4 Global Sensitivity	38
C Additional PCP-Face-v1 Tables	39
C.1 Scaling Data	39
C.2 CelebA Per-Attribute Deltas	39
Bibliography	41

List of Figures

1.1	Overview of the SPOT calibration pipeline	2
3.1	PCP on-device use case	10
3.2	PCP dataset abstraction	11
4.1	Intervention points in CLIP adaptation	18
4.2	Spectral analysis of CelebA attributes	19
5.1	DP-SPOT accuracy versus privacy budget	23
6.1	Data efficiency of SPOT	28
6.2	SPOT parameter sweep	28

List of Tables

2.1	Debiasing supervision taxonomy	8
3.1	PCP-Face-v1 evaluation suite	13
6.1	Standard group-robustness benchmark results	26
6.2	SPOT validation with and without group labels	27
6.3	Baseline failure on UTKFace under PCP	28
6.4	PCP-Face-v1 main results	29
6.5	PCP-Face-v1 scaling with oracle budget	29
6.6	Per-attribute CelebA diagnostic summary	30
C.1	PCP-Face-v1 scaling data	39
C.2	CelebA per-attribute WGA deltas	39

Chapter 1

Introduction

Vision-language models (VLMs) have changed visual recognition by allowing a classifier to be specified with language rather than task-specific training labels. CLIP-style models learn image and text encoders whose outputs lie in a shared feature space; at test time, one can encode prompts such as “a photo of a blond hair person” or “a photo of a black hair person” and classify an image by whichever text embedding has the highest similarity to the image embedding [9, 20]. This zero-shot mechanism is attractive because it requires no new network head, no gradient-based training, and no labeled dataset for every new task.

The central weakness of this mechanism is that the pretrained text axis is treated as if it were already calibrated for the deployment distribution. In practice, downstream feature distributions differ from the heterogeneous pretraining distribution. The mismatch can arise from the collection process, from domain shift, or from correlations between class labels and nuisance attributes. For human-sensing tasks, the risk is particularly acute: face datasets may contain correlations between attributes such as hair color, gender, race, age, or image style. A zero-shot classifier may then score minority or underrepresented groups poorly even when its average accuracy appears high.

This thesis studies a simple question:

Can we correct the decision axes of a CLIP-like zero-shot classifier using only target-domain embedding geometry and sparse validation feedback, without retraining the encoders and without requiring explicit group labels during adaptation?

The answer developed here is SPOT, Spectral Preconditioning of Text Embeddings. SPOT is a post-training calibration method. Given frozen image and text encoders, class prompts, unlabeled target-domain image embeddings, and a small validation set, SPOT estimates the covariance of the target image embeddings and uses its eigenspace as a data-aware coordinate system. Each text axis is then filtered by a smooth spectral response over the eigenvalue spectrum. The calibrated axis remains close to the original prompt-derived axis in class-relevant directions but is allowed

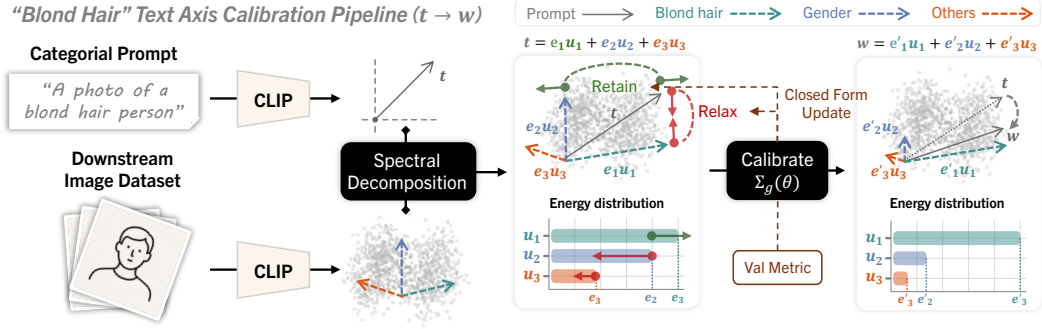


Figure 1.1: Overview of the SPOT calibration pipeline for a prompt-derived text axis. A category prompt and a target image set are encoded by frozen CLIP encoders. The target image embeddings define a covariance spectrum. The original text axis is decomposed in that eigenspace, filtered by a validation-selected spectral response, and reconstructed as a calibrated axis. The update keeps energy in class-relevant bands and attenuates off-band nuisance energy.

to shrink directions that appear as nuisance or confounding variation. The update is closed form.

1.1 Motivation: Zero-Shot Recognition and Subpopulation Failure

In a standard zero-shot CLIP classifier, an image embedding \mathbf{z} and a class text embedding \mathbf{t}_c are ℓ_2 normalized, and the score for class c is

$$s_c(\mathbf{x}) = \langle \mathbf{z}, \mathbf{t}_c \rangle. \quad (1.1)$$

The predicted class is $\hat{y} = \arg \max_c s_c(\mathbf{x})$. This classifier has no task-specific parameters other than the text axes defined by the prompts. That is precisely why it is useful; it is also why it can be brittle. The text axes inherit semantic structure from pretraining, but they are not optimized for the downstream distribution.

Consider a face-attribute task that distinguishes hair color. If blond hair is correlated with female images in a dataset, and black hair is correlated with male images, then the “blond hair” and “black hair” text axes may partially encode gender. The result is not merely random noise: failures concentrate in particular class-attribute cells, such as black-haired female faces. This motivates measuring not only average accuracy but also worst-group accuracy, where a group is a class-attribute pair and

$$\text{WGA} = \min_{g \in \mathcal{G}} \Pr(\hat{y} = y \mid g). \quad (1.2)$$

The slide deck motivating this thesis reports large zero-shot best-worst gaps in face datasets: a 14.0 percentage point gap for a CelebA hair-color example, a 48.6 point gap for a FairFace race example, and a 63.4 point gap for a UTKFace ethnicity example. These gaps indicate that a model can appear reasonable on average while failing specific subpopulations.

1.2 Why Existing Debiasing Assumptions Are Limiting

Debiasing and group-robustness methods typically use one of three supervision regimes. First, group-supervised methods such as GroupDRO, DFR, and FairerCLIP use class-attribute group labels to rebalance, reweight, or directly optimize a group objective [4, 12, 21]. These methods can be powerful, but they assume that practitioners know which bias axes matter and can annotate them reliably. Second, pseudo-group and weakly supervised methods infer groups from losses, biased classifiers, confidence cues, or residual representations [15, 18, 19, 25, 28]. These methods reduce annotation requirements but depend on proxy signals that can become fragile when biases are multiple, overlapping, or combinatorial. Third, language-guided projection methods use text-specified or language-model-generated bias directions [1, 3, 8]; they require bias directions to be stated or discovered in a way that remains semantically faithful.

The intended deployment setting in this thesis is stricter. A user or device may only be able to provide a small number of corrections to the model’s own errors. The user should not need to name sensitive attributes, supply group labels, or expose the full image album to a server. This setting motivates both the SPOT method and the Personal Calibration Protocol introduced in [chapter 3](#).

1.3 Thesis Statement and Contributions

The thesis statement is as follows:

A zero-shot vision-language classifier can be made more robust and more suitable for private on-device adaptation by spectrally calibrating its text axes in the covariance eigenspace of target-domain images, using sparse validation feedback rather than explicit group annotations.

The thesis makes four contributions.

1. A closed-form spectral calibration method. SPOT transforms the prompt-derived text axes by

$$\mathbf{W}^* = \mathbf{U} \text{diag}(G(\mathbf{e})) \mathbf{U}^\top \mathbf{T}, \quad (1.3)$$

where $\mathbf{U} \text{diag}(\mathbf{e}) \mathbf{U}^\top$ is the target-domain image covariance and G is a validation-selected spectral response. The method is deterministic and requires no gradient descent.

2. A geometry-aware objective for text-axis recalibration. SPOT is derived from a convex objective that keeps the calibrated axis close to the original text axis under a covariance-induced metric while preferring a small-norm solution. This objective admits a per-eigen-direction shrinkage interpretation and motivates the log-Gaussian spectral response used in experiments.

3. A personal calibration protocol for sparse correction-based adaptation. The thesis introduces PCP, a protocol that uses an unlabeled album, zero-shot task retrievals, and sparse oracle corrections of model errors. PCP-Face-v1 evaluates this protocol on face tasks and highlights a gap between conventional few-shot adaptation and correction-based personalization.

4. Empirical and privacy analysis. Experiments show that SPOT improves worst-group accuracy on standard benchmarks and under PCP-Face-v1. The thesis also develops DP-SPOT by deriving sensitivity bounds for the spectral calibration map and adding Gaussian noise to the released calibrated axes.

1.4 Organization

[Chapter 2](#) reviews zero-shot CLIP classification, group robustness, and prior debiasing methods. [Chapter 3](#) introduces the Personal Calibration Protocol and PCP-Face-v1. [Chapter 4](#) develops SPOT and its closed-form solution. [Chapter 5](#) describes the differential privacy extension. [Chapter 6](#) presents results on standard group-robustness benchmarks and PCP-Face-v1. [Chapter 7](#) discusses limitations and design implications, and [chapter 8](#) concludes.

Chapter 2

Background and Related Work

This chapter summarizes the technical context for SPOT. The goal is not to replace the broad literature on vision-language models or fairness, but to identify the assumptions that are most relevant to the thesis: how zero-shot text axes define decision boundaries, how group robustness is measured, and what supervision prior debiasing methods require.

2.1 CLIP Zero-Shot Classification

Let $f : \mathcal{X} \rightarrow \mathbb{R}^d$ be a frozen image encoder and $h : \mathcal{Y} \rightarrow \mathbb{R}^d$ be a frozen text encoder. Given an image x , the normalized image embedding is

$$\mathbf{z} = \frac{f(x)}{\|f(x)\|_2}. \quad (2.1)$$

Given a class name or natural-language label prompt y_c , the normalized text axis is

$$\mathbf{t}_c = \frac{h(y_c)}{\|h(y_c)\|_2}. \quad (2.2)$$

The zero-shot score for class c is the cosine similarity

$$s_c(x) = \langle \mathbf{z}, \mathbf{t}_c \rangle, \quad (2.3)$$

and the prediction is $\hat{y}(x) = \arg \max_c s_c(x)$. For a binary task, the decision boundary is determined by the difference of two text axes. For a multi-class task, each class prompt contributes one prototype, and the classifier compares the image embedding to all prototypes.

The defining property of this classifier is that it uses no downstream training. The text encoder maps language to a semantic axis, and that axis becomes a classifier. This is useful for flexible human-sensing interfaces: a user can search a personal album for “blond hair”, “smiling”, or “wearing glasses” without training a new model for every query. However, the decision boundary is inherited from pretraining and may not align with the actual target distribution.

2.2 Bias, Groups, and Worst-Group Accuracy

A model is group-robust when it performs well not only on average but also on subpopulations formed by combinations of target labels and attributes. In this thesis, a group is typically a pair

$$g = (y, a), \quad (2.4)$$

where y is the target class and a is an attribute such as gender, background, ethnicity, data source, or another factor used for evaluation. The worst-group accuracy is

$$\text{WGA} = \min_{g \in \mathcal{G}} \frac{1}{|\mathcal{D}_g|} \sum_{(x_i, y_i) \in \mathcal{D}_g} \mathbf{1}\{\hat{y}_i = y_i\}. \quad (2.5)$$

Average accuracy can be high while WGA is low. This mismatch is the core measurement problem in biased recognition: majority groups dominate averages, but deployment risks often concentrate on minority groups.

In the standard group-robustness benchmarks used later in this thesis, groups are defined by known evaluation metadata. CelebA uses hair label and gender, Waterbirds uses bird type and background, BAR uses action class and background, and CIFAR-10.02 uses object class and data source. PCP-Face-v1 uses group metadata only for evaluation; adaptation is driven by sparse corrections rather than group labels.

2.3 Debiasing with Group Annotation

Group-supervised methods use explicit group labels during training or model selection. GroupDRO minimizes the worst-group risk when group identity is known [21]. DFR retrains a final layer on a group-balanced subset and shows that the final classifier can often correct shortcut reliance when the representation is fixed [12]. FairerCLIP adjusts CLIP zero-shot predictions using a debiasing formulation in a reproducing-kernel Hilbert space [4].

These methods are relevant because they show that many robustness failures live in the classifier or decision boundary rather than requiring a new encoder. However, they assume that the relevant groups are enumerated and annotated. In face and personal-album settings, this assumption is restrictive. Attributes can be numerous, overlapping, sensitive, and ambiguous. The number of intersectional groups can grow combinatorially [11]. Moreover, a user may not want to disclose or label sensitive attributes merely to calibrate a local search model.

2.4 Debiasing with Known Bias Directions

A second family assumes that the bias direction is known or can be specified with language. Projection-based methods remove or suppress subspaces associated with biased prompts; examples include OrthCali and RoboShot [1, 3]. SANER neutralizes societal attributes in text representations while attempting to preserve task utility [8]. These approaches are compatible with CLIP because language can express both target semantics and nuisance concepts.

The limitation is that robustness depends on correctly specifying the nuisance direction. If the harmful direction is incomplete, prompt-dependent, or entangled with the useful semantic direction, projection can remove useful information or miss the true bias. In multi-bias settings, the interaction between target semantics and nuisance factors is not always captured by a small set of hand-written descriptors.

2.5 Weakly Supervised and Unsupervised Debiasing

Weakly supervised methods avoid explicit group labels by inferring proxy signals. JTT trains an initial model, identifies high-loss examples, and then retrains with those examples emphasized [15]. Learning from Failure (LfF) trains a biased classifier and uses its failures to guide a debiased classifier [18]. Correct-N-Contrast uses contrastive structure around error-prone examples [25]. Automatic Feature Reweighting (AFR) uses confidence-based weights to improve group robustness [19]. GEORGE clusters representations into subgroups and trains for subclass robustness [22]. PruSC prunes spurious clusters, and PPA projects out class information, probes residuals, and aggregates group-aware weights [14, 28].

These methods are attractive because they reduce annotation requirements, but they inherit the reliability of their proxies. Loss, confidence, clustering, and residual probes may align with the true minority groups on binary single-bias tasks, but they can become unreliable when biases are multiple, overlapping, or unevenly represented. The PCP-Face-v1 results in [chapter 6](#) show that several such methods are unstable under sparse correction-based adaptation.

2.6 Few-Shot Adaptation of CLIP

A related line of work adapts CLIP using small labeled sets. Tip-Adapter builds a cache model from few-shot examples [26]. CLIP-Adapter adds a residual feature

Table 2.1: Representative debiasing and adaptation families, organized by the supervision signal they require. A checkmark means the method family typically requires group labels for training or validation.

Family	Representative methods	Core idea	Train GL	Val GL
Group-supervised adaptation	GroupDRO, DFR, FairerCLIP	Optimize group-balanced or worst-group objectives.	✓	✓
Pseudo-group inference	JTT, LfF, CnC, AFR, PPA	Infer proxy groups, then adapt or reweight.	–	✓
Self-supervised clustering	GEORGE, PruSC	Cluster features and rebalance or prune.	–	–
Language-guided projection	RoboShot, OrthCali, SANER	Remove text-specified bias directions.	–	–
SPOT	This thesis	Filter text axes using target covariance.	–	optional

adapter to image embeddings [7]. WiSE-FT interpolates between zero-shot and fine-tuned weights to preserve robustness [24]. Linear probing and related final-layer updates are standard baselines.

These methods are not designed specifically for fairness or sparse corrections. They often optimize average validation accuracy and can overfit when the labeled set is small or biased. In the thesis setting, the labeled examples are not an independent balanced training set; they are oracle corrections of a reference model’s own errors. This distributional conditioning makes naive few-shot adaptation especially risky.

2.7 Positioning of SPOT

SPOT differs from the methods above in three ways. First, it calibrates text axes rather than training a new feature adapter or final layer. Second, it uses unlabeled target-domain covariance to define a coordinate system and a spectral filter, so adaptation is grounded in deployment geometry. Third, it is closed form and therefore amenable to privacy analysis: the released state is a calibrated matrix rather than a sequence of training updates. These choices make SPOT suitable for the PCP setting, where group labels are unavailable, corrections are sparse, and the calibration state should be lightweight enough to cache on device.

Chapter 3

Personal Calibration Protocol

The original SPOT paper evaluates spectral calibration on standard group-robustness datasets with validation labels and, in some settings, validation group labels. The slide deck extends the thesis toward a more realistic deployment problem: on-device personalization for human sensing. In this setting, the user has a large private album, a reference zero-shot model produces task-specific retrievals, and the user supplies sparse corrections when the model is wrong. This chapter formalizes that setting as the Personal Calibration Protocol (PCP).

3.1 Protocol Motivation

A personal device can run a frozen VLM to search a photo library with language prompts. The user might ask for images of people with a given hair color, race-related attribute, age-related category, or other personal concept. The device can retrieve candidates using zero-shot CLIP scores, but the same bias issues described in [chapter 1](#) can cause systematic errors. A user may correct a handful of mistaken results, but they will not label the entire album or annotate sensitive group membership for every image.

This produces a supervision signal different from standard few-shot learning. The labeled examples are not randomly sampled from the task distribution; they are conditioned on the reference model being wrong. This makes the corrections highly informative about the model’s failure modes, but also sparse and distribution-shifted relative to ordinary training data. PCP evaluates whether an adaptation method can use this signal without requiring explicit group labels. [Figure 3.1](#) grounds this protocol in a product-style workflow: a personal album supplies the target-domain pool, a search prompt induces task-specific retrievals, and the user corrects only the wrong results.

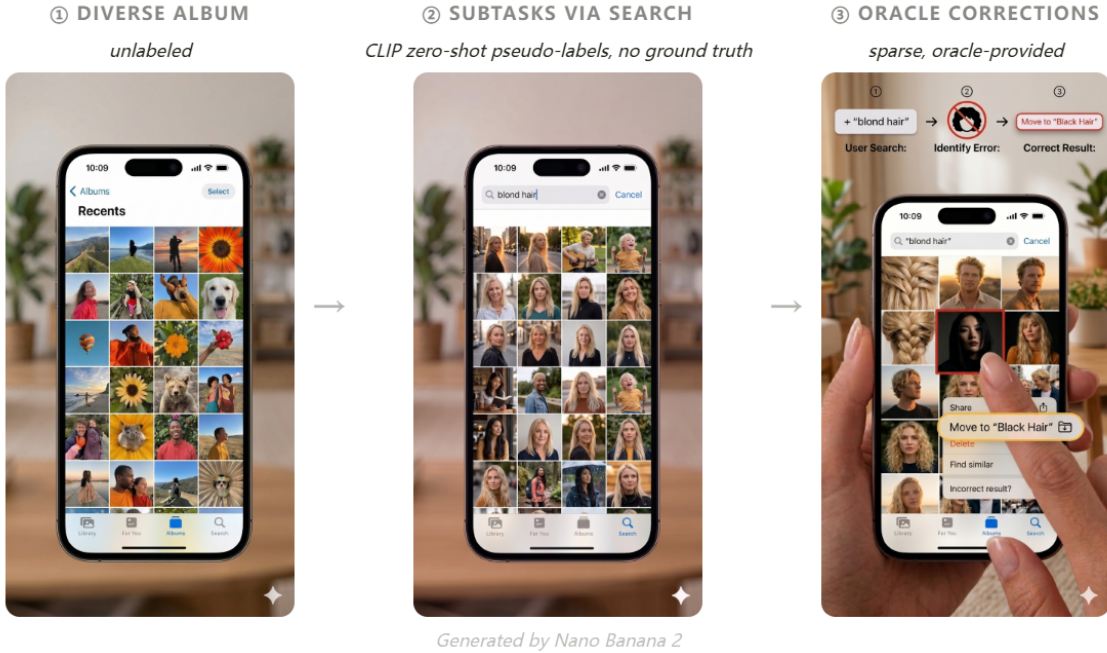


Figure 3.1: PCP as an on-device personalization workflow. A user begins with a large unlabeled album, obtains task-specific retrievals from a frozen zero-shot model, and supplies sparse oracle corrections only for mistaken results. This workflow motivates using error corrections rather than complete class or group annotations as the calibration signal.

3.2 Dataset Abstraction

PCP decomposes the available data into three components:

$$\mathcal{D} = \mathcal{D}_{\text{album}} \cup \mathcal{D}_{\text{task}}^{\text{retr}} \cup \mathcal{D}_{\text{task}}^{\text{user}}. \quad (3.1)$$

The components have different information content and privacy implications.

Unlabeled album. The album is an unlabeled pool

$$\mathcal{D}_{\text{album}} = \{x_i\}_{i=1}^n \sim p_{\text{data}}. \quad (3.2)$$

It represents the user’s full photo library or a large private subset. It is image-only and can be out-of-distribution relative to benchmark validation data. In PCP-Face-v1, this pool is represented by YFCC100M images [23].

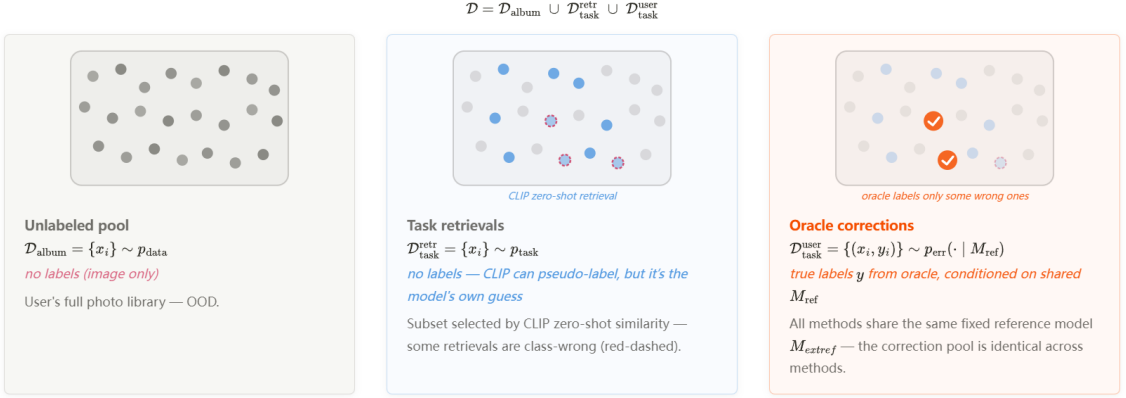


Figure 3.2: Dataset abstraction for PCP. The available information is divided into an unlabeled album $\mathcal{D}_{\text{album}}$, task retrievals $\mathcal{D}_{\text{task}}^{\text{retr}}$, and sparse user corrections $\mathcal{D}_{\text{task}}^{\text{user}}$. The figure emphasizes the asymmetry of the protocol: abundant image-only data define the target geometry, while the scarce oracle corrections provide precise but limited supervision.

Task retrievals. A fixed reference model M_{ref} uses zero-shot similarity to retrieve a task-relevant subset:

$$\mathcal{D}_{\text{task}}^{\text{retr}} = \{x_i\}_{i=1}^{k_{\text{retr}}} \sim p_{\text{task}}. \quad (3.3)$$

These examples are unlabeled. The reference model may assign pseudo-labels, but the pseudo-labels are its own guesses and may be biased or wrong. Thus this set provides task relevance but not trustworthy labels.

Oracle corrections. The sparse labeled signal is

$$\mathcal{D}_{\text{task}}^{\text{user}} = \{(x_i, y_i)\}_{i=1}^{k_{\text{user}}} \sim p_{\text{err}}(\cdot | M_{\text{ref}}). \quad (3.4)$$

These labels are corrections of errors made by the fixed reference model. All methods share the same M_{ref} , so the correction pool is identical across methods. The source slides specify the typical budget relation

$$|\mathcal{D}_{\text{album}}| \gg k_{\text{retr}} > k_{\text{user}}, \quad (3.5)$$

with k_{retr} and k_{user} generally in the range 32–256.

Figure 3.2 also clarifies why PCP is not equivalent to ordinary few-shot learning. The retrievals are selected by M_{ref} , so they are already biased toward the reference model’s view of the task. The corrected examples are even more selective: they come from the error distribution $p_{\text{err}}(\cdot | M_{\text{ref}})$ rather than from the overall task distribution.

A robust calibration method should therefore use the album and retrievals to estimate broad target-domain structure, while using the corrections only as a small validation signal that identifies which adaptation choices actually repair the reference model’s mistakes.

3.3 PCP Rules

A method evaluated under PCP receives the following shared inputs:

1. a fixed reference model M_{ref} ;
2. an unlabeled album $\mathcal{D}_{\text{album}}$;
3. task retrievals $\mathcal{D}_{\text{task}}^{\text{retr}}$ generated by M_{ref} ;
4. oracle corrections $\mathcal{D}_{\text{task}}^{\text{user}}$;
5. class names or prompts for the task.

The method may tune hyperparameters using the correction labels, but it may not use explicit group labels for training or validation. Group labels are reserved for final evaluation of WGA. This restriction is what distinguishes PCP from standard group-robustness benchmarks.

3.4 PCP-Face-v1 Suite

PCP-Face-v1 instantiates the protocol on face recognition tasks. The slide suite specifies three benchmarks and their evaluation group structure, summarized in [table 3.1](#). The suite uses a shared unlabeled album pool and task-specific retrieval/correction sets. The protocol slide lists FairFace as a 7-way race task and UTKFace as a 5-way ethnicity task; the reported PCP-Face-v1 main-results table in the slides uses 4-way variants for FairFace and UTKFace. This thesis preserves that distinction: [table 3.1](#) describes the suite specification, while [table 6.4](#) reports the exact experiments shown in the slides.

3.5 Why PCP Is Hard for Standard Baselines

In PCP, the only true labels are sparse corrections. For a budget such as $k_{\text{user}} = 64$, a method can easily overfit or learn a correction-specific artifact. Methods that assume

Table 3.1: PCP-Face-v1 suite specification from the slide deck. Group labels are used for evaluation, not for adaptation.

Dataset	Task	Evaluation groups	Metric
CelebA [16]	Hair color (4-way)	35 attributes	WGA _m
FairFace [10]	Race (7-way specification)	Gender cells	WGA
UTKFace [27]	Ethnicity (5-way specification)	Gender cells	WGA

balanced few-shot samples may treat the corrections as ordinary labeled examples, but the corrections are drawn from the reference model’s error distribution. Methods that infer groups from losses or clusters may fail because the data are too sparse to reveal stable group structure. Methods that require validation group labels are inapplicable.

The slide deck illustrates this difficulty on UTKFace at $k_{\text{user}} = k_{\text{retr}} = 64$. The zero-shot reference WGA is 32.46. LP++, Tip-Adapter, CLIP-Adapter, JTT, and GEORGE fall below or near the zero-shot baseline, while WiSE-FT reaches 36.8. These results motivate a method that changes the classifier with a low-dimensional, geometry-aware update instead of fitting a high-capacity adapter or cache.

3.6 Role of SPOT in PCP

SPOT is well matched to PCP for three reasons. First, it uses the unlabeled album and retrievals through second-order embedding statistics, which are abundant relative to the correction labels. Second, it uses the corrections only to select a small number of spectral hyperparameters and thresholds, reducing overfitting risk. Third, its calibrated state is just a matrix of text axes and optional margins, making it natural to store locally or release with differential privacy.

Chapter 4

Spectral Preconditioning of Text Embeddings

This chapter develops SPOT as a closed-form calibration of CLIP text axes. The method keeps the encoders frozen and modifies only the axes used for similarity scoring. The derivation follows three steps: estimate target-domain geometry, define a geometry-aware objective for a calibrated axis, and choose a spectral response that selectively preserves useful directions.

4.1 Preliminaries

Let $\mathcal{C} = \{1, \dots, C\}$ denote the class set. For each class c , the frozen text encoder maps a prompt y_c to a normalized text axis $\mathbf{t}_c \in \mathbb{R}^d$. The matrix of text axes is

$$\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_C] \in \mathbb{R}^{d \times C}. \quad (4.1)$$

For target-domain images x_1, \dots, x_n , the frozen image encoder yields normalized embeddings $\mathbf{z}_i \in \mathbb{R}^d$. The empirical mean and centered covariance are

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i, \quad \boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{z}_i - \boldsymbol{\mu})(\mathbf{z}_i - \boldsymbol{\mu})^\top. \quad (4.2)$$

Let the eigendecomposition be

$$\boldsymbol{\Sigma} = \mathbf{U} \text{diag}(\mathbf{e}) \mathbf{U}^\top, \quad (4.3)$$

where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_d]$ is orthogonal and $\mathbf{e} = (e_1, \dots, e_d)$ contains nonnegative eigenvalues. The eigenvectors provide a data-driven basis for target-domain variation.

4.2 Motivation for Data-Aware Calibration

A Euclidean constraint $\|\mathbf{w}_c - \mathbf{t}_c\|_2^2$ treats every orthogonal direction equally. Downstream embedding geometry is not isotropic. Some directions may encode class-relevant attributes, while others encode nuisance factors such as background, illumination, demographic correlations, or data source. A useful calibration should therefore keep

the text axis close to the original along important directions, while allowing more deviation in directions that hurt the validation objective.

To express this idea, define a nonnegative spectral weight profile $g : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$. Applied elementwise to the covariance spectrum, it induces the matrix

$$\Sigma_g = \mathbf{U} \text{diag}(g(\mathbf{e})) \mathbf{U}^\top. \quad (4.4)$$

The geometry-aware deviation is

$$\|\mathbf{w}_c - \mathbf{t}_c\|_{\Sigma_g}^2 = \sum_{j=1}^d g(e_j) \langle \mathbf{u}_j, \mathbf{w}_c - \mathbf{t}_c \rangle^2. \quad (4.5)$$

A large value of $g(e_j)$ forces the calibrated axis to remain close to the original axis along \mathbf{u}_j ; a small value relaxes that constraint.

4.3 Objective

For each class c , SPOT solves

$$\min_{\mathbf{w}_c} \|\mathbf{w}_c - \mathbf{t}_c\|_{\Sigma_g}^2 + \lambda \|\mathbf{w}_c\|_2^2, \quad \lambda > 0. \quad (4.6)$$

The first term preserves geometry-aware consistency with the prompt-derived axis. The second term chooses a small-norm solution among axes that remain close to the original. This small-norm preference is important when the validation signal is sparse: it prevents the calibrated classifier from becoming unnecessarily expressive.

Two limiting cases clarify the objective. If $\Sigma_g = \mathbf{I}$, then the solution is a scalar shrinkage of \mathbf{t}_c , which leaves cosine-based zero-shot predictions unchanged when all class axes are normalized. If $\Sigma_g = \Sigma$, then the geometry term preserves the original in-distribution scores:

$$\|\mathbf{w}_c - \mathbf{t}_c\|_{\Sigma}^2 = \frac{1}{n} \sum_{i=1}^n \langle \mathbf{z}_i - \boldsymbol{\mu}, \mathbf{w}_c - \mathbf{t}_c \rangle^2 \quad (4.7)$$

$$= \frac{1}{n} \left\| \tilde{\mathbf{Z}}^\top \mathbf{w}_c - \tilde{\mathbf{Z}}^\top \mathbf{t}_c \right\|_2^2, \quad (4.8)$$

where $\tilde{\mathbf{Z}} = [\mathbf{z}_1 - \boldsymbol{\mu}, \dots, \mathbf{z}_n - \boldsymbol{\mu}]$. Thus the Mahalanobis choice preserves the scores that the original axis assigns on the target distribution while selecting a smaller-norm axis.

Algorithm 1 SPOT calibration

Require: Image embeddings $\{\mathbf{z}_i\}_{i=1}^n$; text axes $\mathbf{T} \in \mathbb{R}^{d \times C}$; spectral response $G(\cdot)$.

- 1: $\boldsymbol{\mu} \leftarrow \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i$
 - 2: $\tilde{\mathbf{Z}} \leftarrow [\mathbf{z}_i - \boldsymbol{\mu}]_{i=1}^n$
 - 3: $\mathbf{U} \text{diag}(\mathbf{e}) \mathbf{U}^\top \leftarrow \text{PCA}(\tilde{\mathbf{Z}})$
 - 4: $\mathbf{W}^* \leftarrow \mathbf{U} \text{diag}(G(\mathbf{e})) \mathbf{U}^\top \mathbf{T}$
 - 5: **return** \mathbf{W}^*
-

4.4 Closed-Form Solution

The objective in [equation \(4.6\)](#) is strictly convex because $\boldsymbol{\Sigma}_g + \lambda \mathbf{I} \succ 0$. Differentiating gives

$$(\boldsymbol{\Sigma}_g + \lambda \mathbf{I}) \mathbf{w}_c = \boldsymbol{\Sigma}_g \mathbf{t}_c, \quad (4.9)$$

so the unique minimizer is

$$\mathbf{w}_c^* = (\boldsymbol{\Sigma}_g + \lambda \mathbf{I})^{-1} \boldsymbol{\Sigma}_g \mathbf{t}_c. \quad (4.10)$$

In the eigenbasis of $\boldsymbol{\Sigma}$, this is a direction-wise shrinkage:

$$\mathbf{w}_{c,j}^* = G(e_j) \mathbf{t}_{c,j}, \quad G(e_j) = \frac{g(e_j)}{g(e_j) + \lambda}, \quad (4.11)$$

where $\mathbf{t}_{c,j} = \langle \mathbf{u}_j, \mathbf{t}_c \rangle$ and $\mathbf{w}_{c,j}^* = \langle \mathbf{u}_j, \mathbf{w}_c^* \rangle$. Applying the same spectral response to all classes yields the matrix update

$$\mathbf{W}^* = \mathbf{U} \text{diag}(G(\mathbf{e})) \mathbf{U}^\top \mathbf{T} = F(\boldsymbol{\Sigma}) \mathbf{T}. \quad (4.12)$$

4.5 Designing the Spectral Response

It is more convenient to design the shrinkage profile G directly than to design g . For fixed $\lambda > 0$, the mapping

$$G(e) = \frac{g(e)}{g(e) + \lambda} \iff g(e) = \lambda \frac{G(e)}{1 - G(e)} \quad (4.13)$$

is a one-to-one correspondence between nonnegative g and shrinkage values $G(e) \in [0, 1)$. Thus choosing a valid shrinkage profile preserves the convex objective and the minimum-norm interpretation.

The original paper observes that the monotone Mahalanobis response $G(e) = e/(e + \lambda)$ is not uniformly reliable. It assumes that class-relevant information lies in high-variance directions and that nuisance lies in low-variance directions. But spurious factors can appear in either tail of the spectrum. SPOT therefore uses a log-Gaussian band-pass response:

$$G(e) = \exp\left(-\frac{(\log(e + \varepsilon) - m)^2}{2\sigma^2}\right), \quad (4.14)$$

where $\varepsilon > 0$ is a numerical stabilizer. The parameter m controls the center of the passband on the log-eigenvalue axis, and σ controls its bandwidth. A validation set selects m , σ , and optional class margins.

4.6 Scoring and Margins

After calibration, the score for class c becomes

$$s_c(x) = \langle \mathbf{z}, \mathbf{w}_c^* \rangle. \quad (4.15)$$

To correct class-level bias, SPOT also allows class-specific margins τ_c :

$$\hat{y}(x) = \arg \max_{c \in \{1, \dots, C\}} (s_c(x) + \tau_c). \quad (4.16)$$

The margins are selected on validation data. Setting all margins to zero recovers the pure spectral calibration rule.

4.7 Why Text-Axis Calibration

The slide deck compares three intervention points in the CLIP pipeline. CLIP-Adapter modifies image features with a residual MLP; LP++ fine-tunes or learns text-like classifier weights directly; SPOT inserts a spectral filter between the original text embedding and the similarity computation. Figure 4.1 visualizes these alternatives. The key distinction is that SPOT changes the text side through a structured operator $F(\Sigma)$ rather than learning a free-form image residual or directly fitting unconstrained prototypes.

The intervention point in Figure 4.1 explains why SPOT is well suited to PCP. Under sparse oracle corrections, image-side adapters can overfit because they introduce additional trainable capacity on top of the frozen image representation. Direct prototype fitting is lower-dimensional but can still move the class axis without respecting

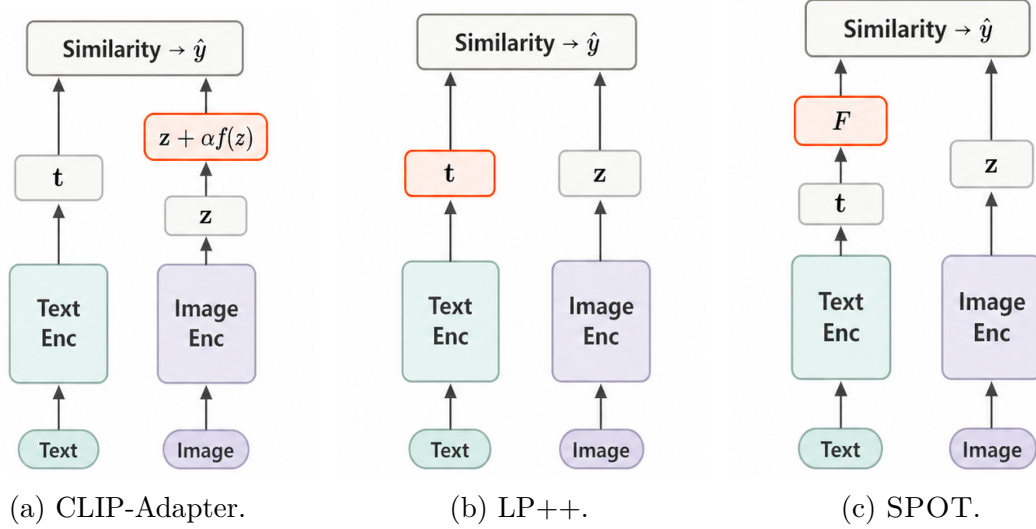


Figure 4.1: Intervention points for CLIP-Adapter, LP++, and SPOT. CLIP-Adapter optimizes an image-side residual, LP++ optimizes the text prototype directly, and SPOT optimizes a spectral filter applied to the text axis before similarity scoring. The highlighted boxes indicate the primary optimization target.

target-domain covariance. SPOT constrains adaptation to a covariance-dependent spectral filter. This preserves the frozen encoders and the prompting interface, avoids gradient-based updates, and produces a single linear transformation shared across classes. Most importantly, it uses the target image distribution to determine which components of the text axes should be retained.

4.8 Spectral Interpretation

The CelebA spectral analysis in the source material maps attribute text axes into the covariance eigenspace. For a text axis \mathbf{t} , define normalized energy over eigen-directions by

$$p_j = \frac{\langle \mathbf{u}_j, \mathbf{t} \rangle^2}{\sum_k \langle \mathbf{u}_k, \mathbf{t} \rangle^2}. \quad (4.17)$$

The spectral centroid and spread on the log-eigenvalue axis are

$$\mu_{\text{spec}} = \sum_j p_j \log(e_j), \quad \sigma_{\text{spec}}^2 = \sum_j p_j (\log(e_j) - \mu_{\text{spec}})^2. \quad (4.18)$$

In the CelebA blond-hair example, the “Blond Hair” text footprint sits in a different spectral region than the “Male” footprint. A log-Gaussian response can therefore retain the target footprint while suppressing a nuisance footprint. Figure 4.2 reproduces the

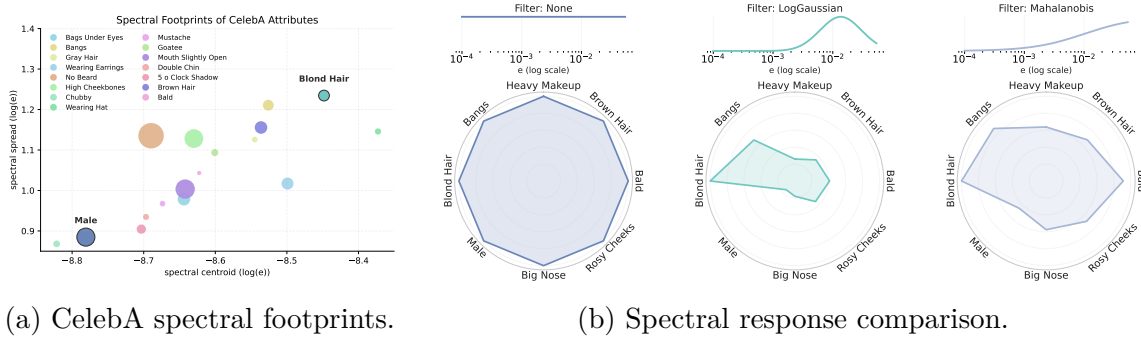


Figure 4.2: Spectral analysis of CelebA attributes. The target attribute and nuisance attributes occupy different regions of the covariance spectrum; the log-Gaussian response preserves the target band more selectively than an identity or monotone Mahalanobis response.

paper’s spectral visualization and filter comparison.

4.9 Practical Properties

SPOT is lightweight because calibration is a matrix multiplication after a PCA. It is label-efficient because labels are needed only for validation, not for fitting the covariance. It is data-efficient because unlabeled target embeddings are sufficient to estimate the eigenspace. It is easy to integrate because the output is a calibrated matrix of class axes that replaces the original text prototypes. Finally, its closed-form structure is compatible with differential privacy, as developed in the next chapter.

Chapter 5

Differentially Private SPOT

Any adaptation method that depends on user data can leak information through its released parameters or predictions. In on-device settings, a calibrated model may be cached locally, shared across processes, or synchronized. The source paper therefore develops a differentially private variant, DP-SPOT, that releases a noisy version of the calibrated axes while preserving the inference pipeline.

5.1 Differential Privacy Background

Differential privacy (DP) controls the effect of a single record on a randomized output [5, 6]. A randomized mechanism \mathcal{M} satisfies (ϵ, δ) -DP if, for any adjacent datasets D and D' differing in one example and any measurable output set S ,

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \Pr[\mathcal{M}(D') \in S] + \delta. \quad (5.1)$$

A smaller ϵ means stronger privacy, and δ bounds a small failure probability. DP is closed under post-processing: once a private output is released, deterministic transformations of it do not consume additional privacy budget.

In this thesis, the released output is the calibrated axis matrix. The non-private SPOT map is

$$f_W(D) = \mathbf{W}^*(D) = F(\Sigma_D)\mathbf{T}. \quad (5.2)$$

DP-SPOT releases

$$\widetilde{\mathbf{W}} = F(S_D)\mathbf{T} + \mathcal{N}(0, \sigma_W^2 \mathbf{I}), \quad (5.3)$$

where S_D is a non-centered second moment, and the noise scale is calibrated to the global sensitivity of f_W .

5.2 Why Use the Non-Centered Second Moment

The centered covariance is natural for spectral analysis, but its mean subtraction complicates sensitivity. The DP derivation uses the non-centered second moment

$$S(D) = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top. \quad (5.4)$$

If embeddings are clipped to $\|\mathbf{z}_i\|_2 \leq B$, then under replace-one adjacency,

$$\Delta_2(S) = \sup_{D \sim D'} \|S(D) - S(D')\|_F \leq \frac{2B^2}{n}. \quad (5.5)$$

This simple bound is the starting point for the release mechanism.

5.3 Sensitivity of the Spectral Map

Let $F(S) = \mathbf{U} \text{diag}(G(e_1), \dots, G(e_d)) \mathbf{U}^\top$ be the spectral functional calculus applied to a positive semidefinite matrix $S = \mathbf{U} \text{diag}(e) \mathbf{U}^\top$. Suppose the spectrum lies in $[0, B^2]$ and the scalar response G is continuously differentiable with

$$L_G = \sup_{e \in [0, B^2]} |G'(e)| < \infty. \quad (5.6)$$

The paper's key spectral Lipschitz lemma gives

$$\|F(S) - F(S')\|_F \leq L_G \|S - S'\|_F. \quad (5.7)$$

Combining this with the second-moment sensitivity yields the global sensitivity of the calibrated weight matrix:

$$\Delta_2(\mathbf{W}^*) \leq \frac{2B^2}{n} L_G \|\mathbf{T}\|_F. \quad (5.8)$$

For the log-Gaussian response

$$G(e) = \exp\left(-\frac{(\log(e + \varepsilon) - m)^2}{2\sigma^2}\right), \quad (5.9)$$

the derivative is

$$G'(e) = -\frac{\log(e + \varepsilon) - m}{\sigma^2(e + \varepsilon)} G(e). \quad (5.10)$$

Algorithm 2 DP-SPOT release

Require: Embeddings $\{\mathbf{z}_i\}_{i=1}^n$; text axes \mathbf{T} ; clip radius B ; spectral response G ; privacy budget $(\varepsilon_W, \delta_W)$.

- 1: Clip each embedding so $\|\mathbf{z}_i\|_2 \leq B$.
- 2: $S \leftarrow \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top$.
- 3: Compute $F(S)$ using the spectral response G .
- 4: $\mathbf{W}^* \leftarrow F(S)\mathbf{T}$.
- 5: Compute L_G and $\Delta_W \leftarrow \frac{2B^2}{n} L_G \|\mathbf{T}\|_F$.
- 6: $\sigma_W \leftarrow \Delta_W \sigma_{\text{AGM}}(\varepsilon_W, \delta_W)$.
- 7: $\widetilde{\mathbf{W}} \leftarrow \mathbf{W}^* + \mathcal{N}(0, \sigma_W^2 \mathbf{I})$.
- 8: Optionally normalize columns of $\widetilde{\mathbf{W}}$.
- 9: **return** $\widetilde{\mathbf{W}}$.

A conservative bound is

$$L_G \leq \frac{1}{\sigma^2 \varepsilon} \max \left\{ |\log \varepsilon - m|, |\log(B^2 + \varepsilon) - m| \right\}. \quad (5.11)$$

The supplementary material also introduces a small eigenvalue floor $\eta \geq 0$ to stabilize the derivative:

$$G_\eta(e) = \exp \left(-\frac{(\log(e + \eta + \varepsilon) - m)^2}{2\sigma^2} \right). \quad (5.12)$$

5.4 Analytic Gaussian Mechanism

The source paper uses the Analytic Gaussian Mechanism (AGM) [2]. For a deterministic statistic $f(D)$ with ℓ_2 sensitivity Δ , the Gaussian mechanism releases

$$\mathcal{M}(D) = f(D) + Z, \quad Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad (5.13)$$

with $\sigma = \Delta \sigma_{\text{AGM}}(\varepsilon, \delta)$, where σ_{AGM} is the smallest normalized scale satisfying the AGM calibration equation. In DP-SPOT,

$$\sigma_W = \sigma_{\text{AGM}}(\varepsilon_W, \delta_W) \cdot \frac{2B^2}{n} L_G \|\mathbf{T}\|_F. \quad (5.14)$$

Column normalization, margin application, and inference with $\widetilde{\mathbf{W}}$ are post-processing steps.

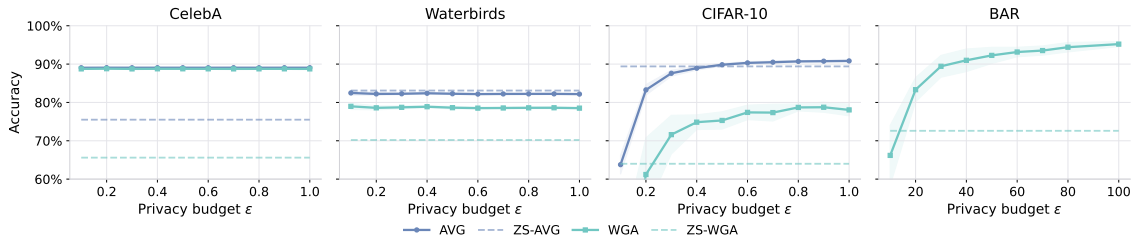


Figure 5.1: Test accuracy versus privacy budget ϵ for DP-SPOT with a ViT-B/16 backbone. Solid curves show average and worst-group accuracy; dashed lines show zero-shot baselines. The results indicate that the closed-form calibration can be privatized while preserving substantial WGA gains.

5.5 DP-SPOT Algorithm

5.6 Privacy-Utility Observations

The original experiments show that DP-SPOT remains close to non-private SPOT on several datasets under strict budgets. On CelebA, Waterbirds, and CIFAR-10.02, DP-SPOT with $\epsilon = 1$ achieves performance close to the non-private calibration. For BAR, the paper uses a larger budget because the per-class sample size is smaller and the required noise would otherwise harm utility. Figure 5.1 reproduces the paper’s privacy-budget sweep for a ViT-B/16 backbone.

5.7 Scope of the Privacy Claim

The DP guarantee protects the released calibrated axes under the stated adjacency definition and clipping assumptions. Hyperparameter tuning consumes privacy if it is performed on private data and the selected hyperparameters are released. The source paper notes this as an important direction for future work and focuses the formal guarantee on releasing the calibrated matrix for fixed hyperparameters. In an on-device deployment where tuning remains local and no calibrated state leaves the device, the operational privacy risk differs from a public release; nevertheless, the DP-SPOT derivation is valuable because it bounds how much a single image can affect the cached calibration state.

Chapter 6

Experiments

This chapter reports two sets of experiments. The first follows the original SPOT paper and evaluates standard group-robustness benchmarks: CelebA, Waterbirds, BAR, and CIFAR-10.02. The second follows the slide deck and evaluates PCP-Face-v1, the personal calibration setting defined in [chapter 3](#).

6.1 Standard Group-Robustness Setup

Datasets. The standard benchmark set contains four datasets. CelebA evaluates a binary hair-color task, with gender as the spurious attribute [16]. Waterbirds evaluates waterbird versus landbird classification with background as the spurious attribute [21]. BAR evaluates six action classes with background correlations. CIFAR-10.02 combines CIFAR-10 and CIFAR-10.2 and treats the data source as a spurious or domain-shift attribute [13, 17].

Metrics. The paper reports average accuracy (AVG) and worst-group accuracy (WGA), where groups are class-attribute pairs. On BAR, the source table reports AVG for each backbone.

Validation. The source paper uses two validation regimes: with group labels, hyperparameters are selected to maximize WGA; without group labels, hyperparameters are selected using worst-class accuracy (WCA). The search grid is $m \in [-10, 0]$ with step 0.1, $\sigma \in [0.5, 10]$ with step 0.1, and class thresholds $\tau_c \in [-0.1, 0.1]$ with step 0.01.

Backbones and prompts. Experiments use CLIP backbones including ViT-L/14, ViT-B/16, and ResNet-50. For each class, prompts are encoded by the CLIP text encoder and averaged to obtain a class prototype.

6.2 Main Results on Standard Benchmarks

Table 6.1 reproduces the main result table from the source paper. SPOT improves over zero-shot CLIP on all datasets and is particularly strong in multi-class settings. On BAR, many trainable or proxy-based baselines fail to improve over zero-shot, while SPOT improves the reported AVG on both backbones. On CIFAR-10.02, SPOT reaches the top WGA among the listed methods for both ViT-L/14 and RN50. DP-SPOT remains close to the non-private variant, showing that spectral calibration can be privatized with limited loss in many settings.

Table 6.1: Results on CelebA, Waterbirds, BAR, and CIFAR-10.02 with ViT-L/14 and ResNet-50 backbones. Values are accuracies in percentage points. The table follows the source paper. BAR columns report AVG.

Method	CelebA				Waterbirds				BAR		CIFAR-10.02			
	L/14		RN50		L/14		RN50		L/14	RN50	L/14		RN50	
	AVG	WGA	AVG	WGA	AVG	WGA	AVG	WGA	AVG	AVG	AVG	WGA	AVG	WGA
CLIP-Zero-Shot	78.1	72.3	84.5	80.2	86.5	77.9	79.7	62.7	93.3	69.9	93.6	71.5	70.4	20.5
ERM	94.7	28.3	94.7	11.9	97.6	65.9	93.5	7.9	83.8	60.9	95.9	82.5	78.5	54.0
OrthCali	86.2	76.1	84.4	82.2	84.5	68.8	78.7	74.0	63.0	50.6	91.7	73.5	68.8	36.5
AFR	85.2	70.0	94.3	53.4	88.2	73.4	89.3	48.4	86.8	62.0	95.4	83.0	74.6	47.5
JTT	93.3	75.6	90.6	61.7	97.3	83.6	90.6	61.7	83.0	60.7	96.0	81.0	78.1	55.0
CnC	89.3	79.2	90.3	63.9	97.5	84.5	87.1	61.2	35.9	29.4	95.7	84.5	72.4	38.0
FairerCLIP	87.8	85.2	85.0	81.5	92.2	86.0	84.3	75.4	96.6	74.4	96.1	81.0	77.9	44.0
CFR	84.5	73.6	93.4	73.9	76.7	58.2	73.1	50.0	89.1	59.9	95.6	82.5	80.1	57.5
PPA	82.8	77.8	86.0	78.3	81.3	64.9	75.4	54.0	93.3	68.7	95.0	81.5	75.9	44.5
SPOT	89.6	83.3	90.6	84.4	86.9	83.1	80.3	78.8	96.6	82.6	95.0	85.0	76.6	58.5
DP-SPOT	89.3	84.2	90.6	84.1	86.9	82.4	80.1	78.6	97.1	82.9	95.6	84.9	75.2	55.4

Table 6.2: SPOT with and without group labels on validation. Without group labels, the validation objective is class-balanced or worst-class accuracy.

Dataset	Validation	ViT-L/14		ViT-B/16		RN50	
		AVG	WGA	AVG	WGA	AVG	WGA
CelebA	CLIP-ZS	78.1	72.3	75.5	62.6	84.5	80.2
CelebA	SPOT w/ GL	89.6	83.3	89.1	88.7	90.6	84.4
CelebA	SPOT w/o GL	90.7	70.6	90.3	75.0	91.2	70.0
Waterbirds	CLIP-ZS	86.5	77.9	83.1	70.2	79.7	62.7
Waterbirds	SPOT w/ GL	86.9	83.1	82.1	78.5	80.3	78.8
Waterbirds	SPOT w/o GL	87.1	81.7	84.3	74.6	82.7	73.7

6.3 Validation Without Group Labels

The source paper also evaluates SPOT when the validation split lacks group labels. In this case, the spectral parameters and thresholds are selected by WCA rather than WGA. Table 6.2 shows that performance is lower than group-supervised selection in some cases, especially CelebA WGA, but remains competitive and stable. This result matters because the PCP protocol removes group labels from adaptation altogether.

6.4 Data Efficiency and Hyperparameter Smoothness on Standard Benchmarks

Figures 6.1 and 6.2 reports the two standard-benchmark diagnostics from the source paper. The left panel studies data efficiency: AVG and WGA are plotted as functions of the fraction of target-domain training data used for spectral calibration, with dashed horizontal lines indicating the zero-shot baselines. The curves rise quickly and then plateau, showing that SPOT can recover most of its gains from a relatively small subset of target-domain features.

The right panel studies hyperparameter smoothness. Accuracy changes smoothly as a function of the filter center m and bandwidth σ on CelebA and Waterbirds. The fair region marks settings where AVG and WGA are closely aligned, supporting the practical claim that SPOT is not highly sensitive to exact filter-parameter choice.

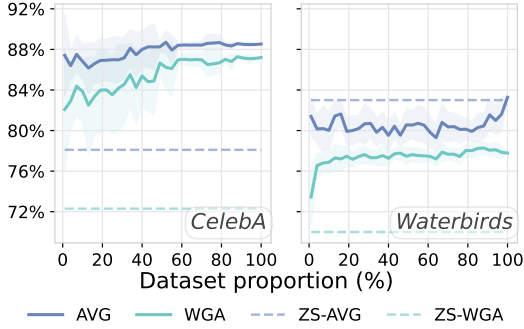


Figure 6.1: AVG and WGA as functions of varying % of training data. Shaded regions denote standard deviations obtained by sampling random subsets of the training set.

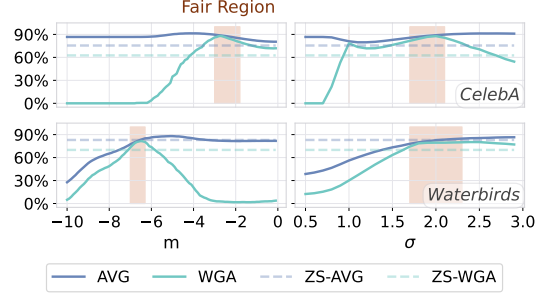


Figure 6.2: Accuracy (AVG and WGA) as a function of filter parameter m (left) and σ (right) on CelebA (top) and Waterbirds (bottom). *Fair Region* denotes where AVG and WGA are closely aligned.

Table 6.3: UTKFace WGA at $k_{\text{user}} = k_{\text{retr}} = 64$ under PCP. Zero-shot WGA is 32.46.

Method	LP++	Tip-Adapter	CLIP-Adapter	WiSE-FT	JTT	GEORGE
WGA (%)	15.6	1.3	17.4	36.8	16.8	21.2

6.5 PCP-Face-v1: Baseline Failure Under Sparse Corrections

The slide deck first evaluates group-label-free methods on UTKFace at $k_{\text{user}} = k_{\text{retr}} = 64$. [Table 6.3](#) summarizes the slide. The dashed zero-shot reference is 32.46 WGA. Several adaptation methods fall below this reference, showing that naive use of sparse corrections can degrade robustness.

6.6 PCP-Face-v1 Main Results

[Table 6.4](#) reproduces the PCP-Face-v1 main-results table from the slide deck. All methods use CLIP ViT-L/14 with $k_{\text{user}} = 256$, $k_{\text{retr}} = 256$, and $|\mathcal{D}_{\text{album}}| = 2000$. Values are WGA mean \pm standard deviation over 20 trials. SPOT achieves the largest margin over zero-shot on all three datasets: +26.7 on CelebA, +17.1 on FairFace, and +24.1 on UTKFace. Some baselines improve one dataset but degrade another; SPOT is the only listed method with large positive improvements across all three.

Table 6.4: PCP-Face-v1 WGA at $k_{\text{user}} = k_{\text{retr}} = 256$ with CLIP ViT-L/14. Values are mean \pm standard deviation over 20 trials. The FairFace and UTKFace columns are the 4-way variants reported in the slide table.

Method	CelebA hair 4-way		FairFace race 4-way		UTKFace ethnicity 4-way	
	WGA	Δ	WGA	Δ	WGA	Δ
Zero-shot	41.2	–	35.7	–	32.5	–
LP++	45.7 ± 12.8	+4.5	43.7 ± 4.5	+8.1	26.0 ± 11.5	-6.5
Tip-Adapter	12.0 ± 5.2	-29.2	20.4 ± 9.9	-15.3	1.9 ± 1.8	-30.6
CLIP-Adapter	47.2 ± 9.6	+6.0	44.2 ± 5.1	+8.6	25.0 ± 10.2	-7.5
WiSE-FT	41.3 ± 14.1	+0.1	40.0 ± 11.6	+4.4	47.5 ± 10.0	+15.0
JTT	34.9 ± 9.0	-6.3	38.9 ± 7.9	+3.3	32.8 ± 8.5	+0.3
GEORGE	38.2 ± 8.8	-3.0	44.7 ± 4.8	+9.1	41.8 ± 7.9	+9.3
SPOT	67.9 ± 3.5	+26.7	52.7 ± 4.8	+17.1	56.6 ± 3.5	+24.1

Table 6.5: PCP-Face-v1 WGA as a function of oracle budget $k_{\text{user}} = k_{\text{retr}}$. Data come from the slide deck.

Dataset	SPOT				CLIP-Adapter			
	$k = 32$	$k = 64$	$k = 128$	$k = 256$	$k = 32$	$k = 64$	$k = 128$	$k = 256$
CelebA	45.8	59.7	61.0	67.9	43.7	49.5	48.3	47.2
FairFace	27.7	42.4	45.9	52.7	37.1	36.9	41.5	44.2
UTKFace	30.3	38.6	50.6	56.6	11.8	17.4	23.3	25.0

6.7 Scaling with Oracle Budget

The slide deck compares SPOT and CLIP-Adapter across correction budgets $k_{\text{user}} = k_{\text{retr}} \in \{32, 64, 128, 256\}$. Table 6.5 plots the WGA curves. SPOT improves substantially as the correction budget grows, especially on FairFace and UTKFace, while CLIP-Adapter often plateaus or remains below zero-shot. The main point is not that SPOT needs many labels, but that its low-dimensional spectral tuning can productively use additional corrections without overfitting as quickly as a feature adapter.

6.8 Per-Attribute Diagnostic on CelebA

The slide deck also includes a per-attribute diagnostic for CelebA. The task is 4-way hair-color classification with $k_{\text{user}} = 64$. The plot reports Δ WGA relative to zero-shot across 35 attributes, excluding Bald. The summary is shown in table 6.6: SPOT improves 32 of 35 attributes, has a mean improvement of +15.6, and improves 29

Table 6.6: Summary of the CelebA per-attribute diagnostic from the slide deck.

Diagnostic	Mean Δ	Median	Max	Min	Improved	$\Delta > 5$
Δ WGA vs. zero-shot	+15.6	+19.3	+24.4	-7.8	32/35	29/35

attributes by more than five percentage points. The few regressions are concentrated in attributes with negative deltas: Chubby, Sideburns, and Mustache.

6.9 Experimental Takeaways

The experiments support four claims. First, text-axis calibration is sufficient to improve group robustness in many cases; the encoders need not be retrained. Second, spectral calibration is especially useful when target tasks are multi-class or when correction labels are sparse. Third, the same closed-form structure allows privacy analysis and DP release, which is important for the on-device scenario motivating PCP.

Chapter 7

Discussion and Limitations

7.1 Why Spectral Calibration Works

The central empirical observation behind SPOT is that CLIP text axes have structured energy when expressed in the target image covariance eigenspace. If a target semantic attribute and a nuisance attribute occupy different spectral regions, a band-pass response can keep one and shrink the other. This is visible in the CelebA blond-hair analysis: the target hair-color footprint differs from the gender footprint, so the log-Gaussian response can retain hair-color energy while attenuating gender-associated energy.

This interpretation also explains why a monotone Mahalanobis response is insufficient as a universal solution. A high-pass filter assumes that high-variance directions are informative and low-variance directions are noise. That assumption can hold for some tasks, but it is not guaranteed. A nuisance can occupy a high-variance direction, and a target semantic can live in a middle or low spectral band. The log-Gaussian response is therefore a more flexible two-parameter mechanism for matching the target semantic footprint.

7.2 Why PCP Changes the Evaluation Problem

Standard group-robustness benchmarks evaluate whether a method can improve WGA given either group labels or enough validation information to infer robust hyperparameters. PCP changes the nature of the supervision. The labels are corrections of the reference model’s errors. This is realistic for personal calibration because a user normally interacts with a model by correcting mistakes, not by producing a balanced labeled dataset.

This conditioning makes the learning problem harder. The correction set is not representative of the full task distribution. It is biased toward failure modes, may contain rare groups, and may be small. In such a regime, high-capacity adapters can overfit. SPOT limits this risk by making the trainable part of adaptation a small number of spectral parameters and margins while using unlabeled data to estimate

geometry.

7.3 Limitations

Linear feature-space calibration. SPOT calibrates text axes through a linear transformation in the frozen embedding space. If semantic and nuisance directions are inseparable in that space, or if they occupy the same spectral band, direction-wise shrinkage cannot disentangle them. A possible extension is a kernelized or locally adaptive variant that separates nonlinear semantic and nuisance structure.

Dependence on covariance quality. The method relies on a useful estimate of target-domain covariance. With very small unlabeled samples, heavy distribution shift between the unlabeled pool and the evaluation distribution, or strong outliers, the eigenspace may be unstable. Robust covariance estimators or streaming updates could improve reliability.

Validation dependence. The log-Gaussian response and class margins are selected on validation data or sparse corrections. If the validation distribution does not represent the test distribution, the selected passband can degrade performance. The PCP-Face-v1 results suggest that SPOT is less fragile than several baselines, but validation mismatch remains a limitation.

Group labels still used for evaluation. PCP removes group labels from adaptation, but the reported WGA metrics require group labels at evaluation time. This is appropriate for benchmarking but does not solve the broader challenge of auditing unknown or hidden groups in deployment.

Differential privacy and hyperparameter tuning. DP-SPOT protects the released calibrated axes for fixed hyperparameters. If the tuning process itself is performed on private data and exposed, its privacy cost must be accounted for. The source paper identifies private hyperparameter tuning as future work.

7.4 Design Implications

The thesis suggests several principles for fair personal adaptation. First, use abundant unlabeled local data to estimate geometry rather than relying only on sparse labels. Second, constrain adaptation capacity when labels are corrections rather than

representative samples. Third, preserve the language interface of zero-shot models: users should be able to specify tasks with prompts while the system handles calibration. Fourth, treat privacy as a design constraint, not an afterthought; a calibration state should be small, auditable, and, when necessary, privatizable.

Chapter 8

Conclusion

This thesis presented SPOT, a closed-form spectral calibration method for CLIP-like zero-shot classifiers. SPOT uses unlabeled target-domain image embeddings to estimate covariance geometry, decomposes prompt-derived text axes in that eigenspace, and applies a validation-selected spectral response to produce calibrated decision axes. The method requires no gradient-based training, no new encoder, and no change to the prompting interface.

The thesis also introduced the Personal Calibration Protocol, which models a realistic on-device scenario where a user has a private unlabeled album, task retrievals generated by a fixed reference model, and sparse oracle corrections of the model’s own errors. PCP-Face-v1 shows that conventional few-shot and adaptation baselines can be unreliable in this sparse-correction regime, while SPOT improves worst-group accuracy across CelebA, FairFace, and UTKFace.

The differential privacy analysis shows that SPOT’s closed-form map admits sensitivity bounds, enabling a DP release of the calibrated axis matrix. This property is important for human-sensing systems, where calibration data may be personal and sensitive.

Overall, SPOT supports a simple thesis: fair and private zero-shot adaptation need not require retraining large models or collecting explicit group labels. By using target-domain spectral geometry and sparse validation feedback, it is possible to recalibrate text axes in a lightweight, interpretable, and deployable way.

Chapter A

Implementation Details

This appendix collects implementation details from the source manuscript.

A.1 Prompt Design

For all vision-language experiments, prompts are fixed before validation. The text encoder embeds each prompt, and prompt embeddings for the same class are averaged to form a class prototype.

BAR. The BAR experiment uses six action classes: climbing, diving, fishing, racing, throwing, and vaulting. Each class is associated with four short prompts in a generic photographic context, such as “a photo of rock climbing” and “a person climbing a rock wall” for climbing, or “a photo of pole vaulting” and “athlete vaulting over bar” for vaulting.

CIFAR-10.02. The CIFAR-10.02 experiment uses twelve CLIP-style prompts per class, varying viewpoint and photographic style while keeping the class name fixed. Examples include “photo of airplane”, “side view photo of airplane on plain background”, and “isolated airplane on white background.”

Waterbirds. Waterbirds prompts are constructed from a Cartesian product of generic bird contexts and waterbird or landbird species names. Contexts include phrases such as “a photo of the {}” and “the {} taking off.”

CelebA. CelebA hair-color prompts use attribute templates and synonyms. Positive hair-color prompts include phrases such as “blond hair”, “fair hair”, and “golden hair”. Negative prompts include “dark hair”, “black hair”, and “deep brown hair”. These are inserted into templates such as “a photo of a person featuring {}” and “a close-up headshot highlighting {}.”

A.2 Hyperparameter Search

For the standard benchmark experiments, the source manuscript uses grid search over

$$m \in [-10, 0], \quad \Delta m = 0.1, \quad \sigma \in [0.5, 10], \quad \Delta \sigma = 0.1, \quad (\text{A.1})$$

and class thresholds

$$\tau_c \in [-0.1, 0.1], \quad \Delta \tau = 0.01. \quad (\text{A.2})$$

With group labels, WGA is the validation objective. Without group labels, WCA is the validation objective. The shared spectral response uses one pair (m, σ) for all classes.

A.3 Shared Versus Classwise Spectral Parameters

A classwise variant assigns each class its own spectral parameters (m_c, σ_c) . This creates a $2C$ -dimensional search space, so the source manuscript uses Tree-structured Parzen Estimator Bayesian optimization with a cap of 100 trials. The source results suggest that classwise parameters can improve some test WGA values, but they also increase the validation-test WGA gap. The shared parameterization is therefore more conservative and stable across backbones.

Chapter B

Additional Theory and Proofs

This appendix expands the theoretical results used in [chapters 4](#) and [5](#).

B.1 Convexity and Closed Form

Fix a class and write \mathbf{t} and \mathbf{w} for the original and calibrated axes. The objective is

$$J(\mathbf{w}) = (\mathbf{w} - \mathbf{t})^\top \Sigma_g (\mathbf{w} - \mathbf{t}) + \lambda \mathbf{w}^\top \mathbf{w}. \quad (\text{B.1})$$

Since $\Sigma_g \succeq 0$ and $\lambda > 0$, the Hessian is $2(\Sigma_g + \lambda \mathbf{I}) \succ 0$. Therefore J is strictly convex and has a unique minimizer. Setting the gradient to zero gives

$$2\Sigma_g(\mathbf{w} - \mathbf{t}) + 2\lambda\mathbf{w} = 0, \quad (\text{B.2})$$

which implies

$$\mathbf{w}^* = (\Sigma_g + \lambda \mathbf{I})^{-1} \Sigma_g \mathbf{t}. \quad (\text{B.3})$$

If $\Sigma_g = \mathbf{U} \text{diag}(g(\mathbf{e})) \mathbf{U}^\top$, the solution in the eigenbasis is

$$\mathbf{w}_j^* = \frac{g(e_j)}{g(e_j) + \lambda} \mathbf{t}_j. \quad (\text{B.4})$$

B.2 Bijection Between Weight Profile and Shrinkage Profile

For $\lambda > 0$, define

$$G(e) = \frac{g(e)}{g(e) + \lambda}. \quad (\text{B.5})$$

If $g(e) \in [0, \infty)$, then $G(e) \in [0, 1)$. Conversely, if $G(e) \in [0, 1)$, then

$$g(e) = \lambda \frac{G(e)}{1 - G(e)} \in [0, \infty). \quad (\text{B.6})$$

The two maps are inverses pointwise. Applying them elementwise to the eigenvalues of a PSD matrix gives the operator-level mapping

$$(\boldsymbol{\Sigma}_g + \lambda \mathbf{I})^{-1} \boldsymbol{\Sigma}_g = \mathbf{U} \text{diag}(G(e_1), \dots, G(e_d)) \mathbf{U}^\top. \quad (\text{B.7})$$

B.3 Spectral Lipschitz Bound

Let $S, S' \succeq 0$ have spectra in $[0, B^2]$. Let $F(S)$ be defined by spectral calculus using scalar response G , and assume G is continuously differentiable with $L_G = \sup_{e \in [0, B^2]} |G'(e)|$. The source proof uses the Frechet derivative of spectral functions. Along the interpolation $S_t = (1-t)S + tS'$, the fundamental theorem of calculus gives

$$F(S') - F(S) = \int_0^1 DF_{S_t}[S' - S] dt. \quad (\text{B.8})$$

The derivative is bounded by L_G in Frobenius norm, yielding

$$\|F(S') - F(S)\|_F \leq L_G \|S' - S\|_F. \quad (\text{B.9})$$

B.4 Global Sensitivity

Assume all embeddings are clipped to $\|\mathbf{z}_i\|_2 \leq B$. For the non-centered second moment

$$S(D) = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top, \quad (\text{B.10})$$

replace-one adjacency changes one summand. Thus

$$\|\mathbf{z}\mathbf{z}^\top - \mathbf{z}'\mathbf{z}'^\top\|_F \leq \|\mathbf{z}\mathbf{z}^\top\|_F + \|\mathbf{z}'\mathbf{z}'^\top\|_F = \|\mathbf{z}\|_2^2 + \|\mathbf{z}'\|_2^2 \leq 2B^2, \quad (\text{B.11})$$

and $\Delta_2(S) \leq 2B^2/n$. For $\mathbf{W}^*(D) = F(S(D))\mathbf{T}$,

$$\|\mathbf{W}^*(D) - \mathbf{W}^*(D')\|_F \leq \|F(S(D)) - F(S(D'))\|_2 \|\mathbf{T}\|_F \quad (\text{B.12})$$

$$\leq L_G \|S(D) - S(D')\|_F \|\mathbf{T}\|_F \quad (\text{B.13})$$

$$\leq \frac{2B^2}{n} L_G \|\mathbf{T}\|_F. \quad (\text{B.14})$$

This is the sensitivity used by DP-SPOT.

Chapter C

Additional PCP-Face-v1 Tables

This appendix records additional numerical details from the slide deck for traceability.

C.1 Scaling Data

[Table C.1](#) lists the SPOT and CLIP-Adapter scaling values plotted in [table 6.5](#).

Table C.1: WGA values used for the PCP-Face-v1 scaling plot.

Dataset	SPOT				CLIP-Adapter			
	$k = 32$	$k = 64$	$k = 128$	$k = 256$	$k = 32$	$k = 64$	$k = 128$	$k = 256$
CelebA	45.8	59.7	61.0	67.9	43.7	49.5	48.3	47.2
FairFace	27.7	42.4	45.9	52.7	37.1	36.9	41.5	44.2
UTKFace	30.3	38.6	50.6	56.6	11.8	17.4	23.3	25.0

C.2 CelebA Per-Attribute Deltas

The slide deck provides 35 CelebA per-attribute WGA deltas at $k_{\text{user}} = 64$. [Table C.2](#) lists the attributes in the same descending order as the slide chart.

Table C.2: CelebA per-attribute zero-shot WGA and SPOT Δ WGA from the slide deck.

Attribute	Zero-shot WGA	Δ WGA
Bags Under Eyes	43.8	+24.4
Straight Hair	45.2	+23.1
Oval Face	46.2	+22.8
Bushy Eyebrows	46.3	+22.1
Wearing Lipstick	43.2	+21.9
Heavy Makeup	45.0	+21.9
Pointy Nose	46.8	+21.2

Attribute	Zero-shot WGA	Δ WGA
Bangs	48.4	+21.1
5 o Clock Shadow	37.1	+21.0
Wearing Earrings	47.6	+20.7
Arched Eyebrows	48.3	+20.6
Smiling	48.1	+20.5
Pale Skin	49.9	+20.3
Mouth Slightly Open	49.5	+19.9
Male	41.2	+19.8
Receding Hairline	45.2	+19.5
Wavy Hair	48.3	+19.4
Wearing Necklace	49.0	+19.3
High Cheekbones	50.0	+19.2
Attractive	49.4	+18.9
Narrow Eyes	50.2	+18.1
Blurry	50.3	+17.9
Big Lips	47.4	+17.6
Big Nose	50.0	+17.3
Rosy Cheeks	50.1	+14.8
Double Chin	45.0	+14.2
Wearing Hat	50.2	+14.0
Eyeglasses	47.3	+13.8
No Beard	41.9	+11.9
Young	45.1	+4.4
Wearing Necktie	34.1	+3.0
Goatee	28.6	+1.1
Chubby	45.9	-6.2
Sideburns	42.4	-7.3
Mustache	50.1	-7.8

Bibliography

- [1] Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://openreview.net/forum?id=fCeUoDr9Tq>.
- [2] Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning (ICML)*, 2018. URL <https://proceedings.mlr.press/v80/balle18a/balle18a.pdf>.
- [3] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. URL <https://arxiv.org/abs/2302.00070>.
- [4] Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. Fairerclip: Debiasing clip’s zero-shot predictions using functions in rkhss. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://openreview.net/pdf?id=HXoq9EqR9e>.
- [5] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/04000000042.
- [6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, volume 3876 of *LNCS*, pages 265–284. Springer, 2006. doi: 10.1007/11681878_14.
- [7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.
- [8] Yusuke Hirota, Min-Hung Chen, Chien-Yi Wang, Yuta Nakashima, Yu-Chiang Frank Wang, and Ryo Hachiuma. Saner: Annotation-free societal attribute neutralizer for debiasing clip. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://arxiv.org/abs/2408.10202>.

- [9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021. URL <https://proceedings.mlr.press/v139/jia21b/jia21b.pdf>.
- [10] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.
- [11] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, 2018. URL <https://proceedings.mlr.press/v80/kearns18a.html>.
- [12] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://openreview.net/forum?id=uyxIY8Q0bV>.
- [13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- [14] Phuong Quynh Le, Jörg Schlötterer, and Christin Seifert. Out of spuriousity: Improving robustness to spurious correlations without group annotations. *Transactions on Machine Learning Research*, 2024. URL <https://openreview.net/forum?id=EEeVYfXor5>.
- [15] Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021.
- [16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [17] Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *International Conference on Machine*

Learning (ICML) Workshop on Uncertainty and Robustness in Deep Learning, 2020.

- [18] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [19] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning (ICML)*, 2023. URL <https://arxiv.org/abs/2306.11074>.
- [20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. URL <https://proceedings.mlr.press/v139/radford21a/radford21a.pdf>.
- [21] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations (ICLR)*, 2019. URL <https://openreview.net/pdf?id=ryxGuJrFvS>.
- [22] Nimit S. Sohoni, Maziar Sanjabi, Nicolas Ballas, Aditya Grover, Shaoliang Nie, Hamed Firooz, Christopher Re, and Christopher G. White. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [23] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [24] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wortsman_Robust_Fine-Tuning_of_Zero-Shot_Models_CVPR_2022_paper.html.

- [25] Michael Zhang, Nimit S. Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning (ICML)*, 2022.
- [26] Renrui Zhang, Ziyu Guo, Xianzheng Ma, Xuming He, Bin Cui, et al. Tip-adapter: Training-free adapter for vision-language models. In *European Conference on Computer Vision (ECCV)*, 2022.
- [27] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [28] Beier Zhu, Jiequan Cui, Hanwang Zhang, and Chi Zhang. Project-probe-aggregate: Efficient fine-tuning for group robustness. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://arxiv.org/abs/2503.09487>.