PROJECT AND MIX: TASK-SEMANTIC PROTOTYPES FOR FEW-SHOT IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models like CLIP are trained with the objective of aligning text and image pairs. Beyond text prompts alone, recent works show that exploiting few-shot image embeddings from a training set is effective for CLIP-based classification. In this work, we analyze mixing image and text prototypes from a bias-variance perspective and show that mixing prototypes acts like a variance shrinkage estimator. Naively mixing text and image prototypes combines two partially aligned spaces since the two modalities are not perfectly aligned. To address this, we project image prototypes onto the principal directions of the semantic text embedding space to obtain a task-semantic image subspace. Mixing the image prototypes with text embeddings in the task-semantic subspace improves few-shot classification. However, when the task-semantic subspace captures insufficient discriminative visual information, relying on this subspace alone can be suboptimal. On extensive experiments over several few-shot classification benchmarks, we show that combining a task-semantic mixed prototype classifier and an anisotropic image-specific classifier systematically outperforms existing methods.

1 INTRODUCTION

Vision-language models (VLMs) like CLIP (Radford et al., 2021) are trained with a symmetric contrastive loss that maximizes the similarity between an image and its corresponding text while pushing it away from negative texts, and vice versa from text to image. This produces a shared embedding space in which semantically corresponding image and text representations are geometrically aligned. As a result, CLIP achieves strong zero-shot performance, where classification is performed using the cosine similarities between the test image representations and the text prompts encoded by the text encoder (text prototypes). However, the text and image embeddings in CLIP are not fully aligned and occupy different regions of the shared embedding space separated by a modality gap (Liang et al., 2022; Schrodi et al., 2025; Levi & Gilboa, 2025).

Multimodal Large Language Models (MLLMs) (Liu et al., 2023) despite strong performance for generative tasks, struggle for discriminative tasks (Zhang et al., 2024; Feng et al., 2026), making CLIP-based models more relevant for few-shot classification task. Several works explore the few-shot setting with CLIP in which only a few training images per class are available and show that the use of image embeddings in addition to text embeddings improves classification performance. These include training-free approaches (Zhang et al., 2022; Wang et al., 2024; Bendou et al., 2025) that combine image and text embeddings using different formulations, as well as methods that adapt the VLM by learning prompts (Khattak et al., 2023; Zhou et al., 2022a) or adapters (Silva-Rodriguez et al., 2024; Zanella & Ben Ayed, 2024). In this paper, we analyze training-free mixed prototype-based classification and show that mixing shrinks the high variance of the few-shot image prototypes thereby improving image prototype estimation by controlling the bias-variance tradeoff.

Although naive mixing is effective, the image and text representations are not perfectly aligned (Liang et al., 2022), resulting in the mixing of two misaligned prototypes. The image modality generally contains more detailed information such as background or instance-specific attributes such as object color which are not available in the text prompts, leading to the information imbalance between modalities (Schrodi et al., 2025). We therefore ask the question: *Can we use the class semantics to disentangle task-relevant image information from other visual variations for reliable few-shot classification?*

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.

We illustrate in Figure 1 that the image space could be exploited differently based on the classification task defined by the class labels. We argue that the space spanned by the text prototypes defines a relevant task-semantic space. We exploit the cross-modal alignment in CLIP and decompose image prototypes into two components: a *task-semantic subspace* to represent only the text label-specific information and a *task-orthogonal subspace* to represent the remaining attributes and instance-specific details. We show that mixing text and image prototypes in the task-semantic subspace improves classification performance over naively mixed prototype classifiers.

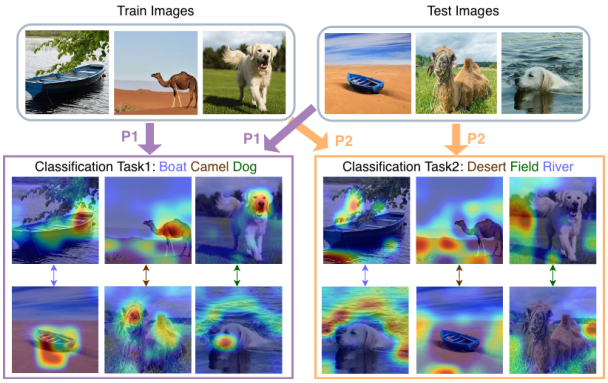

Figure 1: The latent space spanned by text prototypes defines a *task-semantic subspace*. Task semantic regions denoted by attention maps are obtained by the proposed task-specific semantic projection P1 and P2.

However, in out-of-distribution (OOD) datasets, the text-based task-semantic subspace poorly reflects the task relevant structure of the image space due to the poor cross-modal alignment in CLIP. To tackle this, we combine the mixed prototype classifier in the task-semantic subspace with a separate image classifier. We summarize our contributions as follows:

- We analyze the mixing of image and text prototypes in CLIP from a bias–variance perspective for few-shot classification. We show theoretically and empirically (Figure 2) that mixing acts as a shrinkage estimator, trading bias for variance reduction and yielding lower mean squared error than image-only prototypes.
- We use the class text prototypes to define a task-semantic subspace and decompose image prototypes into task-semantic and orthogonal components. Our analysis (Figure 3) shows that the task-semantic subspace can concentrate discriminative visual information, while its effectiveness varies across tasks. Performing prototype mixing in this subspace improves prototype estimation and classification performance.
- We complement the task-semantic classifier with an image-specific classifier that captures discriminative information beyond the task-semantic subspace. Across 15 diverse benchmarks, the proposed training-free method outperforms existing approaches from 1 to 16 shots, with particularly strong gains in the 1–4 shot regime, and generalizes across multiple contrastive vision-language backbones.

2 Related Work

Vision-Language Models. VLMs learn joint image-text representations through large-scale contrastive pre-training. CLIP (Radford et al., 2021) optimizes a symmetric contrastive loss over image-text pairs, producing a shared embedding space that enables strong zero-shot classification and other applications (Baldrati et al., 2022; Sain et al., 2023; Yu et al., 2023; Caselli et al., 2026). Despite this alignment objective, Liang et al. (2022) showed that image and text embeddings occupy distinct regions of the shared space separated by a *modality gap*. Recent works have attempted to close or exploit this gap: Eslami & de Melo (2025) proposed methods to improve cross-modal alignment in CLIP, while Mistretta et al. (2025) exposed intra-modal misalignment via modality inversion. Yu et al. (2024) used text-guided attention to improve zero-shot robustness in CLIP.

Few-Shot Classification. Among training-free methods, Tip-Adapter (Zhang et al., 2022) constructs a key-value cache from few-shot image features and retrieves them at test time to augment zero-shot predictions. TIP-X (Udandarao et al., 2023) extends this idea by populating the cache with support images retrieved or generated from class names. APE (Zhu et al., 2023) refines the cache by selecting only the most informative features via adaptive prior refinement. GDA (Wang et al., 2024) proposed an ensemble of a zero-shot text classifier with an LDA image classifier. ProKeR (Bendou et al., 2025) improves Tip-Adapter using a global regularizer in a reproducing kernel Hilbert space.

Training-based prompt learning methods (Zhou et al., 2022b;a) adapt CLIP by optimizing prompt tokens. MaPLe Khattak et al. (2023) extends this to multi-modal prompt learning across both encoders. Adapter-based methods tune lightweight modules: CLIP-Adapter (Gao et al., 2024) adds residual feature adapters, MMA (Yang et al., 2024) introduces multi-modal adapters, and CLIP-LoRA (Zanella & Ben Ayed, 2024) applies low-rank adaptation to CLIP's encoders. In this work, we propose a training-free method that combines a task-semantic mixed prototype classifier with an image-based classifier. The proposed method can also be applied on top of training-based methods.

3 Cross-Modal Prototype Mixing via Bias–Variance Analysis

A simple training-free method based on the CLIP vision encoder is the **Nearest Class Mean (NCM)** classifier, which estimates class prototypes as the empirical mean of training image features and assigns test samples to the nearest prototype. While effective, NCM performance depends on the quality of estimated visual prototypes. Here we analyze prototype estimation and introduce a **mixed prototype estimator** that improves estimation by controlling the bias–variance tradeoff.

3.1 Image Prototype Estimation in the Few-Shot Regime

Let μ_i^* be the true population image class mean for a given class, and $\hat{\mu}_i$ the empirical sample mean computed from n samples used as the class prototype by the NCM classifier, with $\mathbb{E}[\hat{\mu}_i] = \mu_i^*$, $\mathrm{Cov}[\hat{\mu}_i] = \frac{\Sigma_i^*}{n}$, where Σ_i^* is the population covariance of the class. Recall that, for a generic estimator $\hat{\theta}$ of a parameter θ^*, the Mean Squared Error (MSE) is defined as:

$$\mathrm{MSE}(\hat{\theta}, \theta^*) = \mathrm{Bias}^2(\hat{\theta}, \theta^*) + \mathrm{Var}(\hat{\theta}). \tag{1}$$

The sample mean is unbiased, thus $\mathrm{Bias}(\hat{\mu}_i, \mu_i^*) = \mathbb{E}[\hat{\mu}_i] - \mu_i^* = 0$. Therefore the MSE of the image mean estimator reduces to the variance term:

$$\mathrm{MSE}(\hat{\mu}_i, \mu_i^*) = \|\mathbb{E}[\hat{\mu}_i] - \mu_i^*\|^2 + \mathrm{tr}(\mathrm{Cov}[\hat{\mu}_i]) = \frac{1}{n}\,\mathrm{tr}(\Sigma_i^*). \tag{2}$$

This result shows that the estimation error of the class prototype scales inversely with the number of samples n. This classical decomposition explains why the NCM classifier becomes unreliable in the few-shot regime: the variance of the image mean estimator becomes large, leading to unreliable prototype estimates and degraded classification performance. CLIP also has a text encoder which, when fed with a text prompt, produces text prototypes for zero-shot prediction. We now present a better estimator of the image population mean that exploits these text prototypes.

3.2 Mixed Image and Text Prototype Estimator

Let μ_t^* be the text prototype for class c, obtained deterministically by feeding a fixed prompt template to text encoder, so that $\mathbb{E}[\mu_t^*] = \mu_t^*$ and $\mathrm{Cov}[\mu_t^*] = 0$. We define the *mixed prototype estimator* as:

$$\hat{\mu}_{\text{mix}} = \lambda\hat{\mu}_i + (1-\lambda)\mu_t^*, \tag{3}$$

where $\lambda \in [0, 1]$ is the mixing coefficient controlling the contribution of the image prototype. We propose to use this mixed prototype for each class during classification, since it provides a better estimate of the population class mean than the standard image prototype. Classification is then performed using NCM with the mixed prototype $\hat{\mu}_{\text{mix}}$.

The MSE of the mixed prototype estimator with respect to the true population image mean can be decomposed into the sum of the squared bias and variance (see Appendix A for the full derivation):

$$\mathrm{MSE}(\hat{\mu}_{\text{mix}}, \mu_i^*) = (1-\lambda)^2||\mu_t^* - \mu_i^*||^2 + \lambda^2\frac{1}{n}\mathrm{tr}(\Sigma_i^*). \tag{4}$$

Mixing with text prototype acts as a variance shrinkage operator. Eq. 4 reveals that prototype mixing behaves as a shrinkage estimator. While mixing introduces a bias term determined by the modality gap $||\mu_t^* - \mu_i^*||^2$ between population image mean and text, at the same time it reduces the variance of the estimator through the λ^2 factor scaling the image covariance trace. When the modality gap is moderate compared to the variance term, mixing can reduce overall MSE by trading a small increase in bias for a larger reduction of variance.

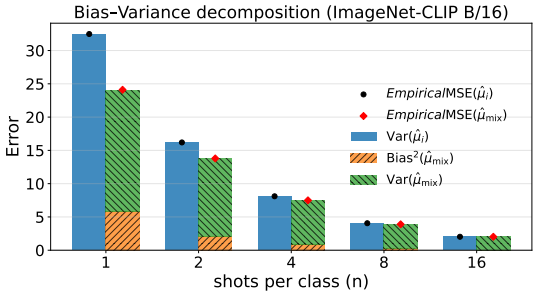

Figure 2: MSE of image and mixed prototypes as a function of the number of shots using CLIP-B/16.

In Figure 2, we analyze this decomposition on ImageNet. We approximate the population statistics μ_i^* and Σ_i^* using all training samples per class. For the text prototype μ_t^*, we use the standard CLIP prompt template "a photo of a {class name}". We then simulate the few-shot setting by repeatedly sampling n examples per class with replacement for $n \in \{1, 2, 4, 8, 16\}$. For each shot number n, we compute the empirical MSE of the image prototype $\hat{\mu}_i(n)$ and the mixed prototype $\hat{\mu}_{\text{mix}}(n)$ with respect to the estimated population mean μ_i^*. The results are averaged across multiple Monte Carlo trials and classes. The empirical MSE values are shown as dots in Figure 2, while the bars illustrate the bias and variance components from the theoretical decomposition. The mixing parameter λ is selected from a fixed grid for each shot level to minimize the empirical MSE. In the lower shots, image prototypes deviate significantly from the population means, resulting in higher MSE. Using mixed prototypes (green+orange) reduces the MSE, providing more accurate estimates of the population mean and a better bias–variance trade-off than the image prototype estimator.

Limitations of naive prototype mixing. While mixed prototypes reduce the estimation error of the class mean, there is a limitation: mixing is performed in the full embedding space, implicitly combining components of the image representation that may not be semantically related to the text prototype. In CLIP, image and text representations are only partially aligned (Liang et al., 2022). As a result, certain directions in the image embedding space correspond to semantic information captured by the text encoder, while others encode image-specific variations (Schrodi et al., 2025). Since text prototypes provide no information about image-specific directions, we propose mixing them only with image components in the text subspace to improve classification.

4 Task-Semantic Projection

CLIP's contrastive training encourages image and text representations to align. Consequently, image class means may be approximately aligned with their corresponding text prototypes, although they can also contain information absent in the text descriptions. At inference, the text prototypes of the C candidate classes define the classification task. We use them to identify the *task-semantic subspace*, allowing to isolate the image mean components along text-defined directions.

Let $T = [\mu_{t,1}^*, \ldots, \mu_{t,C}^*] \in \mathbb{R}^{d \times C}$ be the matrix of all the text prototypes. We compute a truncated Singular Value Decomposition (SVD) of T at rank k, $T_k = U_k \Sigma_k V_k^\top \in \mathbb{R}^{d \times C}$, where $U_k \in \mathbb{R}^{d \times k}$, $\Sigma_k \in \mathbb{R}^{k \times k}$, $V_k \in \mathbb{R}^{C \times k}$ contain the top-k left singular vectors, singular values, and right singular vectors, respectively. We define the matrix

$$P = U_k U_k^\top \in \mathbb{R}^{d \times d} \tag{5}$$

as the *task-semantic projector* onto the k-dimensional principal subspace spanned by left singular vectors of the text prototype matrix. The range U_k forms the *task-semantic subspace*.

For any image prototype $\mu_i \in \mathbb{R}^d$ (for brevity, we write $\mu_i = \mu_{i,c}$), we can decompose it as:

$$\mu_i = P\mu_i + (I_d - P)\mu_i = \mu_i^{\parallel} + \mu_i^{\perp}, \tag{6}$$

where $\mu_i^{\parallel} = P\mu_i$ is the *task-semantic component* of the image prototype, which is its projection onto the subspace defined by the class text prototypes. The residual $\mu_i^{\perp} = (I_d - P)\mu_i$ is the task-orthogonal complement, where $I_d \in \mathbb{R}^{d \times d}$ is the identity matrix. Intuitively, $\mu_i^{\parallel}$ retains image information along the dominant text directions for the given classification task, while $\mu_i^{\perp}$ captures additional information outside those directions that may still aid classification.

To visualize the image regions associated with the task-semantic subspace and its orthogonal complement, we use GradCAM (Selvaraju et al., 2020). Using the text prompts for all classes in a dataset, we compute the projector P. For an image feature f_i, we evaluate the energy $f_i P f_i^\top$ and

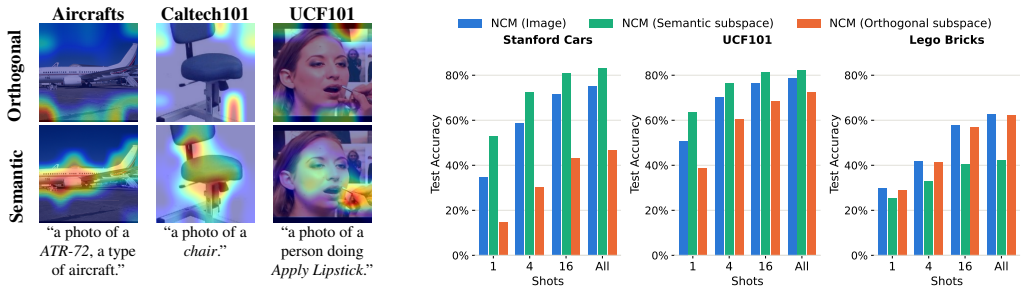

Figure 3: Impact of the *task-semantic* projection: (left) Attention maps reveal that visual features projected on the task-semantic subspace focus on the text information; (right) Using NCM classifier for standard datasets (Stanford Cars (Krause et al., 2013) and UCF101 (Soomro et al., 2012)) on CLIP-ViT-B/16, we show that task-semantic subspace is more discriminative than the full image space while the task-orthogonal subspace performs poorly. For OOD datasets like Lego Bricks (Garcia, 2019), the full image space and the orthogonal subspace is more discriminative than the semantic subspace due to poor cross-modal alignment in CLIP.

backpropagate this score to the last visual layer, highlighting regions that contribute to the task-semantic subspace. In contrast, using the orthogonal projection $(I - P)$ reveals complementary information. As shown in Figure 3 (left), the semantic component focuses on regions identified by the text space, while the orthogonal component captures instance-specific details. We demonstrate in Figure 3 (right) that the accuracy of an NCM classifier after projecting image prototypes onto the task-semantic subspace improves significantly over naive NCM on Stanford Cars and UCF101. This observation holds irrespective of the number of shots, suggesting a better class separation than the original image prototypes.

Although the proposed semantic projection works well for standard in-distribution datasets, we show in Figure 3 (right) that the task-semantic subspace performs poorly for OOD datasets like Lego Bricks where the original image space is more discriminative. We also observe that the task-orthogonal subspace is much less discriminative for Stanford Cars and UCF, but is more discriminative for Lego Bricks. These results suggest that the quality of the semantic subspace varies across tasks or datasets and depends on the alignment between the image and the text embedding space. We now discuss how to measure the existing cross-modal alignment in CLIP.

Analysis of Cross-Modal Alignment. The principal angles between the image and text prototype spaces measure cross-modal alignment, indicating how well the semantic structure in text matches the geometry of the image space. Computing SVD of text and image prototypes T and I, we obtain the cross-modal matrix M as:

$$T = U_t \Sigma_t V_t^\top \in \mathbb{R}^{d \times C}, \quad I = U_i \Sigma_i V_i^\top \in \mathbb{R}^{d \times C}, \quad M = U_t^\top U_i = U_M S_M V_M^\top \in \mathbb{R}^{C \times C} \tag{7}$$

where the singular values in S_M correspond to the *cosines of the principal angles*.

The mean of the principal angle cosines $\mu(S_M)$, shown in Figure 4 for 16-shots, indicates that CLIP-ViT-B/16 exhibits weak alignment for some datasets like EuroSAT, Lego Bricks and stronger alignment for others like Stanford Cars, UCF101. Refer to Appendix C for analysis on more datasets. The poor alignment could be due to poor semantic quality of the class labels (e.g. `Plate 1×2`, `Brick 1×3` in Lego Bricks) or could be due to a strong domain shift from training data distribution (e.g. satellite images in EuroSAT (Helber et al., 2019)). Due to poor alignment in these datasets, the proposed semantic projector does not manage to extract relevant discriminative features from the image space resulting in a weaker task-semantic subspace, and hence poor NCM performance. In such cases, the full image space and the orthogonal component have more discriminative information.

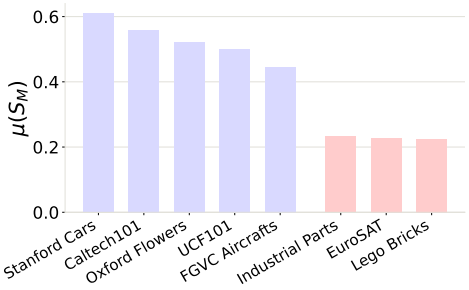

Figure 4: Analysis of $\mu(S_M)$.

Furthermore, even with good cross-modal alignment, the task-orthogonal component may contain discriminative information absent from text (Schrodi et al., 2025). We therefore separate these two sources of information: mix text and semantically projected image prototypes while handling image-specific additional information separately through a classifier in the full image space.

5 Proposed Training-free Classifier for Few-Shot Classification

In this section, we propose a training-free classifier for few-shot settings that combines a classifier in the task-semantic subspace with an image-based classifier.

Mixing Prototypes in the Task Semantic Subspace. The task-semantic projection improves classification by retaining more discriminative information, but the projected few-shot means $\hat{\mu}_i^{||} = P\hat{\mu}_i$ remains noisy. As discussed in Section 3, mixing with a text prototype reduces image mean estimation variance. Unlike mixing in the full image space, the relevant estimation target is now the projected population mean $\mu_i^{||,*} = P\mu_i^*$. We therefore mix prototypes in this subspace and use the resulting **Proj+Mix** prototypes to construct an NCM classifier in the task-semantic subspace (TSS), with logits computed using cosine similarity. We compute the logits s_{TSS} as follows:

$$\mu_{\mathrm{Proj+Mix},c} = \lambda\mu_{i,c}^{\|} + (1-\lambda)\mu_{t,c}^*, \qquad \mathrm{s}_{\mathrm{TSS},c} = \frac{Pf_{\text{test}}}{\|Pf_{\text{test}}\|} \cdot \frac{\mu_{\mathrm{Proj+Mix},c}}{\|\mu_{\mathrm{Proj+Mix},c}\|}, \;\; c = 1, \ldots, C, \tag{8}$$

where $\mu_{i,c}^{\|} = P\mu_{i,c}$ is the projected image prototype of class c, Pf_{test} is the projected test image feature and λ controls the mixing weight. Note that we apply the task-semantic projection to the test image features to classify them using the NCM classifier with Proj+Mix estimated prototypes. In Appendix B, we analyze why mixing prototypes in the full image space is suboptimal.

To incorporate additional information beyond the text-based task semantics, we propose to use a Mahalanobis image-based classifier.

Image-based Classifier. We compute the global covariance Σ using the features centered on the class means. We use a shrinkage regularized form of the covariance as $A = (N-1)\Sigma + \mathrm{tr}(\Sigma)I$ where N is the total number of samples. Now, we compute $A_M = A^{-1/2}$ which will be used to compute the Mahalanobis cosine similarity between the test features f_{test} and the prototypes $\mu_{i,c}$. We compute the logits s_{IM} as follows:

$$\mathrm{s}_{IM,c} = \frac{A_M f_{\text{test}}}{\|A_M f_{\text{test}}\|} \cdot \frac{A_M \mu_{i,c}}{\|A_M \mu_{i,c}\|}, \qquad c = 1, \ldots, C, \tag{9}$$

where A_M is the Mahalanobis projection in the image space. By exploiting the global image covariance, this classifier captures the anisotropy of the image space.

Cross-Modal Prototype (CMP) Classifier. The final classifier combines the logits from task-semantic subspace classifier s_{TSS} and the image-based classifier s_{IM}. The output logits s_{CMP} are calculated as:

$$\mathrm{s}_{CMP} = \mathrm{s}_{TSS} + \gamma \;\; \mathrm{s}_{IM} \tag{10}$$

where γ is a hyperparameter combining the two classifiers. This final classifier combines the strengths of the Proj+Mix estimator, exploiting the text-based semantics and reducing the variance in the lower shots with an anisotropic image classifier that can exploit reliable covariance estimates in the higher-shots.

While λ and γ could be tuned on a validation set following Zhang et al. (2022); Zhu et al. (2023); Wang et al. (2024), we also propose a *validation-free* approach to select λ and γ. Intuitively, the mixing coefficient λ used in Eq. (8) for the task-semantic subspace classifier could be selected based on shots. Lower shots imply lower weight λ for image prototypes due to higher variance (Eq. (2)). We propose using λ: {1: 0.1, 2: 0.2, 4: 0.3, 8: 0.4, 16: 0.5} for respective shots.

To select γ in Eq. (10), we propose to exploit the existing cross-modal alignment as follows:

$$\gamma = (1 - \mu(S_M))^2 * n/(n+1) \tag{11}$$

where $\mu(S_M)$ refer to the mean of principal angle cosines S_M (from Eq. (7)) and n refers to the number of shots. Here, γ gives us a squared cross-modal misalignment score scaled by a few-shot reliability term $n/(n+1)$ since principal angle cosines are computed from few-shot training samples.

Table 1: Few-shot classification performance (mean over 3 seeds) for training-free methods on 15 datasets using CLIP-ViT-B/16. Best results in each group are in **bold** and second best underlined.

		Standard Datasets										OOD Datasets					
Shot	Method	SUN	CARS	FOOD	CAL	FLWR	ImgNet	UCF	INSCT	AIR	PETS	YOGA	DTD	LEGO	ESAT	INDS	Avg
	Zero-Shot	62.6	65.3	86.1	92.9	71.4	68.8	66.8	35.4	24.8	89.1	6.6	44.4	8.9	47.7	27.2	53.2
1	TIP-Adp.	65.1	67.3	85.9	93.6	82.8	69.2	71.3	41.8	28.4	89.4	7.8	51.0	30.7	63.9	51.4	60.0
	APE	**66.7**	68.6	86.1	**94.2**	**87.5**	**69.6**	72.4	45.2	29.0	90.9	9.3	**53.9**	33.7	66.8	57.2	62.1
	GDA	65.6	65.4	**86.2**	93.0	80.9	69.1	72.3	47.0	26.7	89.3	7.4	48.1	31.3	68.9	59.7	60.7
	ProKeR	63.8	66.4	86.0	93.6	84.3	69.3	70.7	47.6	27.6	89.6	9.3	49.7	33.2	61.8	58.7	60.8
	CMP	66.6	**69.9**	85.9	93.7	87.4	68.9	**73.1**	**57.3**	**30.9**	**91.0**	**11.2**	53.7	**34.3**	**72.4**	**65.3**	**64.1**
2	TIP-Adp.	67.3	68.3	86.1	94.0	87.0	69.4	73.4	46.9	29.1	90.1	8.9	54.2	32.5	68.8	59.1	62.3
	APE	69.1	69.8	86.1	94.5	90.8	**70.0**	**77.1**	49.9	30.1	91.6	11.1	58.2	38.1	72.5	66.4	65.0
	GDA	68.6	70.6	**86.3**	93.8	89.2	69.5	75.8	63.5	31.0	90.5	12.6	54.5	49.1	75.0	74.8	67.0
	ProKeR	67.4	67.8	86.2	93.8	91.1	69.9	74.4	60.3	29.8	89.8	12.6	55.4	44.3	72.8	72.5	65.9
	CMP	**70.0**	**73.5**	86.2	**94.6**	**92.7**	69.7	76.5	**69.0**	**34.0**	**92.4**	**16.3**	**59.8**	**51.3**	**77.3**	**76.6**	**69.3**
4	TIP-Adp.	68.8	70.2	86.1	94.2	89.2	69.6	73.1	50.5	31.4	90.0	9.3	56.6	38.5	73.4	69.5	64.7
	APE	70.8	72.7	86.4	94.8	92.5	70.4	76.5	56.0	33.5	91.6	14.1	61.4	45.9	76.6	76.5	68.0
	GDA	72.1	74.6	**86.6**	94.6	94.3	70.6	76.9	74.6	36.9	91.3	20.0	61.4	64.5	80.3	86.0	72.3
	ProKeR	70.1	72.0	86.3	94.3	94.4	**70.9**	76.4	70.0	35.2	90.1	19.3	61.2	54.9	79.4	84.0	70.6
	CMP	**72.8**	**77.2**	86.5	**95.2**	**95.3**	70.5	**78.6**	**76.8**	**39.0**	**92.2**	**25.2**	**64.3**	**65.3**	**82.1**	**87.1**	**73.9**
16	TIP-Adp.	71.5	74.0	86.4	94.5	92.9	70.7	76.0	59.5	36.7	90.0	12.7	62.2	51.8	77.3	72.5	68.6
	APE	73.4	77.5	86.6	95.0	94.9	71.6	78.6	65.3	40.6	92.1	20.8	68.4	61.2	80.9	80.9	72.5
	GDA	75.9	83.1	87.3	95.7	97.6	72.2	83.4	85.6	49.8	93.6	39.7	70.9	85.3	89.2	93.0	80.2
	ProKeR	76.1	82.9	86.5	96.0	**98.4**	**73.3**	**84.2**	85.5	49.3	90.5	36.0	71.7	77.4	88.5	92.1	79.2
	CMP	**76.7**	**83.6**	**87.4**	**96.2**	98.1	72.5	83.8	**87.7**	**50.6**	**94.0**	**41.8**	**72.7**	**85.4**	**89.8**	**93.5**	**80.9**

6 Experimental Results

Datasets. We perform evaluation on 15 datasets comprising of standard in-distribution and out-of-distribution (OOD) datasets. Standard datasets include ImageNet (ImgNet), Caltech101 (CAL), SUN397 (SUN), UCF101 (UCF), FGVC Aircrafts (AIR), Stanford Cars (CARS), Oxford Pets (PETS), Flowers102 (FLWR), Food101 (FOOD) and Flying Insects (INSCT). OOD datasets include a satellite image dataset - EuroSAT (ESAT), a texture dataset - Describable Textures (DTD), a yoga pose dataset (YOGA), a lego bricks dataset (LEGO) and an industrial parts dataset (INDS). We provide more details in Appendix C.

Implementation Details. We select 1, 2, 4, 8, or 16 random samples per class to evaluate few-shot settings. To ensure fair comparison with existing training-free methods, we reproduce the baselines in the same evaluation settings and use three variants of CLIP - ViT-B/16, Siglipv2-B/16 and the recently proposed PE-Core-B/16 (Bolya et al., 2026). For all experiments in Table 1, the hyperparameters of all methods are selected using the validation set following standard practice (Zhang et al., 2022; Zhu et al., 2023; Wang et al., 2024). We select λ and γ for our method (CMP) via grid search over 10 values in the range [0, 1]. We also present the *validation-free* setting performance in Table 2. For the decomposition of image prototypes, we select the top-k components based on the explained variance ratio which covers 99% of the variance. Following existing methods, we apply 10 augmentations per image. All experiments report average accuracy over 3 random seeds.

Baselines. We compare our method with several training-free and training-based approaches. Among training-free methods, we use Zero-Shot CLIP (Radford et al., 2021), Tip-Adapter (Zhang et al., 2022), APE (Zhu et al., 2023), GDA (Wang et al., 2024) and ProKeR (Bendou et al., 2025). We also evaluate our approach on top of the adapted models from MaPLe (Khattak et al., 2023) and CLIP-LoRA (Zanella & Ben Ayed, 2024).

6.1 Comparison with Training-free Methods

Table 1 reports the comparison with training-free approaches across 15 datasets. The results show that the proposed method CMP outperforms existing approaches across all settings with more significant improvements on average in 1-Shot (+2% over APE), 2-shot (+2.3% over GDA) and 4-Shot (+1.6% over GDA) settings. We also improve over GDA in 16-Shot settings by 0.5% on average. Existing methods fare well either on lower- or higher-Shot settings (APE does better than GDA in 1-Shot while GDA is much better than APE in 2-, 4- and 16-Shot) while CMP outperforms them in most cases on all shots across standard and OOD datasets.

Table 2: Few-shot classification accuracy (mean over 3 seeds) on 14 datasets using different models in the *validation-free setting* (Validation on ImageNet for existing methods while ours do not require any validation set). Best results in each group are in **bold** and second best underlined.

	CLIP-ViT-B/16						SigLIPv2-B/16						PE-Core-B/16					
Method	1	2	4	8	16	Avg	1	2	4	8	16	Avg	1	2	4	8	16	Avg
TIP-Adp.	53.9	55.7	57.1	58.0	66.9	58.3	70.1	71.2	72.6	75.5	79.3	73.7	72.2	70.8	76.1	72.8	78.4	74.1
APE	57.3	57.5	61.3	64.1	69.8	62.0	70.6	72.7	74.2	77.4	81.9	75.3	70.0	73.3	74.9	75.5	80.3	74.8
GDA	58.3	65.6	71.3	76.5	80.3	70.4	62.0	75.0	81.9	**85.6**	86.9	78.3	72.8	76.2	80.6	84.4	86.3	80.1
ProKeR	59.7	63.2	66.9	73.4	79.1	68.5	74.6	77.1	80.5	83.3	**87.1**	80.5	74.0	76.1	78.5	82.7	86.0	79.4
CMP	**61.7**	**68.1**	**73.2**	**77.3**	**80.4**	**72.2**	**74.7**	**79.0**	**82.9**	85.5	86.8	**81.8**	**74.3**	**78.3**	**82.6**	**85.8**	**87.7**	**81.7**

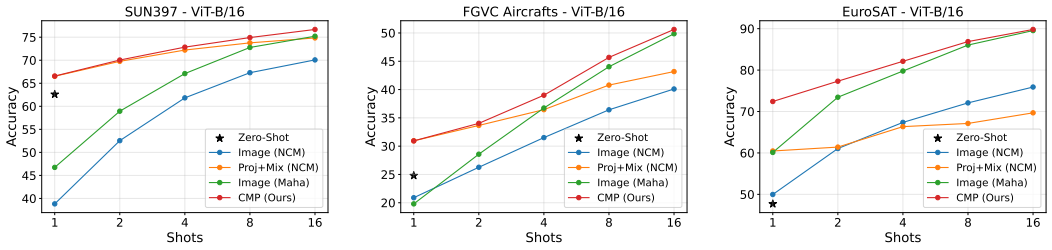

Figure 5: Impact of different classifiers across 3 datasets and varying shots. The proposed method CMP combines Proj+Mix (NCM) and Image (Maha).

Table 2 presents the performance comparison across different models and shots for 14 datasets in the *validation-free* setting where we consider no access to dataset-specific validation sets for our approach. While all existing methods need to tune the hyperparameters on a reference dataset like ImageNet as shown in Bendou et al. (2025), our method CMP works without any reference validation set. As proposed in Section 5, we select λ based on shots and γ using the cross-modal misalignment score (Eq. (11)). CMP outperforms all methods consistently using 3 VLMs on average over all shots (+1.8% over GDA on CLIP-ViT-B/16, +1.3% over ProKeR on SigLIPv2-B/16 and +1.6% over GDA on PE-Core-B/16). This further demonstrates the robustness of our method with reduced reliance on availability of validation sets. We provide more detailed results across all datasets on CLIP-ViT-B/16 in Appendix Section D.4.

6.2 Analysis and Ablation of Classifiers

We analyze the impact of different classifiers for 3 datasets using CLIP-ViT-B/16 in Figure 5. On the standard datasets, we observe a higher alignment (mean of principal angle cosines $\mu(S_M) > 0.4$) while on the OOD datasets, $\mu(S_M) < 0.4$. This alignment affects the performance of the Proj+Mix classifier. We present $\mu(S_M)$ scores for all datasets in Appendix C. On SUN397 having $\mu(S_M) \in [0.84, 0.85]$ across shots and FGVC Aircrafts $\big(\mu(S_M) \in [0.43, 0.45]\big)$, Proj+Mix significantly improves accuracy over Zero-Shot CLIP and NCM on image prototypes, thus validating the impact of using only the task-semantic subspace. In EuroSAT with poor alignment $\big(\mu(S_M) \in [0.2, 0.23]\big)$, Proj+Mix performs rather poorly and thus depends more on the Mahalanobis image classifier. Across all datasets, the anisotropic Mahalanobis classifier (Maha) is better than the isotropic NCM classifier. As the number of shots increases to 16, we see an improvement in the Image (Maha) classifier performance due to better covariance estimates leading to overall better performance of CMP which combines Proj+Mix (NCM) and Image (Maha).

In Figure 6, we present the ablations of using NCM classifiers with Image, Proj, Mix and Proj+Mix variants of prototypes. For datasets with good alignment like Stanford Cars $\big(\mu(S_M) \in [0.59, 0.61]\big)$ and Oxford Flowers $\big(\mu(S_M) \in [0.5, 0.52]\big)$, we observe similar trends - Proj consistently improves over naive Image classifiers, Mix improves over Proj and Image with decreasing impact as the number of shots increases, and finally Proj+Mix improves Mix, Proj and Image classifiers. For OOD datasets like Industrial Parts $\big(\mu(S_M) \in [0.19, 0.23]\big)$, we observe that Proj+Mix improves over Proj while Image and Mix performs better than Proj+Mix. This indicates that the full image space is more useful here compared to the task-semantic subspace due to the poor alignment.

In Appendix D we provide more analysis of classifiers across all datasets, the impact of the number of classes and analysis of hyperparameter selection.

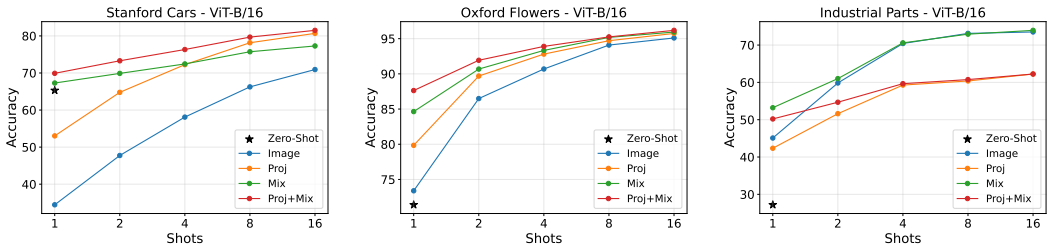

Figure 6: Ablations of using NCM classifiers with image prototypes (Image), image prototypes in task-semantic subspace (Proj), mixed prototypes in the full image space (Mix) and mixed prototypes in the task-semantic subspace (Proj+Mix).

Table 3: Performance comparison of proposed training-free (TF) method CMP with training-based methods using CLIP-ViT-B/16 backbone. Best results denoted in **bold** and second best is underlined.

		Standard Datasets										OOD Datasets					
Shot	Method	SUN	CARS	FOOD	CAL	FLWR	ImgNet	UCF	INSCT	AIR	PETS	YOGA	DTD	LEGO	ESAT	INDS	Avg
	Zero-Shot	62.6	65.3	86.1	92.9	71.4	68.8	66.8	35.4	24.8	89.1	6.6	44.4	8.9	47.7	27.2	53.2
4	CMP (TF)	72.8	77.2	**86.5**	95.2	95.3	70.5	78.6	76.8	39.0	92.2	25.2	64.3	**65.3**	82.1	87.1	73.9
	MaPLe	71.6	69.8	**86.5**	94.4	84.1	70.5	76.4	41.5	29.6	**93.2**	7.4	58.4	21.6	65.3	57.6	61.9
	+ CMP	74.4	76.6	86.3	95.3	94.9	71.7	80.4	76.7	39.8	92.7	20.8	65.0	51.4	82.2	85.2	72.9
	CLIP-LoRA	73.4	76.7	83.5	95.2	93.1	71.6	81.9	71.9	37.3	91.9	23.4	63.6	53.6	84.9	87.8	72.7
	+ CMP	**75.9**	**80.5**	84.4	**95.9**	**96.3**	**72.7**	**84.0**	**83.4**	**42.4**	92.6	**31.1**	**69.4**	55.2	**88.1**	**88.4**	**76.0**
16	CMP (TF)	76.7	83.6	87.4	96.2	98.1	72.5	83.8	87.7	50.6	94.0	41.8	72.7	**85.4**	89.8	93.5	80.9
	MaPLe	74.4	74.7	87.5	95.3	93.7	71.9	80.9	67.1	36.0	93.5	15.2	67.6	42.4	85.1	86.6	71.5
	+ CMP	75.5	85.6	**87.7**	96.2	98.4	72.2	84.7	88.9	51.7	**94.1**	43.3	74.1	76.0	91.9	93.7	80.9
	CLIP-LoRA	76.9	85.1	85.3	96.5	97.9	73.5	86.6	90.9	54.9	93.0	54.6	72.5	80.4	92.3	**96.1**	82.4
	+ CMP	**78.4**	**87.7**	86.6	**96.8**	**98.6**	**74.7**	**88.2**	**92.0**	**59.9**	93.7	**59.3**	**75.5**	81.3	**93.1**	**96.1**	**84.1**

6.3 Comparison with Training-based Methods

We compare with MaPLe (Khattak et al., 2023), which learns multi-modal prompts, and CLIP-LoRA (Zanella & Ben Ayed, 2024), which adds LoRA modules for few-shot adaptation. We show in Table 3 that the proposed training-free method CMP outperforms training-based MaPLe by 12% on 4-shot and 9.4% on 16-shot settings. Training-free CMP also outperforms CLIP-LoRA on 4-shot setting while performing competitively on 16-shots. We also demonstrate that our method can be easily added as a plug-in on top of the embedding space learned using these methods. Using our method with MaPLe improves performance by 11% on 4-Shot and 9.4% in 16-Shot settings on average. Similarly, we see an improved performance of 3.3% on 4-Shot and 1.7% on 16-Shot on average using CMP on the adapted model from CLIP-LoRA.

7 Conclusions

In this work, we introduced a training-free approach to few-shot classification with vision-language models. Our bias–variance analysis shows that mixing image and text prototypes improves classification, but remains suboptimal because the two modalities are only partially aligned. To isolate task-specific semantics, our semantic subspace classifier projects image prototypes onto the principal directions of the text embedding space and mixes only their task-semantic components with text prototypes. Because cross-modal alignment varies across datasets and the image space contains additional discriminative variance, we complement this classifier with an image-specific classifier that models class covariances. Fusing the two classifiers leverages the complementary strengths of both modalities. Across 15 benchmarks, our approach consistently outperforms existing training-free baselines and integrates seamlessly with models updated through training-based methods.

Limitations and Future Work. Our approach relies on the cross-modal alignment of the pre-trained CLIP model. For out-of-distribution datasets with severe domain shift or inadequate text representations (e.g. EuroSAT, Lego Bricks), the task-semantic projection provides limited benefit. Additionally, our image-specific classifier is less effective in low-shot scenarios because robust covariance estimation is unstable. Moving forward, we plan to explore the integration of parameter-efficient fine-tuning to explicitly encourage and improve subspace alignment for challenging downstream tasks.

AI USE STATEMENT

We used generative AI tools solely for limited language editing, including stylistic refinement and grammar correction. We did not use these tools for literature retrieval or discovery, research ideation or execution, drafting sections of the paper, generating synthetic data, or deriving mathematical claims. All AI-assisted edits were reviewed by the authors, who take full responsibility for the paper's final content, claims, and artifacts.

REPRODUCIBILITY STATEMENT

We have taken steps to ensure the reproducibility of our work. The full source code and scripts to reproduce the results in the paper will be published after the review phase is completed. All experiments were performed on publicly available datasets, and details of model architectures, and main training hyperparameters are given in the main paper with additional details included in the Appendix.

REFERENCES

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21466–21474, 2022. 2

Yassir Bendou, Amine Ouasfi, Vincent Gripon, and Adnane Boukhayma. Proker: A kernel perspective on few-shot adaptation of large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25092–25102. IEEE, 2025. 1, 2, 7, 8, 15, 17

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. 7

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 15

Lorenzo Caselli, Marco Mistretta, Simone Magistri, and Andrew D. Bagdanov. SpectralGCD: Spectral concept selection and cross-modal representation learning for generalized category discovery. In *The Fourteenth International Conference on Learning Representations*, 2026. 2

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 15

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in clip. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006. 15

Hengyi Feng, Zeang Sheng, Meiyi Qiang, Yang Li, and Wentao Zhang. Generative giants, retrieval weaklings: Why do multimodal large language models fail at multimodal retrieval? In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15917–15933, 2026. 1

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. *IJCV*, 2024. 3

Francisco Garcia. Lego brick sorting - image recognition. `https://www.kaggle.com/datasets/pacogarciam3/lego-brick-sorting-image-recognition`, 2019. Kaggle. 5, 15

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*, 2019. 5, 15

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal Prompt Learning. In *CVPR*, 2023. 1, 3, 7, 9

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 5, 15

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of clip. In *International Conference on Machine Learning*, pp. 33999–34019. PMLR, 2025. 1

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. 1, 2, 4

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. arXiv:1306.5151. 15

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. 2

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *IEEE Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 15

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 15

Gerry Piosenka. Butterfly & moths image classification 100 species. `https://www.kaggle.com/datasets/gpiosenka/butterfly-images40-species`, 2023. Kaggle. 15

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 *ICML*, 2021. 1, 2, 7

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In *IJCV*, 2015. 15

Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2765–2775, 2023. 2

Shruti Saxena. Yoga pose image classification dataset. `https://www.kaggle.com/datasets/shrutisaxena/yoga-pose-image-classification-dataset`, 2021. Kaggle. 15

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 4, 6

Berit Schuerrle and Venkatesh Sankarappan. Industrial dataset. `https://www.kaggle.com/datasets/beschue/industrial-classification-data-set`, 2023. Kaggle. 15

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Feb 2020. ISSN 1573-1405. 4

Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 23681–23690, 2024. 1

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. 5, 15

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2725–2736. IEEE, 2023. 2

Yunhan Wang, Eshika Khandelwal, Edson Araujo, Walid Bousselham, Nina Shvetsova, and Hilde Kuehne. Decompose, compare, and decide: Multimodal llms are implicit few-shot learners. In *European Conference on Computer Vision*, pp. 157–174. Springer, 2026. 15

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free clip-based adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 6, 7, 15

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 15

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. MMA: Multi-Modal Adapter for Vision-Language Models. In *CVPR*, 2024. 3

Lu Yu, Haiyang Zhang, and Changsheng Xu. Text-guided attention is all you need for zero-shot robustness in vision-language models. *Advances in Neural Information Processing Systems*, 37: 96424–96448, 2024. 2

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 2

Maxime Zanella and Ismail Ben Ayed. Low-Rank Few-Shot Adaptation of Vision-Language Models. In *CVPRW*, 2024. 1, 3, 7, 9

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification. In *ECCV*, 2022. 1, 2, 6, 7, 15

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024. 1

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022a. 1, 3

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022b. 3

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 2, 6, 7, 15

APPENDIX A MEAN SQUARED ERROR DERIVATIONS

We first review the Mean Squared Error (MSE) for a generic estimator, then extend the bias-variance decomposition of MSE to $\mathbb{R}^d$ space. Finally, we apply this analysis to the mixed prototype estimator and to the mixed prototype estimator confined to the semantic subspace.

A.1 MEAN SQUARE ERROR FOR A GENERIC ESTIMATOR

Recall that for a generic estimator $\hat{\theta} : \Omega \to \mathbb{R}$, the MSE with respect to the true value θ^* is defined as

$$\mathrm{MSE}(\hat{\theta}, \theta^*) = \mathbb{E}_{X \sim P_{\theta^*}}\left[(\hat{\theta}(X) - \theta^*)^2\right] = \mathrm{Bias}^2(\hat{\theta}, \theta^*) + \mathrm{Var}(\hat{\theta}), \tag{12}$$

where

$$\mathrm{Bias}(\hat{\theta}, \theta^*) = \mathbb{E}_{X \sim P_{\theta^*}}[\hat{\theta}(X)] - \theta^*, \tag{13}$$

and

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}_{X \sim P_{\theta^*}}\left[(\hat{\theta}(X) - \mathbb{E}_{X \sim P_{\theta^*}}[\hat{\theta}(X)])^2\right]. \tag{14}$$

Here $X \sim P_{\theta^*}$ denotes samples drawn from the data distribution under the true parameter θ^*. From now on, all expectations are taken with respect to $X \sim P_{\theta^*}$, and we omit this dependence for notational simplicity.

A.2 BIAS–VARIANCE DECOMPOSITION IN $\mathbb{R}^d$

We now extend the bias–variance decomposition to the case of vector-valued estimators $\hat{\theta} \in \mathbb{R}^d$. The bias is defined as

$$\mathrm{Bias}(\hat{\theta}, \theta^*) = \mathbb{E}[\hat{\theta}] - \theta^*, \tag{15}$$

which is now a vector in $\mathbb{R}^d$. The squared bias is :

$$\mathrm{Bias}^2(\hat{\theta}, \theta^*) = (\mathbb{E}[\hat{\theta}] - \theta^*)^\top (\mathbb{E}[\hat{\theta}] - \theta^*) = \|\mathbb{E}[\hat{\theta}] - \theta^*\|^2. \tag{16}$$

The covariance matrix of $\hat{\theta}$ is

$$\mathrm{Cov}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\hat{\theta} - \mathbb{E}[\hat{\theta}])^\top\right]. \tag{17}$$

The corresponding variance is defined as:

$$\mathrm{Var}(\hat{\theta}) = \mathrm{tr}(\mathrm{Cov}(\hat{\theta})). \tag{18}$$

Consequently, the Mean Squared Error with the squared Euclidean norm is defined as:

$$\mathrm{MSE}(\hat{\theta}, \theta^*) = \mathbb{E}\Big[\|\hat{\theta} - \theta^*\|^2\Big] = \mathrm{Bias}^2(\hat{\theta}, \theta^*) + \mathrm{Var}(\hat{\theta}) = \|\mathbb{E}[\hat{\theta}] - \theta^*\|^2 + \mathrm{tr}\Big(\mathrm{Cov}(\hat{\theta})\Big). \tag{19}$$

A.3 BIAS–VARIANCE DECOMPOSITION FOR MIXED PROTOTYPE ESTIMATOR

Defining μ_i^* as the true population image class mean for a generic class c (we omit the class index c to avoid cluttered notation) and $\hat{\mu}_i$ as the empirical sample mean computed from n samples, we get:

$$\mathbb{E}[\hat{\mu}_i] = \mu_i^*, \quad \mathrm{Cov}[\hat{\mu}_i] = \frac{\Sigma_i^*}{n} \tag{20}$$

where Σ_i^* is the population covariance for class c. We assume that the text prototype μ_t^* for class c is deterministic, since it comes from a fixed prompt template fed into the text encoder. Hence:

$$\mathbb{E}[\mu_t^*] = \mu_t^*, \quad \mathrm{Cov}[\mu_t^*] = 0. \tag{21}$$

The mixed prototype estimator is defined as:

$$\hat{\mu}_{\text{mix}} = \lambda \hat{\mu}_i + (1 - \lambda)\mu_t^*. \tag{22}$$

We now analyze the bias–variance decomposition of the mixed estimator.

Bias of the mixed estimator. By computing the expectation we have:

$$\mathbb{E}[\hat{\mu}_{\text{mix}}] = \mathbb{E}[\lambda\hat{\mu}_i + (1-\lambda)\mu_t^*] = \lambda\mathbb{E}[\hat{\mu}_i] + (1-\lambda)\mathbb{E}[\mu_t^*] = \lambda\mu_i^* + (1-\lambda)\mu_t^*, \tag{23}$$

where in the last step we have used that $\mathbb{E}[\hat{\mu}_i] = \mu_i^*$ and $\mathbb{E}[\mu_t^*] = \mu_t^*$. The squared bias is given by:

$$\text{Bias}^2(\hat{\mu}_{\text{mix}}, \mu_i^*) = \|\mathbb{E}[\hat{\mu}_{\text{mix}}] - \mu_i^*\|^2 = \|\lambda\mu_i^* + (1-\lambda)\mu_t^* - \mu_i^*\|^2 \tag{24}$$

Covariance of the mixed estimator. Assuming that the image and text prototypes are independent, $\text{Cov}[\hat{\mu}_i, \mu_t^*] = 0$ and since we assumed that the text prototype is deterministic ($\text{Cov}[\mu_t^*] = 0$), we obtain:

$$\text{Cov}[\hat{\mu}_{\text{mix}}] = \lambda^2\text{Cov}[\hat{\mu}_i] + (1-\lambda)^2\text{Cov}[\mu_t^*] = \lambda^2\text{Cov}[\hat{\mu}_i] = \lambda^2\frac{\Sigma_i^*}{n}, \tag{25}$$

where the last equality follows from Eq. (20).

MSE of the mixed estimator. The MSE, obtained by summing the squared bias and the variance, is given by:

$$\begin{aligned}\text{MSE}(\hat{\mu}_{\text{mix}}, \mu_i^*) &= \|\lambda\mu_i^* + (1-\lambda)\mu_t^* - \mu_i^*\|^2 + \lambda^2\frac{1}{n}\text{tr}(\Sigma_i^*) \\ &= \|(\lambda-1)\mu_i^* + (1-\lambda)\mu_t^*\|^2 + \lambda^2\frac{1}{n}\text{tr}(\Sigma_i^*) \\ &= (1-\lambda)^2\|\mu_t^* - \mu_i^*\|^2 + \lambda^2\frac{1}{n}\text{tr}(\Sigma_i^*).\end{aligned} \tag{26}$$

APPENDIX B ANALYSIS ON MIXING IN THE FULL IMAGE SPACE

Recall that in Section 4, we defined the *task-semantic projector* onto the k-dimensional principal subspace spanned by the left singular vectors of the text prototype matrix as:

$$P = U_k U_k^\top \in \mathbb{R}^{d\times d}. \tag{27}$$

This leads to the task-semantic subspace decomposition of the empirical image prototype $\hat{\mu}_i \in \mathbb{R}^d$:

$$\hat{\mu}_i = P\hat{\mu}_i + (I-P)\hat{\mu}_i = \mu_{\text{i}}^{\parallel} + \mu_{\text{i}}^{\perp}. \tag{28}$$

We decompose the mixed prototype estimator as:

$$\begin{aligned}\hat{\mu}_{\text{mix}} &= \lambda\hat{\mu}_i + (1-\lambda)\mu_t^* = \lambda(\mu_{\text{i}}^{\parallel} + \mu_{\text{i}}^{\perp}) + (1-\lambda)\mu_t^* \\ &= \mu_t^* + \lambda(\mu_{\text{i}}^{\parallel} - \mu_t^*) + \lambda\,\mu_{\text{i}}^{\perp}.\end{aligned} \tag{29}$$

This shows that naive mixing interpolates the task-semantic image component $\mu_{\text{i}}^{\parallel}$ with the text prototype, while simultaneously scaling the orthogonal image component $\mu_{\text{i}}^{\perp}$ by λ. Now we will compute the MSE for the mixed prototype estimator.

Now, by applying the linearity of the expectation and recalling that the text prototype is deterministic (Eq. (21)), we obtain the expectation as follows:

$$\mathbb{E}[\hat{\mu}_{\text{mix}}] = \mathbb{E}[\mu_t^* + \lambda(\mu_{\text{i}}^{\parallel} - \mu_t^*) + \lambda\,\mu_{\text{i}}^{\perp}] = \mu_t^* + \lambda(\mu_i^{\parallel,*} - \mu_t^*) + \lambda\mu_i^{\perp,*} \tag{30}$$

where $\mu_i^{\parallel,*} = \mathbb{E}[\mu_{\text{i}}^{\parallel}]$ and $\mu_i^{\perp,*} = \mathbb{E}[\mu_{\text{i}}^{\perp}]$ denote the task-semantic and task-orthogonal components of the population image class mean. The squared bias is given by:

$$\begin{aligned}\text{Bias}^2(\hat{\mu}_{\text{mix}}, \mu_i^*) &= ||\mathbb{E}[\hat{\mu}_{\text{mix}}] - \mu_i^*||^2 \\ &= \left\|\mu_t^* + \lambda(\mu_i^{\parallel,*} - \mu_t^*) + \lambda\mu_i^{\perp,*} - \mu_i^{\parallel,*} - \mu_i^{\perp,*}\right\|^2 \\ &= (1-\lambda)^2\left\|(\mu_t^* - \mu_i^{\parallel,*}) - \mu_i^{\perp,*}\right\|^2.\end{aligned} \tag{31}$$

Since $\mu_t^* - \mu_i^{\parallel,*}$ lies in the task-semantic subspace and $\mu_i^{\perp,*}$ lies in its orthogonal complement, the two vectors are orthogonal. Therefore, $\|a-b\|^2 = \|a\|^2 + \|b\|^2$ and the bias become:

$$\text{Bias}^2(\hat{\mu}_{\text{mix}}, \mu_i^*) = (1-\lambda)^2\left(\|\mu_t^* - \mu_i^{\parallel,*}\|^2 + \|\mu_i^{\perp,*}\|^2\right). \tag{32}$$

Similarly, we obtain the covariance as:

$$\mathrm{Cov}[\hat{\mu}_{\text{mix}}] = \lambda^2 \mathrm{Cov}[\mu_{\text{i}}^{\parallel}] + \lambda^2 \mathrm{Cov}[\mu_{\text{i}}^{\perp}] = \lambda^2 \frac{\Sigma_i^{\parallel,*}}{n} + \lambda^2 \frac{\Sigma_i^{\perp,*}}{n}. \tag{33}$$

Finally, the MSE decomposition is given by:

$$\mathrm{MSE}(\hat{\mu}_{\text{mix}}, \mu_i^*) = (1-\lambda)^2 \left(\|\mu_t^* - \mu_i^{\parallel,*}\|^2 + \|\mu_i^{\perp,*}\|^2 \right) + \lambda^2 \frac{1}{n} \left(\mathrm{tr}(\Sigma_i^{\parallel,*}) + \mathrm{tr}(\Sigma_i^{\perp,*}) \right), \tag{34}$$

where $\mu_i^{\parallel,*}$ and $\mu_i^{\perp,*}$ denote the task-semantic and task-orthogonal components of population image class mean, and $\Sigma_i^{\parallel,*}$, $\Sigma_i^{\perp,*}$ are the corresponding class covariances in the two subspaces. Notably, the factor $(1-\lambda)^2$ also scales the bias contribution $\|\mu_i^{\perp,*}\|^2$ from the orthogonal component, even though the text prototype contains no information about it.

Consequently, naive mixing pulls the text-based semantic component toward the text prototype while also shrinking the orthogonal component toward zero. In particular, reducing λ to control the high variance of few-shot image prototypes also suppresses image-specific information in the orthogonal subspace, although no text information is available to guide this shrinkage. Therefore, mixing with text prototypes in the full image space is sub-optimal and we propose to handle the image-specific information separately while mixing only the task-semantic component of image prototype with the text prototype.

APPENDIX C DATASETS

The datasets used in our experiments include generic object recognition datasets like ImageNet (ImgNet; Russakovsky et al., 2015) having 1000 classes and Caltech101 (CAL; Fei-Fei et al., 2006) having 100 classes, fine-grained image datasets like FGVC Aircrafts (AIR; Maji et al., 2013) having 100 classes, Stanford Cars (CARS; Krause et al., 2013) having 196 classes, Oxford Pets (PETS; Parkhi et al., 2012) having 37 classes, Flowers102 (FLWR; Nilsback & Zisserman, 2008) having 102 classes, Flying Insects (INSCT; Piosenka, 2023) having 100 classes and Food101 (FOOD; Bossard et al., 2014) having 101 classes, a scene recognition dataset - (SUN397; Xiao et al., 2010) having 397 classes, an action recognition dataset - (UCF101; Soomro et al., 2012) having 101 classes. We also include several out-of-distribution datasets like a satellite image dataset - EuroSAT (ESAT; Helber et al., 2019) having 10 classes, a texture dataset - Describable Textures (DTD; Cimpoi et al., 2014) having 47 classes, a lego bricks dataset (LEGO; Garcia, 2019) having 20 classes, a yoga pose dataset (YOGA; Saxena, 2021) having 107 classes and an industrial parts dataset (INDS; Schuerrle & Sankarappan, 2023) having 10 classes. While 11 of these datasets are commonly used in all CLIP-based few-shot classification benchmarks (Zhang et al., 2022; Zhu et al., 2023; Wang et al., 2024; Bendou et al., 2025), we consider 4 new datasets (INSCT, YOGA, LEGO and INDS) following a recent work (Wang et al., 2026) which uses these datasets to evaluate generalization performance of multimodal LLMs.

Cross-Modal Alignment across Datasets. We present the mean of the principal angle cosines $\mu(S_M)$ for all datasets in Figure 7 using CLIP-ViT-B/16 on the 4-shot setting. These values reflect the existing cross-modal alignment in CLIP for different datasets.

APPENDIX D ADDITIONAL ANALYSIS AND ABLATIONS

In this section, we provide more analysis and ablations to complement those in the main paper.

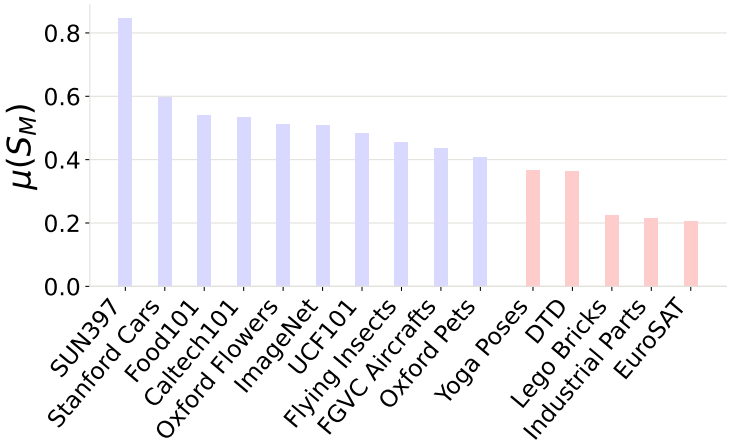

Figure 7: Analysis of the mean of principal angle cosines $\mu(S_M)$ indicating cross-modal alignment across all datasets in the 4-shot setting using CLIP-ViT-B/16.

D.1 Impact of the Number of Classes on the Semantic Projection

To study the impact of the number of classes on the proposed semantic subspace decomposition, we consider the SUN397 dataset which has 397 classes and good cross-modal alignment in CLIP. We selected random subsets of classes of varying number of classes in the 16-shot setting and compared the performance of NCM with image prototypes in the task-semantic subspace and in the task-orthogonal subspace as shown in Figure 8. We observe that despite having very few classes and thereby smaller task-semantic component compared to the orthogonal component, NCM in the proposed semantic subspace consistently performs better. This implies that the proposed task-semantic projector identifies the more discriminative subspace irrespective of the number of classes or the rank of the semantic projector.

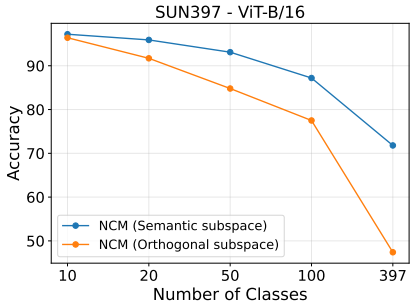

Figure 8: Analysis of the impact of varying subsets of classes on the semantic and orthogonal subspaces.

D.2 Impact of Different Classifiers

We demonstrate the impact of different classifiers in Figure 9 for 12 datasets using CLIP-ViT-B/16. We observe some consistent trends across all datasets. Image (Maha) improves significantly over Image (NCM) since it exploits the image covariances. The Proj+Mix (NCM) classifier performs very well in the lower shots for most datasets except the OOD datasets. The proposed method CMP combining Proj+Mix (NCM) and Image (Maha) always outperforms the other classifiers.

D.3 Impact of Mixing in the Semantic Subspace

We demonstrate the impact of mixing text and image prototypes in the task-semantic subspace on 12 datasets in Figure 10. Across all datasets, we see a significant impact of NCM with naively mixed prototypes (Mix) over image prototypes (Image). NCM using the task-semantic subspace of the image prototypes (Proj) improves over NCM with image prototypes (Image) for all datasets except EuroSAT and Lego Bricks. Finally, we show that mixing the task-semantic component of the image prototypes with the text prototypes (Proj+Mix) improves the accuracy by non-trivial margins for most datasets across all shots.

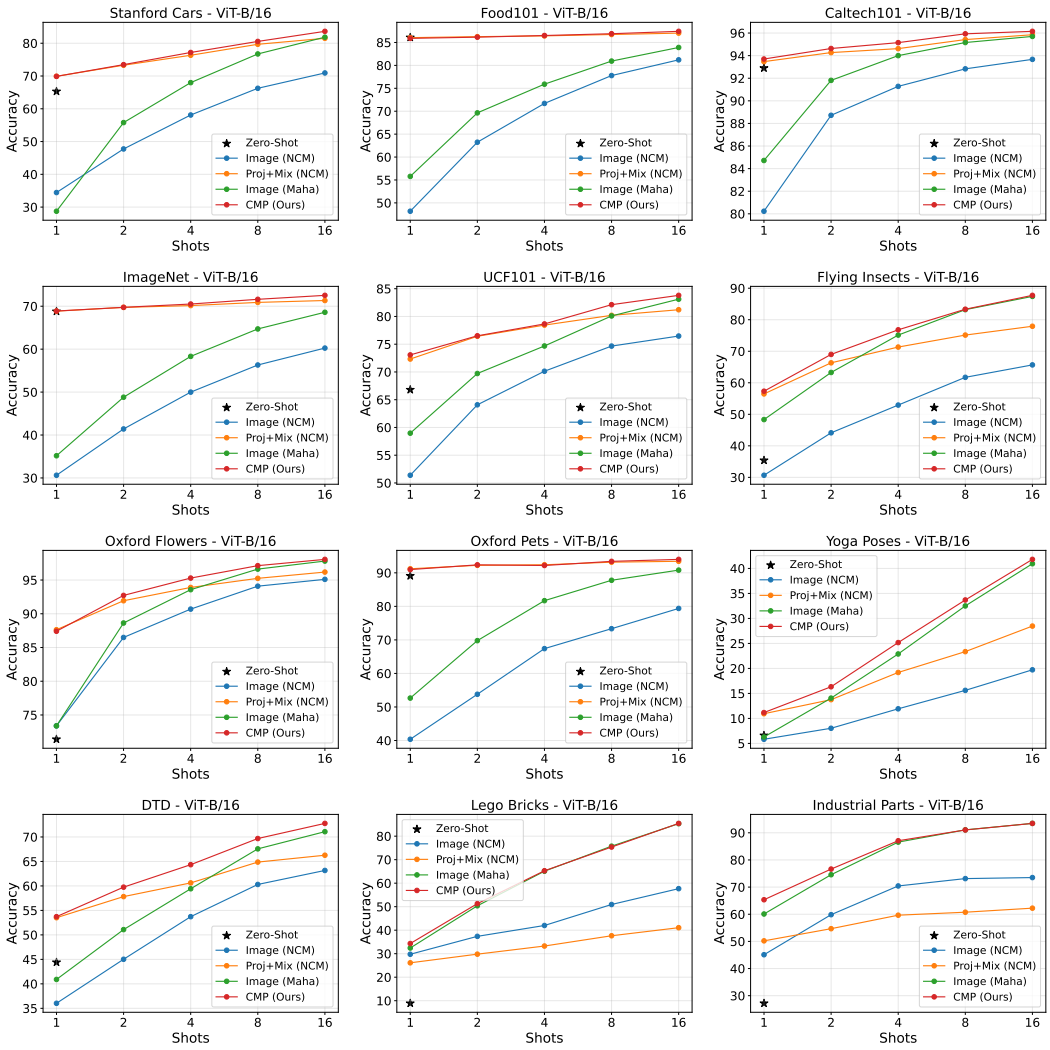

Figure 9: Impact of different classifiers across 3 datasets and varying shots on 12 datasets with CLIP-ViT-B/16.. The proposed method CMP combines Proj+Mix (NCM) and Image (Maha).

D.4 Performance Comparison on validation-free setting

We present the performance comparison with all training-free methods for 14 datasets using CLIP-ViT-B/16 in the *validation-free* setting in Figure 11. All existing methods are very sensitive to hyperparameters and needs to carefully finetune them on a validation set. Following the validation-free setting from (Bendou et al., 2025) where the hyperparameters are selected on the validation set of a reference dataset ImageNet and then the selected values are used for all downstream datasets. We followed this approach for all existing methods. For the proposed CMP classifier, we do not require any such hyperparameter tuning on a reference validation set. For our method CMP, we select the mixing coefficient λ based on shots and γ based on the cross-modal misalignment score from Eq. (10) to combine the Proj+Mix (NCM) and Image (Maha) classifiers. Figure 11 shows that the proposed CMP outperforms other methods in most cases across all datasets. The performance of existing methods vary across datasets suggesting the need to carefully finetune them on dataset-specific validation sets. In contrast, CMP is relatively stable across all datasets suggesting less reliance on dataset-specific hyperparameter selection.

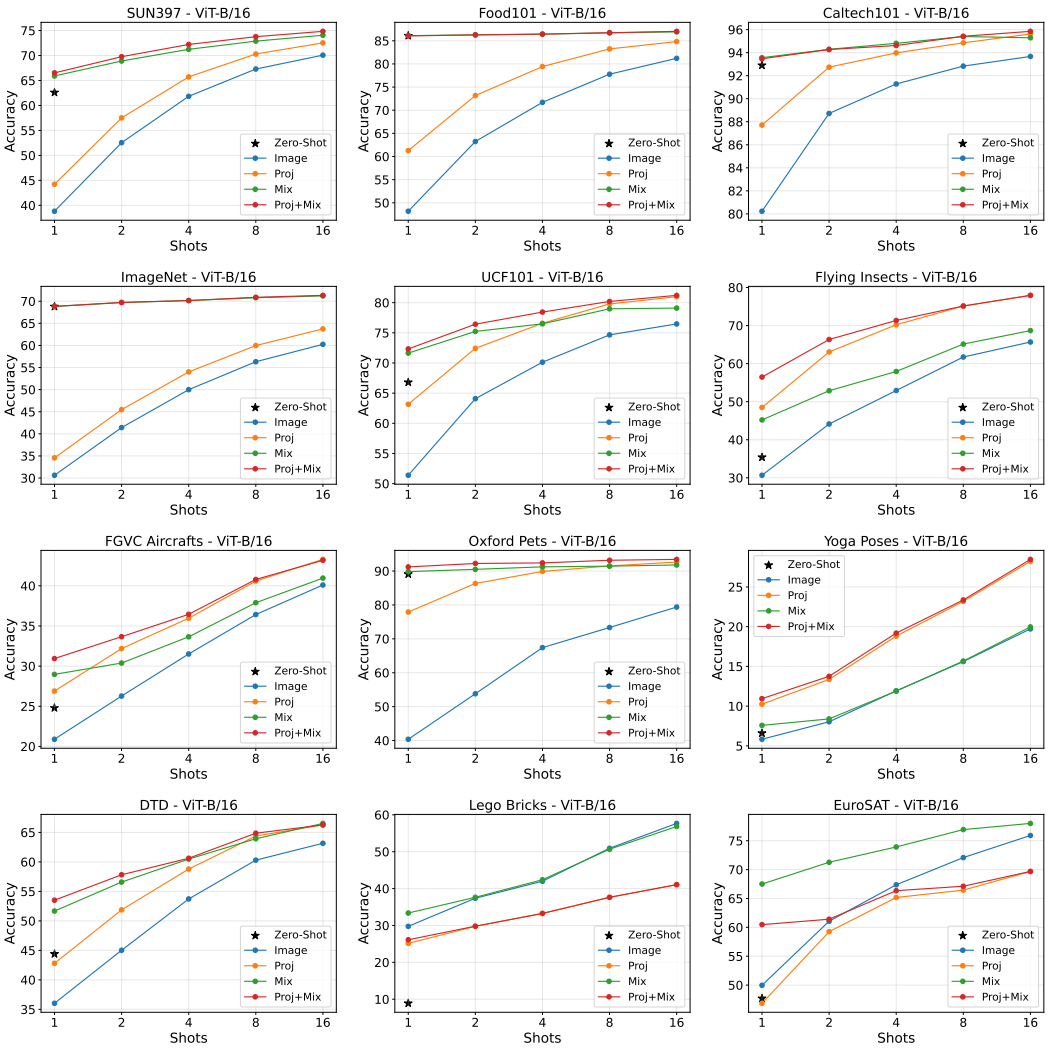

Figure 10: Ablations of using NCM classifiers with image prototypes (Image), image prototypes in task-semantic subspace (Proj), mixed prototypes in the full image space (Mix) and mixed prototypes in the task-semantic subspace (Proj+Mix) on 12 datasets with CLIP-ViT-B/16.

D.5 Analysis of Hyperparameters

We analyze the sensitivity of the proposed method to the hyper-parameters λ and γ in Figure 12 when they are tuned on validation sets of all datasets. $\lambda \in [0, 1]$ is the mixing coefficient of the Proj+Mix classifier. Higher lambda values assign higher weights to the image prototype in the task-semantic space and lower weights to the text prototype. $\gamma \in [0, 1]$ is the weight for the image-based classifier. For SUN397, Stanford Cars and Oxford Flowers, we see similar trends where the γ values corresponding to higher accuracies are lower for few-shots and increases for 16-shot. This suggests that the Mahalanobis image classifier is more significant in 16-shot setting with better covariance estimates. Similarly, for λ, we see that the optimal lambda value increases with more shots where the image prototype estimates are better. Note that we always see an optimal value of $\lambda > 0$ implying the significance of mixing in the semantic subspace. For EuroSAT and Lego Bricks, we see more dependence on the image classifier with higher optimal values of γ. Similar trends across most datasets suggest that the hyperparameters are less sensitive to tasks or datasets.

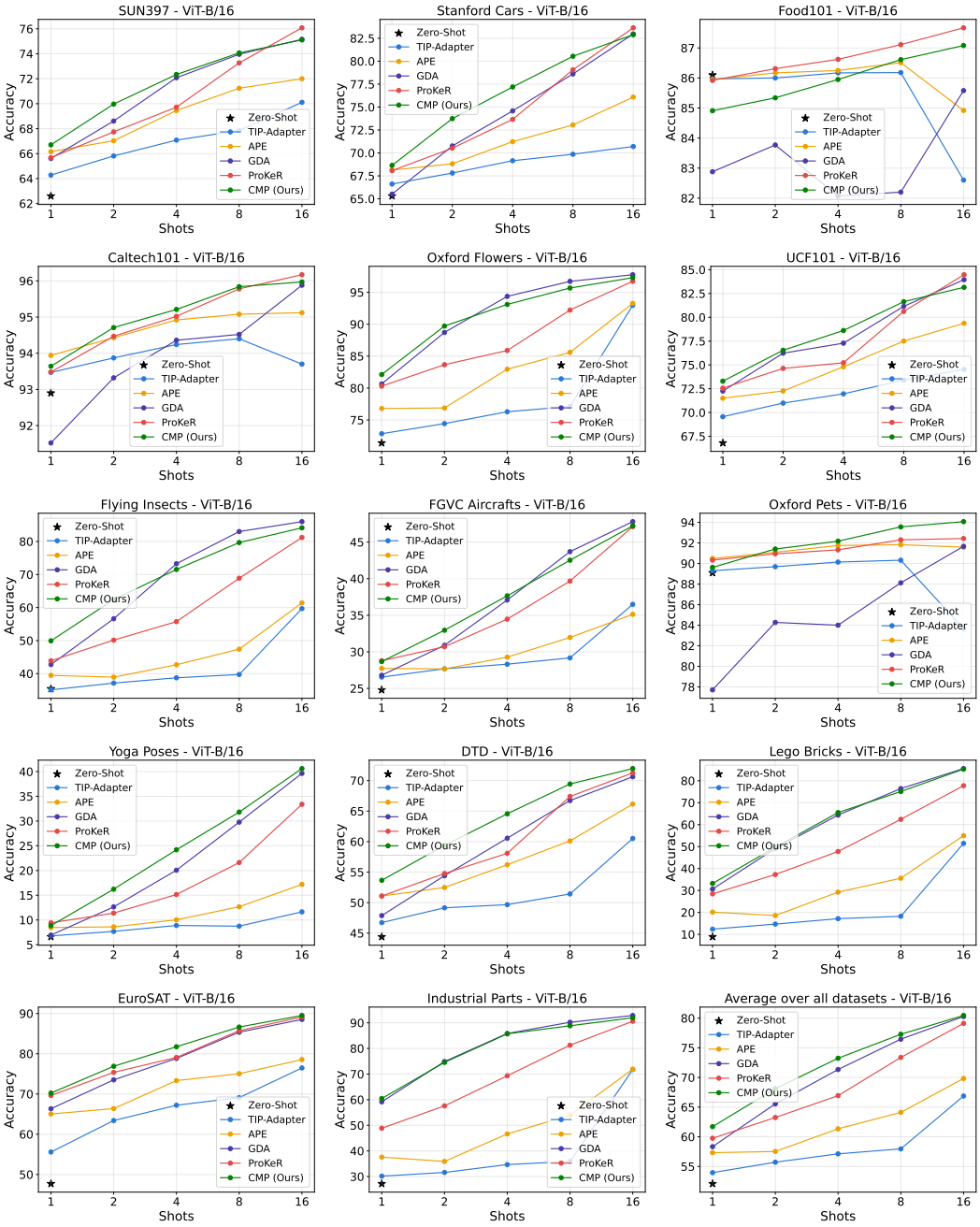

Figure 11: Performance comparison of training-free methods using CLIP-ViT-B/16 in the *validation-free* setting.

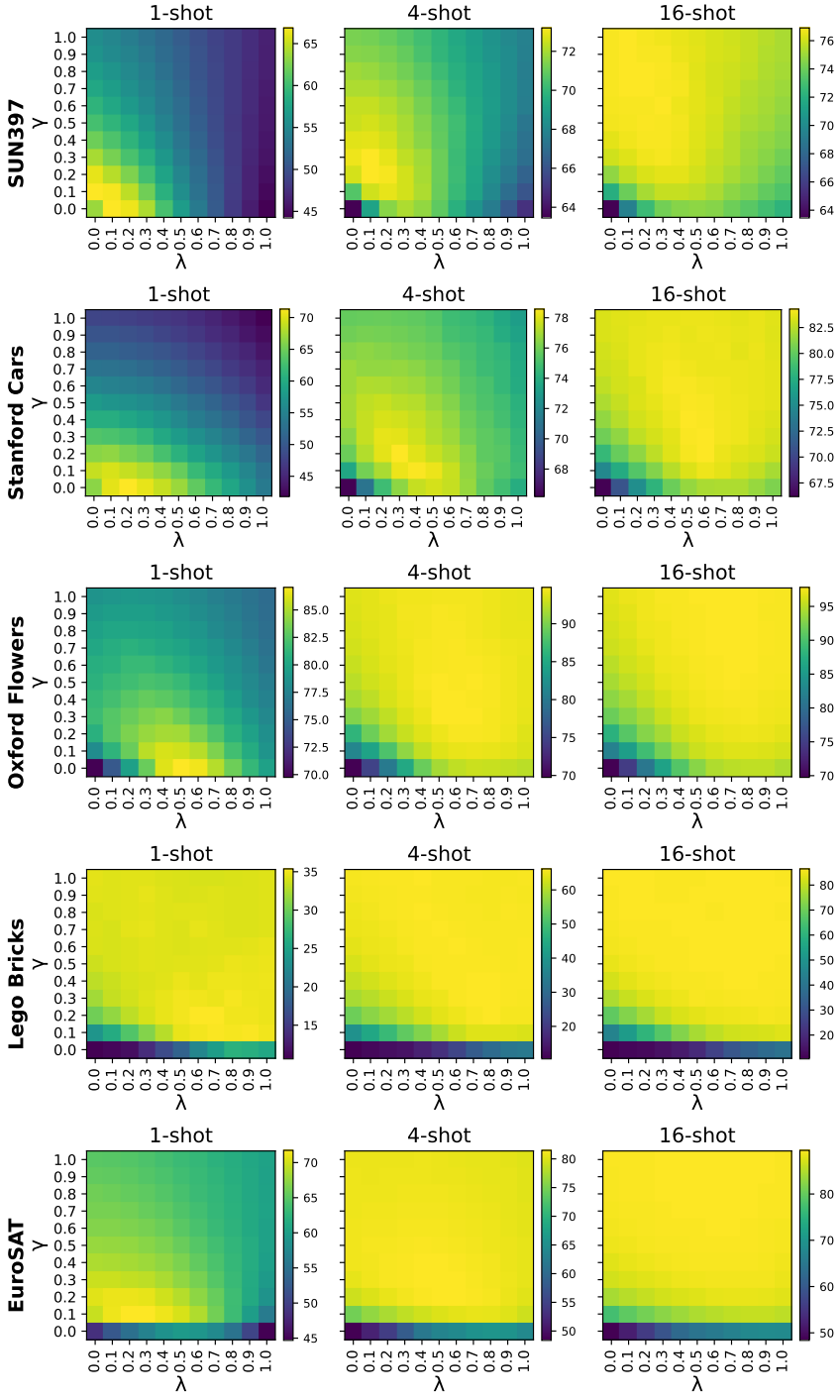

Figure 12: Analysis of the sensitivity of the hyper-parameters λ and γ to the validation set accuracy on 5 datasets with CLIP-ViT-B/16.