GSSP: ELEVATING UNDER-PERFORMING CLASSES IN ZERO-SHOT IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Language models exhibit strong generalized zero-shot image classification capabilities, however, they frequently suffer from severe performance disparities with accuracy in which under-performing classes often fall to near zero accuracy. To address this without the computational burden of retraining, we propose the Gram-Schmidt Subspace Projection (GSSP) a training-free, lightweight framework that operates directly in the post-projection embedding space. Our method constructs orthonormal bases using the corresponding text embeddings of the under-performing classes, applying targeted geometric transformations to the visual features. Because raw projection can cause catastrophic forgetting of a model's generalized knowledge, we introduce the extension GSSP-Boost, which utilizes a parametrized blending mechanism to regularize the method and maintain the overall performance. Through comprehensive empirical evaluations across a diverse suite of datasets, we demonstrate that GSSP-Boost significantly improves the under-performing classes' accuracy while preserving the model's overall performance. Notably, our approach consistently outperforms existing baselines across different architectures, offering a highly effective flexible method for deploying robust zero-shot image classification.

1 INTRODUCTION

Vision-Language Models (VLMs) have fundamentally changed zero-shot image classification by enabling models to recognize unseen categories through natural language Radford et al. (2021); Jia et al. (2021). By aligning visual and textual representations within a shared embedding space, these models enable impressive zero-shot capabilities without the need for task-specific fine-tuning Zhai et al. (2023); Li et al. (2022); Alayrac et al. (2022). However, despite their broad success and scalability, VLMs frequently exhibit severe performance disparities across different semantic concepts Pang et al. (2025); Hirota et al. (2025). While a subset of dominant classes often achieve high—or even near-perfect—accuracy, a distinct subset of classes suffers from systematic under-performance Taori et al. (2020). This skewed performance distribution is well-documented in real-world evaluations, for example, zero-shot models often display degraded accuracy when tested in specialized categories Ju et al. (2025); Luo et al. (2024), or out-of-distribution environments Ming et al. (2022); Oh et al. (2024), revealing deep algorithmic sensitivities to text descriptions and image quality. Addressing these disparities is crucial, as the practical viability of zero-shot image classification depends on eliminating total failures across all classes rather than simply maximizing the average accuracy of the dominant ones.

The rise of CLIP Radford et al. (2021) has seen it employed commonly as a zero-shot classification method. In the zero-shot setting, classification is performed entirely through the similarity between image embeddings and text embeddings. The performance disparity across classes arises from intrinsic properties of the learned embedding space. This could be a result of certain classes being poorly represented by their textual labels or visual embeddings cluster in regions that are less well-aligned with the corresponding text embeddings Liang et al. (2022). This poor alignment may stem from the distribution of the training data Mayilvahanan et al. (2024), from the granularity of the class labels Jeon et al. (2026) or the structure of the shared embedding space Shi et al. (2023). There are two main issues to exploring this further: Firstly, that training datasets are either withheld for closed-source models or too large to determine class statistics. Secondly, the domain gap between the classes used for zero-shot classification and the training captions representing sparse detail across the

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.

image lead to an inability to know how frequently classes (or class concepts) appear in the training data. I.e., captions do not mention every detail in the image, but focus only on key, salient details. Because of these issues, it is hard to utilize training-time techniques with the aim of creating a general model that can adapt to different zero-shot settings at test time.

A growing body of work provides compelling evidence that actively fixing this misalignment directly translated to improved zero-shot generalization Magistri et al. (2026); Mistretta et al. (2025); Qian et al. (2023). AlignCLIP Eslami & de Melo (2025) demonstrated that enforcing tighter geometric alignment between text and image embeddings yields direct, quantifiable gains across a wide variety of zero-shot classification and cross-modal retrieval tasks. Similarly, post-pre-training modality alignment techniques such as those in Yamaguchi et al. (2025) successfully mitigate this gap, empirically proving that structurally bringing the feature clusters closer together improves zero-shot performance on downstream tasks. Other existing methods for improving dual-encoder models primarily focus on prompt optimization, lightweight adapters, or additional training procedures designed to refine the shared embeddings space. Continuous prompt tuning attempts to automate the discovery of optimal semantic contexts Zhou et al. (2022b;a), while adapter-based methods introduce trainable bottleneck layers to recalibrate visual and textual features Gao et al. (2024); Zhang et al. (2022b). While these approaches are undeniably effective at boosting overall accuracy, they present significant operational limitations. They typically require costly gradient updates, auxiliary few-shot data, or intensive test-time optimization Shu et al. (2022).

Inspired by the proven success of geometric gap-reduction, we propose a training-free, lightweight framework that directly manipulates the post-projection embedding space to improve the performance of under-performing classes. By dynamically identifying under-performing classes and utilizing their corresponding text embeddings to construct orthonormal bases, we apply targeted geometric transformations to the visual features. We combine structural operations with parameterized techniques so that our framework enhances the performance of the under-performing classes while maintaining the overall zero-shot performance.

Our contributions can be summarized as follows: (i) **A Novel Geometric Framework**: We introduce Gram-Schmidt Subspace Projection (GSSP) a training-free, computationally lightweight subspace intervention that operates directly in the post-projection embedding space; (ii) **Parametrized Blending**: We develop a parametrized intervention technique that successfully balances the enhancement of under-performing classes while preserving the overall performance; and (iii) **Empirical Evaluation**: We provide a comprehensive evaluation across ten zero-shot image classification datasets, systematically analyzing the impact of GSSP for under-performing classes.

2 Related Work

2.1 Training-Free Enhancement

Recognizing the practical advantages of avoiding any fine-tuning, several works propose training-free improvements to CLIP's zero-shot pipeline. CALIP Guo et al. (2023) introduces a parameter-free attention module that facilitates bidirectional cross-modal interaction between image and text features before classification. Because the contrastive pre-training has already reduced inter-modal distances, the module discards all learnable weights and updates features using only attention-based operations, yielding consistent zero-shot gains without any training and without requiring labeled data. CuPL Pratt et al. (2023) takes a different approach and targets the quality of the textual side of the classifier. Rather than using generic templates, CuPL queries a large language model (GPT-3) to generate a rich set of customized descriptive sentences for each class and aggregates the resulting text embeddings as the class prototype. The resulting improvement requires no changes to the vision encoder. Related approaches include Menon & Vondrick (2023) which exploits LLM-generated descriptors to enrich text features and WaffleCLIP Roth et al. (2023) which utilizes LLM-generated descriptions to augment text rather than altering the visual features, suggesting that much of CLIP's zero-shot performance variance can be attributed to the expressiveness of the class labels rather than the visual features. Building on these foundations, our proposed framework shifts the focus from feature augmentation to geometric alignment of the embedding space. Unlike previous methods, that rely on using LLMs which are prone to hallucinations and high computational cost, our framework is lightweight providing a more robust and efficient alternative.

2.2 TEST-TIME ADAPTATION (TTA)

A parallel line of work seeks to improve zero-shot performance at inference time by exploiting multiple views of each test image, without modifying any model parameters Shanmugam et al. (2021); Kim et al. (2020). The more recent and active direction applies TTA directly to the visual modality by generating multiple stochastic augmented views of a test image at inference time and aggregating the resulting image embeddings or predictions. MEMO Zhang et al. (2022a) established an influential general-purpose baseline for this paradigm. Given a single test sample, multiple augmented versions are generated, and the model's parameters are updated by minimizing the marginal entropy of the predictions across all views. Test-Time Prompt Tuning (TPT) Shu et al. (2022) adapt this augmentation-based consistency objective specifically to the CLIP prompt-learning framework. For each test sample, a batch of augmented views is constructed using random resized crops and other standard transforms. Among these views, TPT retains only the most confident ones and minimizes the marginal entropy of the averaged prediction over the retained views by back-propagating through learnable prompt tokens only, leaving all encoder weights frozen. A key limitation of TPT is that it still requires back-propagation through the text encoder at test time, introducing per-sample computational overhead proportional to the optimization steps. MeanShift for Test-time Augmentation (MTA) Zanella & Ben Ayed (2024) challenges the necessity of this optimization and argues that prompt tuning is not required to achieve strong TTA performance. MTA discards the prompt learning objective entirely and instead operates directly in the image embedding space. Given a set of augmented views of a test image, MTA seeks the mode of the density formed by the corresponding embeddings using a formulation of the MeanShift algorithm. Reinforcement Learning with CLIP Feedback (RLCF) Zhao et al. (2024) takes a different approach by framing test-time adaptation as a feedback-driven optimization problem. A frozen CLIP model serves as a reward model that evaluates the similarity between input images and samples generated or selected by the VLM being adapted. Contrary to these methods, our work operates through geometric manipulation, offering a lightweight method that allows for the targeted enhancement of under-represented classes while maintaining the integrity of the model's global zero-shot performance bypassing the computational overhead typically associated with test-time augmentation.

2.3 MODALITY GAP

Liang et al. (2022) first characterized the **modality gap** in CLIP. Even after training, image and text features occupy separate cones in the embedding space. They show this gap comes from initialization as well as the contrastive loss and that altering the gap distance (e.g. via temperature) can improve zero-shot classification.Optimization-based Textual Inversion (OTI) Baldrati et al. (2023) functions as a mapping technique that takes an extracted visual feature and iteratively optimizes a set of pseudo-tokens in the textual space until they accurately represent the visual anchor. Optimization-based Visual Inversion (OVI) Mistretta et al. (2025) performs the reverse operation, mapping textual features into the visual embedding space. It achieves this by iteratively refining a set of pseudo-visual features via gradient descent until they align with the target text concept. These methods significantly improve performance on intra-modal tasks by converting the problem into CLIP's native text-vs-image setting. Similarly, IsoCLIP Magistri et al. (2026) performs a spectral analysis on CLIP's learned projection matrices. They find an approximately isotropic subspace where the two modalities align and suggests removing modality-specific (anisotropic) directions. IsoCLIP requires no extra training, it simply decomposes CLIP's projectors and projection features onto the aligned subspace. Experiments show this training-free method reduces intra-modal misalignment and outperforms existing approaches for retrieval and classification. SSP Zhu et al. (2024) also addresses alignment but in the context of few-shot. It constructs vision and language specific subspaces spanned by local image features and projects both modalities into these subspaces for better alignment. Intra-Modal Proxy Learning (InMaP) Qian et al. (2023) proves that the contrastive loss in CLIP cannot fully close the modality gap and that for visual classification tasks the optimal class proxy may reside entirely within the vision modality. Given unlabeled target images, InMaP learns vision-space class proxies initialized from the corresponding text embeddings and refined via pseudo-label based self-training. Unlike optimization-heavy methods like OVI or OTI, our approach avoids per-sample back-propagation, instead leveraging targeted geometric transformations. While existing frameworks such as SSP and IsoCLIP rely on modality-specific subspaces or spectral decomposition, our approach introduces a targeted geometric transformation specifically for under-performing classes while maintaining the overall zero-shot performance, providing a more efficient, optimization-free alternative.

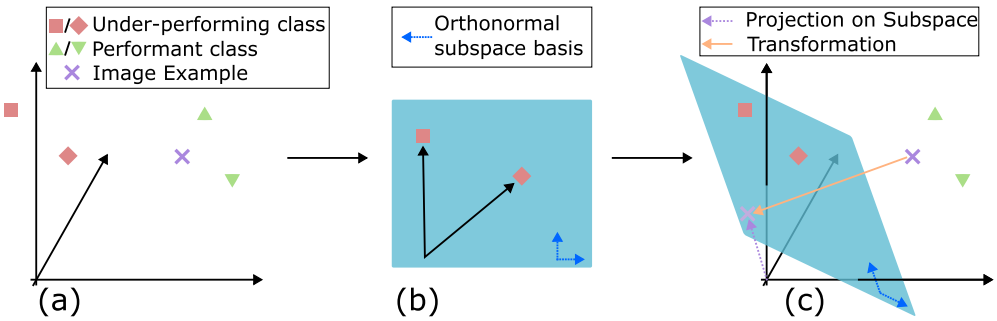

Figure 1: An overview of the Gram-Schmidt Subspace Projection Method. (a) given performant and under-performing classes, we wish to project image examples closer to under-performing classes to help with zero-shot classification. (b) First, the subspace is found using the Gram-Schmidt Process, defined as an orthonormal basis. (c) Image examples can be projected onto the subspace and the corresponding transformation applied.

3 METHODOLOGY

Our approach aims to improve the zero-shot classification performance of dual-encoder style models on under-performing classes. By treating the text embeddings of the class labels as defining geometric subspaces within the shared representation space, we manipulate the image embeddings prior to classification. This allow us to amplify the semantic signal of the under-performing classes while mitigating catastrophic forgetting of the broader class distribution. An overview of our approach is given in Figure 1.

3.1 PROBLEM FORMULATION AND BASELINE EVALUATION

In standard zero-shot classification using dual-encoder style models (e.g. CLIP), both images and class labels (i.e. captions/prompts) are first embedded using image and text encoders. The representations are then passed through a learned linear projection layer that maps these into a fixed D-dimensional space and l2-normalised. Once embedded, the image feature vector $x \in \mathbb{R}^D$ is classified by computing its cosine similarity against a set of text label embeddings $T = \{t_1, t_2, \ldots, t_C\}$, where C is the total number of classes and $t_i \in \mathbb{R}^D$ the embedding of the ith class. The predicted class $\hat{y}$ is given by:

$$\hat{y} = \underset{i \in \{1,\ldots,C\}}{\arg\max} (x \cdot t_i) \tag{1}$$

How are Under-Performing Classes defined? To identify under-performing classes, there are many potential methods, however, when assuming no knowledge of the test set, this can be found with a small labeled validation set; similarities/distances of the unlabeled images with the classes within the embedding space; or frequency of predictions on the unlabeled images given the set of classes T. We investigate two strategies for identifying under-performing classes. The original accuracy-based selection ranks classes by their baseline zero-shot classification accuracy over the full dataset and selects the lowest performing $N\%$. Because this ranking includes evaluation on images and their labels, it serves as an oracle reference rather than a selection strategy.

The similarity-based selection instead uses a separated calibration subset, constructed by reserving 10% of the images of the dataset. The remaining images form the evaluation set. We sample 5% of the calibration subset three times, corresponding to approximately 5% of the full dataset for each draw. Sampling is without replacement for each draw, although different draws may overlap. For each class in the dataset we calculate the average cosine similarity between the text label embedding and the sampled image embedding, then average these class-level scores across draws. Classes with the lowest averaged similarities make up the similarity-based Bottom $N\%$. Averaging across multiple draws reduces reliance on a single sampled subset, while keeping selection separate from evaluation. For either selection strategy, let $T_{UP} \subset T$ represent the text embeddings of these *Under-Performing classes*, and $T_P = T - T_{UP}$ represent the remaining, *Performant* classes. We evaluate both selections on the same held-out evaluation images and examine performance across a range of different values of $N\%$. Figure 1(a) shows this setting in a 3-dimensional version in which two under-performing and

two performant classes are shown with an image embedding. Our main results are reported on the similarity-based selection and results for the accuracy-based selection can be found in B.

3.2 Subspace Construction via Gram-Schmidt

The first step of our method is to find the subspace defined by the under-performing classes. Before constructing the basis, we compute the mean embedding across all C classes. This centering emphasizes deviations from the dataset-wide semantic mean. Let $v_i \in \mathbb{R}^D$ denote the class-name embedding of class i, and let $\mathcal{U}$ denote the indices of the selected Bottom $N\%$ classes:

$$\mu = \frac{1}{C} \sum_{i=1}^{C} v_i \tag{2}$$

We then center each selected class embedding:

$$v_i' = v_i - \mu \quad i \in \mathcal{U} \tag{3}$$

Mean centering is applied only during the text-basis construction, the image embedding itself is not centered.

Subtracting the dataset-wide mean allows basis to represent deviations of the selected classes from the shared class centroid. The mean is computed over all classes, rather than only the under-performing classes. As shown in Figure 1(b), we construct the orthonormal bases spanning the centered embeddings. Given the subset of under-performing classes and their respective text embeddings e.g., T_{UP}, we apply the Gram-Schmidt orthogonalization process to generate a basis matrix $B_{UP} \in \mathbb{R}^{|T_{UP}| \times D}$.

For a vector v_i' added to the basis, we iteratively remove its projection onto all previously established basis vectors b_j:

$$r_i = v_i' - \sum_j (v_i' \cdot b_j) b_j \tag{4}$$

The resulting vector is L2-normalized to get the new basis vector b_i. This yields orthonormal bases B_{UP}, which capture the purely geometric subspaces representing the target, under-performing classes and the non-target, performant classes.

3.3 Gram-Schmidt Subspace Projections

GSSP forces the image representation entirely into the subspace of the under-performing classes. As shown in Figure 1(c), by taking the dot product of the image feature with the bottom basis, we retain only the semantic features relative to T_{UP}:

$$\pi_{UP}(x) = \sum_{k=1}^{|T_{UP}|} (x \cdot b_k) b_k \tag{5}$$

We can get the final, normalized projection x_{proj} by applying L2 normalization:

$$x_{proj} = \frac{\pi_{UP}(x)}{\|\pi_{UP}(x)\|_2} \tag{6}$$

3.4 Parametrized Blending

While our method effectively boosts the performance of the Bottom $N\%$ classes, it inherently distorts information critical for recognizing the remaining classes, significantly damaging the overall accuracy. To resolve this, we introduce a parametrized blending method to blend our method with the original feature space. Specifically, we add a scaled version of the under-performing projected features back to the original image embedding, amplifying the under-performing classes performance without completely distorting the performant classes which we call GSSP-Boost:

$$x_{boost} = \frac{x + \lambda \cdot \pi_{UP}(x)}{\|x + \lambda \cdot \pi_{UP}(x)\|_2} \tag{7}$$

where λ is a hyperparameter that controls the degree of blending between the original feature space with the subspace calculated via GSSP. At $\lambda = 0$, it recovers the original normalized image embedding. As λ increases, the selected component receives greater relative weight; the representation approaches the normalized GSSP projection. The theoretical grounding for GSSP and GSSP-Boost, detailing their geometric effects can be found in A. A full ablation on the choice of λ and the trade-off between performance of the under-performing classes and the performant classes can be seen in 4.3.

3.5 GSSP-Boost Prompted

Radford et al. (2021) demonstrated that prompt engineering and prompt ensembling can improve zero-shot classification performance. Motivated by these findings, we incorporate richer textual representations into GSSP-Boost through a variant termed GSSP-Boost Prompted that combines our subspace with CuPL class descriptions Pratt et al. (2023). This variant follows the same GSSP-Boost procedure, while replacing the class labels with the averaged CuPL description embeddings for the under-performing classes and the boosted image embeddings are compared against averaged CuPL description embeddings for the selected under-performing classes, while the remaining classes retain their baseline text embeddings. This extension combines GSSP-Boost with richer class descriptions without modifying the encoders or introducing additional trainable parameters.

4 Results

In this section we first present the experimental settings in Sec. 4.1, before discussing the main results in Sec. 4.2. We next ablate $N\%$ and λ in Sec. 4.3 and finally present qualitative results in Sec. 4.4

4.1 Experimental Settings

Datasets To evaluate the robustness and generalizability of our subspace method, we perform our experiments across a suite of 10 image classification datasets that are commonly used to benchmark zero-shot performance. The datasets are Caltech101 Fei-Fei et al. (2004) (101 common object categories), OxfordPets Parkhi et al. (2012) (Fine-grained cat and dog breeds), StanfordCars Krause et al. (2013) (Fine-grained car makes, models and years), Flowers102 Nilsback & Zisserman (2008) (Fine-grained flower species), SUN397 Xiao et al. (2010) (Large-scale scene recognition), EuroSAT Helber et al. (2019) (Satellite imagery for land cover and use), and UCF101 Soomro et al. (2012) (Human action recognition in YouTube videos), Food101 Bossard et al. (2014) (101 popular food dishes), FGVC Aircraft Maji et al. (2013) (Fine-grained aircraft models and variants), and DTD Cimpoi et al. (2014) (Textual patterns).

Baselines We evaluate our approach using both zero-shot baselines and methods selected for their methodological similarity. Our CLIP-based comparison include IsoCLIP, Optimization Visual Inversion (OVI), and Optimization Textual Inversion (OTI). These baselines operate across different architectural insertion points and serve as diverse and robust comparators for our strictly post-projection method. We additionally compare against Bottom-N variants of uniform prompt ensembling (UPE) which averages the normalized dataset-specific prompt templates provided by Radford et al. (2021), CuPL and CALIP, applying their text-feature or score replacements only to the selected classes. As a simple methodological control baseline, we include Bottom $N\%$ weighted average, which adds a weighted, normalized centroid of the Bottom $N\%$ classes to each image embedding (WSB). Finally we implement GSSP, GSSP-Boost and GSSP-Boost Prompted on top of the Perception Encoder (PE) Bolya et al. (2026) and SigLIP2 Tschannen et al. (2025) and compare against their respective baselines to assess applicability beyond CLIP.

Implementation Details and Evaluation Metrics We conduct the experiments building on top of both CLIP (ViT-B/32) Perception Encoder PE (PE-Core-L14-336) and SigLIP2 (siglip2-base-patch16-256). For the majority of our experiments, we choose a value of 20% as the size of the under-performing classes, representing the Bottom $N\%$. This was done to account for the large disparity in class counts between different datasets, allowing for the selection of a range of under-performing classes. We later show results with $N\%$ ranging from 5% to 30%. We initially chose $\lambda = 0.5$ as it represents an equal weighting between the GSSP projection and the original embedding—we didn't want to tune this for different datasets/models given the zero-shot assumption. We then sweep through predefined grids for λ which controls the magnitude of the boosting blending method.

For each experiment, we report the Harmonic Mean between the under-performing class accuracy and the remaining classes accuracy as well as the accuracy for the under-performing classes. We report both metrics to track performance trade-off between the two groups of classes.

Final results and ablations used 4 $\times$ NVIDIA GH200 totaling ~6 hours of run time for evaluating ALL of our methods across ALL datasets. Due to the nature of the framework given the same labels, bases can be cached allowing for faster subsequent processing times.

4.2 Main Results

The main results can be seen in Table 1. As mentioned, our primary objective is to improve the average accuracy of the Bottom $N\%$ under-performing classes without severely compromising the model's overall predictive power represented by the harmonic mean.

Our initial baseline results highlight a common problem in the zero-shot models, performance on the under-performing classes is often very low. For instance, the baseline CLIP accuracy on the under-performing classes of DTD is 6.9% and FGVCAircraft is 4.4%. To isolate and address these problems, we first tested the standard GSSP method. While this approach successfully delivers a large boost to accuracy of the under-performing classes across all datasets, it completely collapses the overall capabilities of the model. The Harmonic Mean drops severely under this setting, demonstrating that raw projection destroys the model's generalized knowledge representation.

Our parameterized blending method GSSP-Boost mitigates the catastrophic forgetting, across all models. CLIP GSSP-Boost significantly improves the under-performing classes while retaining competitive overall accuracy leading to similar values for the Harmonic Mean. With respect to models with better base performance, such as PE and SigLIP2, we see that GSSP-Boost is able to increase performance over both the Harmonic Mean and under-performing classes significantly across all 10 datasets. As an example, Flowers102 sees an overall harmonic mean accuracy increase from 46.3$\%$ to 60.5$\%$ whilst also increasing the performance of under-performing classes greatly from 30.3$\%$ to 44.7$\%$.

Comparing other zero-shot baselines, GSSP-Boost performs competitively with complex baselines like CALIP (which requires 2D grid searches over spatial attention weights) or significant external data such as CuPL (which requires generating hundreds of auxiliary LLM descriptions per dataset) and UPE (which requires handcrafted prompt ensembles). Furthermore, GSSP-BP demonstrates that our geometric subspace framework can seamlessly incorporate this extra descriptive information by compounding image subspace boosting with CuPL's LLM descriptions for the bottom classes. Rather than plateauing, GSSP-BP consistently outperforms pure CuPL across all backbones and outpaces all other zero-shot, prompt-tuning and cross-attention methods across the benchmarks, proving that our geometric approach not only competes effectively with data and tuning heavy baselines on its own but also readily ingests auxiliary data to further improve its performance. Finally, both GSSP-B and GSSP-BP consistently outperforms IsoCLIP, OVI and OTI across all evaluated metrics in all datasets, but especially improves under-performing classes.

4.3 Ablations

In this section, we investigate the influence of the hyperparameter λ as well as the percentage of under-performing classes, $N\%$ on the GSSP-Boost method.

How to Choose λ?

The parameter λ controls the intensity of the boosting applied to the under-performing classes and we sweep λ across values of [0.10, 0.25, 0.50, 0.75, 1.00], the results of which can be seen in Figure 2. We assess this balance using the harmonic mean, which rewards strong performance across both performant and under-performing classes. Although the changes across the sweep are small, they demonstrate that λ doesn't have a large effect and provides control over their trade-off (see additional results in Figure 5 in Supp). Based on our empirical sweep, we identify $\lambda = 0.5$ as a good value regardless of dataset and model. At $\lambda = 0.5$, GSSP-Boost displays substantial and meaningful gains for the Bottom N$\%$ under-performing classes. This flexibility distinguishes GSSP-Boost from zero-shot methods that lack a mechanism to control this trade-off to suit application needs.

Table 1: Zero-shot performance comparison across all datasets. GSSP-B refers to our full GSSP approach with boosting and GSSP-BP refers to the GSSP-Boosting Prompted approach, both approaches are reported for $\lambda = 0.5$ and Bottom $N\% = 20$. We report Harmonic Mean (HM) between the under-performing classes and the performant classes and accuracy over the under-performing classes (T_{UP}). We highlight **best** and runner-up results.

	Caltech101		OxfordPets		StanfordCars		Flowers102		SUN397		EuroSAT		UCF101		Food101		FGVCAircraft		DTD	
	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}
CLIP	83.7	77.7	72.3	61.3	43.1	32.2	24.0	14.3	49.5	39.0	27.2	21.3	44.4	32.7	73.1	65.8	7.4	4.4	12.1	6.9
OVI	73.3	68.0	60.1	57.5	31.0	24.8	23.3	15.7	33.6	27.5	28.6	24.5	26.3	17.6	56.7	50.0	0.1	0.1	2.7	1.4
OTI	77.9	69.9	73.1	65.8	37.3	27.2	26.4	16.2	40.5	30.3	13.1	7.6	45.6	35.3	71.7	66.1	9.7	6.4	12.4	7.1
IsoCLIP	78.8	72.5	74.6	70.1	40.9	32.5	23.2	14.4	36.7	28.5	16.7	12.0	42.0	31.9	69.2	63.0	5.5	3.5	12.6	7.4
CALIP	85.4	85.2	84.9	84.4	49.7	41.3	40.5	28.4	57.4	51.2	0.8	**54.4**	57.3	55.7	**78.5**	76.7	17.7	17.4	29.7	22.3
CuPL	**86.5**	83.9	**88.8**	93.5	**55.2**	57.8	48.2	38.8	58.9	54.5	**40.2**	46.4	**62.2**	66.4	77.5	74.5	**21.3**	24.1	42.9	40.7
UPE	83.5	77.7	83.3	80.3	48.5	39.3	39.4	27.1	51.8	42.0	2.7	35.5	57.1	54.5	**78.1**	75.8	16.8	15.0	26.6	18.8
CLIP-WSB	34.4	59.1	4.3	69.4	14.0	15.3	3.0	15.1	17.1	25.2	24.1	51.4	27.0	24.9	32.7	59.3	5.2	6.6	11.7	20.9
CLIP-GSSP	2.2	86.1	1.3	83.8	32.4	51.2	0.3	29.5	31.1	60.2	16.9	47.8	2.7	56.3	7.9	81.6	5.6	17.6	1.0	38.1
CLIP-GSSP-B	85.7	82.6	81.2	78.2	48.0	39.7	35.6	24.2	55.4	48.5	30.7	26.6	52.7	46.3	**78.0**	77.0	10.6	7.3	27.0	18.7
CLIP-GSSP-BP	85.0	**86.9**	83.3	**96.7**	52.9	**60.3**	38.2	**41.1**	**60.1**	**62.0**	**39.7**	48.4	56.0	**73.4**	**78.2**	**82.5**	20.0	**33.1**	**44.5**	**52.2**
PE	**93.2**	89.1	93.5	90.7	91.8	89.1	46.3	30.3	69.8	60.6	7.9	4.2	78.2	69.4	92.6	89.8	47.8	35.1	47.0	33.7
PE-CuPL	92.9	88.4	93.2	90.4	**92.0**	89.9	68.6	53.5	69.9	61.0	21.5	12.9	83.4	78.4	91.0	86.8	60.2	50.6	61.9	52.8
PE-UPE	**93.2**	89.0	94.3	92.4	91.7	88.9	57.2	40.5	70.6	62.0	26.4	**95.9**	81.9	75.7	93.1	91.1	62.8	55.7	58.2	48.9
PE-WSB	79.6	86.9	55.4	83.4	78.6	75.2	32.0	26.9	46.4	45.2	0.5	63.2	60.9	56.2	68.8	82.6	17.0	25.3	42.2	44.9
PE-GSSP	0.8	**90.4**	13.0	**96.2**	53.1	**92.8**	0.0	59.9	27.2	**76.1**	20.8	61.7	10.6	**84.7**	4.8	**95.6**	12.1	59.0	1.1	**72.3**
PE-GSSP-B	**93.6**	**90.1**	**95.1**	95.0	**92.4**	90.7	60.5	44.7	**73.8**	69.3	14.6	8.3	82.3	77.2	**94.1**	94.0	55.2	44.7	60.9	52.3
PE-GSSP-BP	**93.2**	89.8	93.7	93.4	**92.0**	90.5	**73.5**	**63.1**	**73.4**	69.6	**40.0**	29.0	**84.5**	82.7	**93.7**	93.0	**69.0**	**66.1**	**65.5**	62.9
SigLIP2	92.8	88.6	92.4	90.6	91.8	90.3	44.9	29.4	61.8	51.0	0.1	0.1	53.6	39.9	86.7	82.2	0.0	0.0	46.5	35.2
SigLIP2-CuPL	91.0	85.2	89.1	83.9	91.2	89.2	52.6	36.4	54.4	41.3	3.0	1.5	**71.5**	67.8	85.3	79.6	**41.5**	40.1	52.0	43.1
SigLIP2-UPE	91.2	85.6	91.9	89.5	92.3	91.3	58.2	42.3	62.5	51.9	**20.5**	**63.3**	61.1	49.3	**88.8**	86.2	6.8	3.7	57.1	55.0
SigLIP2-WSB	80.4	89.9	62.6	68.5	71.3	73.3	23.2	16.0	36.5	35.8	6.9	35.7	35.1	25.0	57.6	69.7	2.4	3.2	47.0	44.1
SigLIP2-GSSP	2.6	**90.7**	4.2	**97.8**	68.7	**96.7**	2.1	45.5	30.4	**67.7**	10.2	28.2	0.2	69.7	2.6	**93.5**	1.7	5.2	0.5	**59.3**
SigLIP2-GSSP-B	**93.0**	90.0	**93.1**	95.8	**93.1**	94.1	50.2	35.2	**66.4**	60.1	0.4	0.2	67.0	58.4	**89.3**	88.6	0.1	0.1	**57.9**	51.4
SigLIP2-GSSP-BP	**93.0**	89.4	92.2	91.7	92.4	92.6	**67.1**	**54.3**	**66.6**	59.8	3.8	2.0	70.0	**72.7**	**89.1**	87.6	40.6	**53.2**	**58.2**	55.9

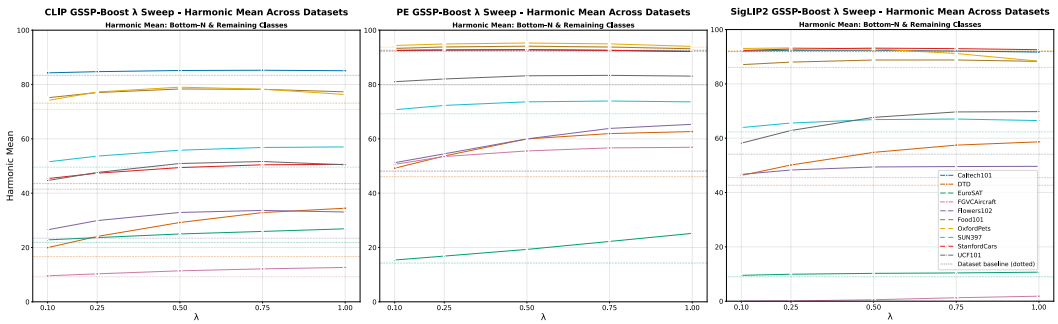

Figure 2: λ ablation (CLIP/PE/SigLIP2) showing trade-off between Bottom N% under-performing classes and performant classes through the harmonic mean. Dashed lines indicate baseline.

How Many Under-Performing Classes? To evaluate the effectiveness of GSSP-Boost across various percentages of under-performing classes, we analyzed the change in mean accuracy relative to the baseline for a range of $N\%$ performing classes. Figure 3 shows this analysis conducted across all architectures as N increases from 5 to 30%.

Across all configurations and the vast majority of datasets, there is a clear inverse relationship between $N\%$ and the relative performance boost. The method yields the most substantial accuracy improvements for the absolute lowest-performing classes, e.g. N = 10%. This is also proportional to the number of classes associated with each dataset i.e if a dataset has a small number of classes a bigger increase will be achieved at a higher $N\%$ compared to a dataset with a much larger number of classes. As N increases to include comparatively easier classes, the average steadily declines.

4.4 Qualitative Results

To provide deeper intuition into the performance gains achieved by our approach, we present a qualitative comparison between the predictions of base CLIP and GSSP-Boost. Fig. 4 illustrates several examples from Caltech101 and Food101 datasets. We see that CLIP struggles to determine differences between fine-grained classes with high visual similarities such as "Bruschetta"/"Pizza" and "Stapler/"Camera". In other cases, CLIP misclassifies images which have high(er) visual differences to other images within their classes, e.g. "Chandelier" and "Lamp". GSSP-Boost can improve

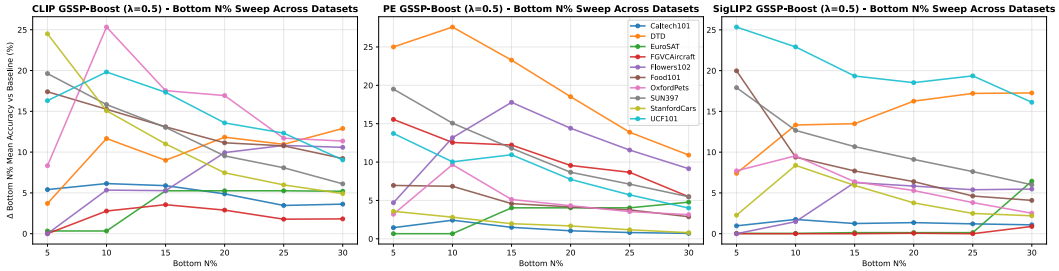

Figure 3: Ablation of N% of under-performing classes for GSSP-Boost across CLIP, PE and SigLIP2.



Figure 4: Qualitative results of GSSP-Boost compared to the base CLIP implementation highlighting three success cases and a failure case. Correct classes are shown bold in green alongside the classification rank change for each class.

performance on these under-performing classes, though can still suffer from visual confusion across classes, i.e. in the last example of "Wild Cat"/"Leopards".

5 Conclusion

In this work, we introduced Gram-Schmidt Subspace Projection (GSSP) and its parameterized extensions, GSSP-Boost and GSSP-Boost Prompted, as a training-free, computationally lightweight framework designed to address the severe performance disparities in zero-shot image classification. While current foundational models exhibit strong generalized capabilities, our baseline analysis highlighted a critical flaw where performance on under-performing classes often is worse than the performant classes. To solve this, our geometric approach directly manipulates the post-projection embedding space constructing orthonormal bases from text embeddings of these under-performing classes to apply targeted transformations to the visual features.

Our comprehensive empirical evaluation across a diverse suite of datasets demonstrates the effectiveness of this approach. While standard GSSP delivers massive accuracy improvements for the under-performing classes, it suffers from catastrophic forgetting, severely degrading the model's overall predictive power. GSSP-Boost successfully resolved this by introducing a blending parameter λ, to regularize the transformation. Through systematic ablations, we identified $\lambda = 0.5$ as the optimal configuration. At this balanced setting, the method delivers substantial gains for under-performing classes while preserving the zero-shot model's broader, generalized knowledge. To further resolve fine-grained visual confusions and demonstrate the framework's extensibility, we developed GSSP-BP, which couples image subspace boosting with rich, LLM-generated descriptions (CuPL) for under-performing classes. This hybrid formulation proves that GSSP can readily absorb auxiliary semantic data, simultaneously refining both visual representations and textual decision boundaries to unlock compound performance gains that neither visual boosting nor prompt engineering can achieve in isolation.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15292–15301. IEEE, 2023.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in clip. In *The Thirteenth International Conference on Learning Representations*, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

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.

Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 746–754, 2023.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Yusuke Hirota, Ryo Hachiuma, Boyi Li, Ximing Lu, Michael Ross Boone, Boris Ivanovic, Yejin Choi, Marco Pavone, Yu-Chiang Frank Wang, Noa Garcia, et al. Bias in gender bias benchmarks: How spurious features distort evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8634–8644, 2025.

Mingyu Jeon, Sunjae Yoon, Jonghee Kim, and Junyeong Kim. Granalign: Granularity-aware alignment framework for zero-shot video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5287–5295, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc 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*, pp. 4904–4916. PMLR, 2021.

Lie Ju, Sijin Zhou, Yukun Zhou, Huimin Lu, Zhuoting Zhu, Pearse A Keane, and Zongyuan Ge. Delving into out-of-distribution detection with medical vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 133–143. Springer, 2025.

Ildoo Kim, Younghoon Kim, and Sungwoong Kim. Learning loss for test-time augmentation. *Advances in neural information processing systems*, 33:4163–4174, 2020.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12289–12301, 2024.

Simone Magistri, Dipam Goswami, Marco Mistretta, Bartłomiej Twardowski, Joost van de Weijer, and Andrew D. Bagdanov. Isoclip: Decomposing clip projectors for efficient intra-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does clip's generalization performance mainly stem from high train-test similarity? In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 44153–44187, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/c089f19038fba90268b46b316a14cf00-Paper-Conference.pdf.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *ICLR*, 2023.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D. Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VVVfuIcmKR.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Changdae Oh, Hyesu Lim, Mijoo Kim, Dongyoon Han, Sangdoo Yun, Jaegul Choo, Alexander Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models. *Advances in Neural Information Processing Systems*, 37:12677–12707, 2024.

Bo Pang, Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Cabin: Debiasing vision-language models using backdoor adjustments. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 484–492, 2025.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15691–15701, 2023.

Qi Qian, Yuanhong Xu, and Juhua Hu. Intra-modal proxy learning for zero-shot visual categorization with clip. *Advances in Neural Information Processing Systems*, 36:25461–25474, 2023.

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.

Karsten Roth, Jae Myung Kim, Andrew Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15746–15757, 2023.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1214–1223, 2021.

Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in CLIP. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4256–4266, 2025.

Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23783–23793, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38629–38642. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/fc28053a08f59fccb48b11f2e31e81c7-Paper-Conference.pdf.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022b.

Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kIP0duasBb.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Selective vision-language subspace projection for few-shot clip. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3848–3857, 2024.

AI USE STATEMENT

In this work, we used generative AI tools for code completion only. All AI-assisted code was reviewed and tested. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

APPENDICES

This appendix provides further analysis of GSSP and GSSP-Boost. We first present their theretical grounding, describing the geomteric effects of subspace projection and boosting and the relationship between the two methods. We then report results using oracle Bottom N% selection, which identifies the under-performing classes from their baseline per-class accuracies. Finally, we examine how the boosting parameter λ controls the trade-off between overall accuracy and the accuracy of the under-performing classes.

A THEORETICAL GROUNDING OF PROJECTION AND BOOSTING

Let $x \in \mathbb{R}^D$ be an L2-normalized image embedding and $t_j \in \mathbb{R}^D$ the normalized text embedding for class j, with zero-shot prediction $\hat{y} = \arg\max_j x^\top t_j$.

Subspace construction. Let $\mathcal{U}$ be the selected under-performing class indices and v_i the normalized class-name embedding of class i. We define

$$\mu = \frac{1}{C}\sum_{i=1}^{C} v_i, \qquad \mathcal{S}_{UP} = \operatorname{span}\{v_i - \mu : i \in \mathcal{U}\}, \tag{8}$$

(an uncentered variant drops μ). Let $B_{UP} \in \mathbb{R}^{r\times D}$ hold an orthonormal basis for $\mathcal{S}_{UP}$, where $r = \dim(\mathcal{S}_{UP}) \le \min(|\mathcal{U}|, D)$ is the number of linearly independent basis vectors retained after Gram–Schmidt orthogonalization. The corresponding projector is $P_{UP} = B_{UP}^\top B_{UP}$, with $x = x_\parallel + x_\perp$ where $x_\parallel = P_{UP}x$. This subspace is text-derived and applied to image embeddings without centering x itself.

Projection (GSSP). GSSP retains only the in-subspace component:

$$x_{\text{proj}} = \frac{P_{UP}x}{\|P_{UP}x\|_2}, \qquad P_{UP}x \neq 0. \tag{9}$$

Since normalization is a positive per-image scalar,

$$\hat{y}_{\text{proj}} = \arg\max_j x^\top P_{UP} t_j, \tag{10}$$

i.e. projecting the image is equivalent to scoring against projected, unnormalized text embeddings $P_{UP}t_j$. Components orthogonal to $\mathcal{S}_{UP}$ are discarded entirely, which helps only if that subspace holds the discriminative evidence.

Boosting (GSSP-Boost). Rather than discarding $x_\perp$, boosting rescales the in-subspace component:

$$x_{\text{boost}}(\lambda) = \frac{x + \lambda P_{UP}x}{\|x + \lambda P_{UP}x\|_2} = \frac{(1+\lambda)x_\parallel + x_\perp}{\sqrt{(1+\lambda)^2\|x_\parallel\|_2^2 + \|x_\perp\|_2^2}}, \quad \lambda \ge 0. \tag{11}$$

The pre-normalization map $I + \lambda P_{UP}$ scales $\mathcal{S}_{UP}$ directions by $1+\lambda$ and leaves $x_\perp$ unchanged, so

$$\frac{\|(I - P_{UP})x_{\text{boost}}(\lambda)\|_2}{\|P_{UP}x_{\text{boost}}(\lambda)\|_2} = \frac{1}{1+\lambda}\frac{\|x_\perp\|_2}{\|x_\parallel\|_2}. \tag{12}$$

Larger λ increases the subspace's relative weight without eliminating $x_\perp$ at any finite λ.

Projection as a limit. Projection and boosting form one family: $x_{\text{boost}}(0) = x$ and $\lim_{\lambda\to\infty} x_{\text{boost}}(\lambda) = x_{\text{proj}}$. With $\alpha = \lambda/(1+\lambda) \in [0, 1)$,

$$x_{\text{boost}}(\lambda) = \frac{(1-\alpha)x + \alpha P_{UP}x}{\|(1-\alpha)x + \alpha P_{UP}x\|_2}, \tag{13}$$

interpolating between the original embedding and its unnormalized projection; GSSP is the $\alpha \to 1$ endpoint, and GSSP-Boost controls the degree of subspace emphasis via λ. Both methods rest on the same hypothesis that directions spanned by under-performing classes carry discriminative signal worth up-weighting which is tested empirically rather than guaranteed by the construction.

B ORACLE BOTTOM N% RESULTS

We additionally evaluate all our methods and baselines using the accuracy-based Bottom N%. We refer to this as the *oracle selection*, since it uses class-level performance information to identify the true under-performing classes for each model. The under-performing accuracies reported for these classes are generally lower than those obtained with similarity-based selection. This is expected as the oracle selection explicitly targets the classes on which the baseline performs worst, while similarity-based selection may include classes with higher baseline accuracy. The lower accuracies therefore reflect differences in selection and should be interpreted relative to their corresponding baselines. The same qualitative performance trends across methods remain evident for CLIP, PE and SigLIP2. The oracle results thus provide further evidence that the observed behavior of projection and boosting is consistent across the evaluated models and is not specific to a class selection method.

Table 2: Zero-shot performance comparison across all datasets. GSSP-B refers to our full GSSP approach with boosting and GSSP-BP refers to the GSSP-Boosting Prompted approach, both approaches are reported for $\lambda = 0.5$ and Bottom $N\% = 20$. We report Harmonic Mean (HM) between the under-performing classes and the performant classes and accuracy over the under-performing classes (T_{UP}). We highlight **best** and runner-up results.

	Caltech101		OxfordPets		StanfordCars		Flowers102		SUN397		EuroSAT		UCF101		Food101		FGVCAircraft		DTD	
	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}	HM	T_{UP}
CLIP	71.3	56.8	62.1	47.0	24.2	14.7	5.1	2.7	35.7	23.8	1.2	0.6	14.6	8.1	69.8	59.8	0.2	0.1	9.0	4.9
OVI	54.9	40.3	51.8	42.5	21.9	14.6	10.5	5.9	28.0	20.3	0.0	0.0	12.6	7.2	52.5	43.0	0.1	0.1	13.0	8.7
OTI	65.4	50.5	64.6	52.0	25.0	15.7	15.2	8.5	31.6	21.0	31.7	24.6	33.7	22.4	68.3	59.7	4.9	2.8	12.4	7.1
IsoCLIP	66.1	51.8	67.0	56.1	27.7	18.1	3.0	1.5	33.9	25.1	12.5	8.0	19.5	11.5	65.6	56.3	4.4	2.7	16.4	10.2
CALIP	79.5	69.5	80.4	74.4	39.1	27.9	29.2	18.3	45.2	33.4	1.0	**56.8**	54.3	48.7	**76.7**	72.6	13.0	9.8	32.2	25.6
CuPL	**82.6**	75.9	**83.5**	79.4	**51.4**	50.5	**44.2**	33.9	55.9	48.4	33.9	29.5	**56.9**	53.2	74.5	68.4	18.8	17.6	42.1	39.0
UPE	76.6	64.7	77.4	68.3	35.6	24.3	26.4	16.1	39.4	27.2	1.7	**56.7**	53.3	46.3	76.2	71.4	11.8	8.3	27.2	19.3
CLIP-WSB	35.3	35.1	14.3	47.8	8.1	5.4	3.5	5.0	16.3	17.4	0.1	0.0	20.5	16.1	33.4	36.1	4.7	4.4	11.5	21.1
CLIP-GSSP	2.3	79.5	1.8	62.4	31.0	39.5	0.4	17.4	28.4	57.5	18.1	37.6	2.8	46.1	4.5	76.6	5.9	14.1	3.1	38.4
CLIP-GSSP-B	79.0	69.8	75.3	67.5	34.4	23.3	19.7	11.5	47.6	37.0	1.6	0.8	32.6	21.7	**76.7**	73.8	1.1	0.6	25.8	17.7
CLIP-GSSP-BP	**82.5**	**81.8**	82.7	**87.1**	50.4	**54.0**	35.6	**36.5**	**59.1**	**59.0**	**35.4**	33.1	53.2	**58.9**	76.2	**78.7**	**19.6**	**27.3**	**44.9**	**51.9**
PE	89.0	80.3	90.0	83.0	84.4	73.7	39.5	24.7	60.5	46.7	0.1	0.0	62.0	45.9	91.5	87.4	25.3	15.0	31.7	19.8
PE-CuPL	88.8	80.2	91.8	87.0	85.5	75.9	64.5	48.4	64.9	52.5	1.0	0.5	75.5	63.8	89.6	83.8	49.6	36.2	57.0	45.0
PE-UPE	88.4	79.3	91.2	85.2	84.6	74.1	51.8	35.2	61.9	48.6	20.4	70.1	71.2	57.0	91.9	88.2	54.8	43.2	55.0	43.3
PE-WSB	72.7	61.5	72.0	91.4	72.7	63.5	32.1	23.8	44.3	38.3	0.0	50.0	45.2	32.7	71.5	69.0	18.5	16.8	41.7	42.8
PE-GSSP	4.8	**83.4**	9.0	92.3	50.9	**77.8**	0.0	58.8	28.3	**68.1**	**23.9**	**85.2**	9.8	74.7	5.7	**93.4**	11.6	**48.9**	3.8	**72.5**
PE-GSSP-B	**89.8**	81.7	**93.6**	91.3	**85.6**	76.0	60.9	44.8	67.5	57.6	1.8	0.9	75.9	65.8	**93.1**	91.6	40.6	27.8	49.5	36.5
PE-GSSP-BP	**89.3**	81.6	92.9	**93.3**	**85.7**	76.4	**71.2**	**59.4**	**70.3**	62.6	3.1	1.6	**79.8**	**75.3**	92.5	90.1	**59.2**	**49.3**	**62.4**	55.5
SigLIP2	87.6	78.0	88.2	81.1	83.2	72.3	32.8	19.7	52.6	38.8	0.1	0.1	31.6	19.3	83.9	76.3	0.0	0.0	24.6	14.8
SigLIP2-CuPL	82.8	70.7	84.6	75.2	78.2	65.2	47.2	31.2	44.8	30.8	3.0	1.5	**67.9**	60.9	82.3	73.5	41.1	39.5	52.3	41.8
SigLIP2-UPE	82.9	70.8	87.4	80.0	83.3	72.6	50.2	34.0	52.8	39.1	**20.5**	**63.3**	49.3	34.8	86.6	81.0	15.9	9.8	51.9	43.1
SigLIP2-WSB	68.7	56.0	57.4	60.4	68.0	60.5	19.0	12.0	32.8	27.2	6.9	35.7	25.7	16.3	54.0	47.5	0.3	0.2	29.2	19.5
SigLIP2-GSSP	6.3	**84.6**	23.3	84.3	69.1	**76.0**	2.6	44.8	30.2	**61.5**	10.2	28.2	2.7	53.8	2.1	**88.0**	3.3	12.3	8.1	38.4
SigLIP2-GSSP-B	**89.8**	81.9	**89.3**	**85.2**	**84.6**	75.0	48.3	33.0	**59.7**	48.6	0.4	0.2	54.9	41.1	**87.1**	83.7	4.0	2.1	36.5	24.6
SigLIP2-GSSP-BP	**88.9**	80.2	**89.1**	84.2	81.2	70.1	**68.8**	**55.6**	59.1	47.5	3.8	2.0	**67.7**	**67.3**	**86.7**	82.3	**42.1**	**51.4**	**59.9**	**55.8**

B.1 λ TRADE-OFF CONTROL

The parameter λ controls the intensity of the boosting applied to the under-performing classes and we sweep λ across values of [0.10, 0.25, 0.50, 0.75, 1.00] evaluating the trade-off between the mean Bottom $N\%$ Accuracy and the Overall Macro Accuracy. Figure 5 shows that λ acts as a dial for the trade-off between Bottom N% performance and Overall accuracy. Based on our empirical sweep, we identify λ = 0.5 as the optimal value for balancing this trade-off regardless of dataset. At λ = 0.5, GSSP-Boost displays substantial and meaningful gains for the Bottom N% under-performing classes. Crucially, it achieves this without resulting in a significant drop in overall accuracy – improving the performance of the under-performing classes without compromising the foundational knowledge of the majority classes. Interestingly, we observe remarkable robustness to the performance trade-off. Even as λ increases, the overall accuracy across most datasets remains highly stable.

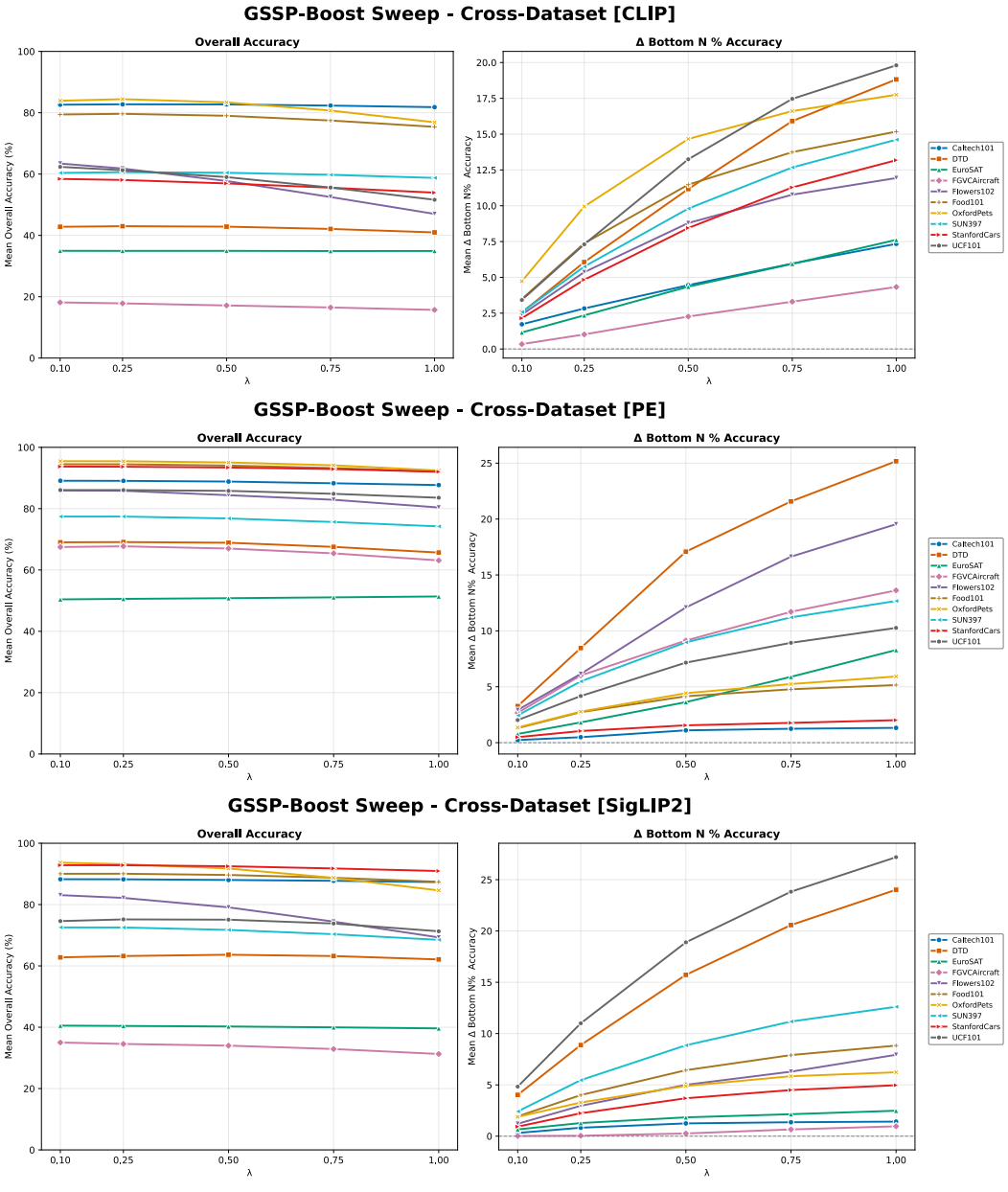

Figure 5: λ ablation (CLIP/PE/SigLIP2) showing trade-off between Bottom N% under-performing classes Δ acc. and overall acc. preservation.