CEKALA: CKA-Guided Layer Selection for Efficient Vision–Language Adaptation

Tasnimul Hossain Tomal[1,*], **Md Fahim**[1,2,*,†],
Mir Sazzat Hossain[1,2,*], **Md Farhad Alam Bhuiyan**[1]

[1]Penta Global Limited, Bangladesh
[2]Center for Computational & Data Sciences
[*]Equal Contribution [†]Project Lead
Correspondence: {tomal11902012, fahimcse381}@gmail.com

Abstract

Pre-trained vision-language models (VLMs) such as CLIP offer strong generalization but face challenges in few-shot adaptation, particularly in identifying which layers to adapt and how to align cross-modal representations effectively. Existing multimodal adaptation methods uniformly apply adapters across fixed layers, assuming homogeneous layer importance and implicit depth-wise alignment between vision and text encoders. This assumption neglects layer-wise heterogeneity and cross-modal semantic misalignment. To overcome these limitations, we propose **Ce**ntered **K**ernel **A**lignment based **L**ayer **A**dapter (CEKALA), a representation measurement framework that leverages CKA to guide selective layer adaptation and cross-modal alignment. CEKALA first computes layer-wise CKA scores to quantify each layer's contribution to downstream performance, then identifies semantically aligned vision–text layer pairs based on CKA scores. Shared cross-modal adapters are injected only into aligned layer pairs, while unpaired layers receive modality-specific adapters, ensuring both semantic consistency and efficient parameter usage. CEKALA enables fine-grained, interpretable, and performance-aware layer selection for vision-language models. Empirical results demonstrate that CEKALA improves few-shot generalization and cross-modal alignment while maintaining strong parameter efficiency.

1 Introduction

The success of deep learning on massive datasets (Deng et al., 2009a; Russakovsky et al., 2015) has profoundly impacted computer vision, leading to exceptional results in core tasks like image classification (He et al., 2016; Huang et al., 2017; Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Szegedy et al., 2015; Zagoruyko & Komodakis, 2016), object detection (Girshick et al., 2014; He et al., 2017; Redmon et al., 2016; Ren et al., 2015; Liu et al., 2016), semantic segmentation (Chen et al., 2017; Long et al., 2015; Ronneberger et al., 2015; Zhao et al., 2017), and re-identification (Sun et al., 2018; Wang et al., 2018b; Zheng et al., 2015; Zhong et al., 2017). Building on this powerful foundation, Vision-Language Models (VLMs) (Jia et al., 2021; Li et al., 2021; Radford et al., 2021; Zhang et al., 2022a; Zhou et al., 2022b) have emerged as foundational models capable of unified comprehension by processing visual and textual data together.

Central to this progress is Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021). Trained contrastively (Oord et al., 2018) on over 400 million image-text pairs, CLIP learns a joint representation space that aligns semantically related concepts across modalities while pushing unrelated ones apart. This holistic approach, utilizing distinct encoders for vision and text, grants CLIP remarkable *zero-shot generalization* abilities across various downstream applications, including medical diagnostics (Huang et al., 2021; Muller et al.,

2022; Wang et al., 2022), image generation and description (Li et al., 2022; Mokady et al., 2021; Wang et al., 2021), and complex question answering (Gui et al., 2021; Hu et al., 2022; Su et al., 2019).

Despite their versatility, adapting VLMs to specific downstream tasks remains challenging. The sheer size of these models makes full fine-tuning computationally expensive and often results in *overfitting* (Khattak et al., 2023a; Yang et al., 2024; Guo & Gu, 2025), particularly in low-data (few-shot) regimes. To navigate this, adaptation techniques have become essential. Initially, adaptation relied on *manual prompt engineering* (Radford et al., 2021) (e.g., using "a photo of a [CLASS]"), which is time-consuming, requires specialized knowledge, and is sub-optimal. This led to the development of `prompt learning` (Lester et al., 2021). CoOp (Zhou et al., 2022b) introduced the concept of learning continuous, trainable prompt vectors that are optimized while the VLM backbone remains frozen, offering efficient dataset-specific adaptation. Subsequent work focused on refining prompt placement within the text encoder (Khattak et al., 2023b; Yao et al., 2023; Zhou et al., 2022b;a), the vision encoder (Jia et al., 2022; Zhu et al., 2023), or both, as seen in MaPLe (Khattak et al., 2023a), which proposed multi-modal deep prompts and a coupling function to align both encoders.

In parallel, *adapter-based methods* (Chen et al., 2022; Gao et al., 2021; Houlsby et al., 2019; He et al., 2022; Liu et al., 2022; Pfeiffer & Gurevych, 2020; Sung et al., 2022; Zhang et al., 2021) emerged as a structurally agnostic alternative. These approaches insert small, lightweight modules (e.g., MLPs) within the frozen VLM architecture to refine feature representations. Examples include CLIP-Adapter (Gao et al., 2021), which fine-tunes features via an MLP with residual connections, and AdaptFormer (Chen et al., 2022), which places adapters in transformer blocks. Recently, the field has converged on *multi-modal adaptation*. MMA (Yang et al., 2024) proposed an adapter that refines cross-modal alignment by aggregating features across modalities, noting that lower layers encode generalizable novel-class features while higher layers encode discriminative base-class features.

Existing multimodal prompt learning methods such as MMRL, MaPLe, and MMA (Guo & Gu, 2025; Khattak et al., 2023a; Yang et al., 2024) typically adopt a uniform adapter insertion strategy, where adapters are applied starting from a manually chosen layer index J and extended to deeper layers. While simple and effective in some cases, this depth-based heuristic overlooks the intrinsic structure of learned representations across layers. This uniform strategy suffers from several key limitations: 1) it assumes deeper layers contribute uniformly despite the heterogeneous nature of representations across depth; 2) it overlooks redundancy across layers, failing to identify which layers provide unique information and thus leading to inefficient adaptation; and 3) it ignores cross-modal alignment, missing layers that encode shared semantics between modalities and thereby limiting effective multimodal adaptation.

To address these limitations, we propose ***Ce**ntered **K**ernel **A**lignment-based **L**ayer **A**dapter (*CeKALA*)*, a principled and data-driven framework for selective adapter placement in Vision–Language Models. Our approach leverages Centered Kernel Alignment (CKA) to analyze the similarity structure of representations across layers. Specifically, we first compute layer-wise representations over a dataset and measure pairwise similarities using CKA. Based on this, we define a *diversity score* for each layer, which captures how dissimilar a layer is from the rest of the network. Layers with higher diversity scores encode more unique information and are therefore selected for adaptation. We independently perform this selection for the image and text encoders to obtain modality-specific informative layers. Furthermore, we identify *multimodal layers* as those that are selected in both encoders, indicating shared representational importance. These layers are natural candidates for cross-modal adaptation. The remaining layers are treated as modality-specific, enabling targeted unimodal adaptation.

We performed comprehensive evaluations across three rigorous settings—base-to-novel generalization (11 datasets), cross-dataset evaluation (10 target datasets), and domain generalization (4 ImageNet variants)—demonstrating CeKALA's superior adaptation and generalization capabilities by consistently achieving the best results. Specifically, CeKALA obtained an average 80.45% Harmonic Mean (HM) on base-to-novel generalization, leading on 10 out of 11 datasets; a 67.65% average accuracy on cross-dataset evaluation, leading

on 6 out of 10 target datasets; and secured the best results on 3 out of 4 ImageNet domain generalization variants.

2 Related Work

2.1 Efficient Transfer Learning for VLMs

Previous research (Yang et al., 2024; Guo & Gu, 2025; Khattak et al., 2023a) has explored the role of efficient transfer learning in VLMs over fully finetuning. Work in this direction can be broadly categorized into two main approaches: prompt learning and adapter-based learning.

Prompt Learning aims to adapt pretrained VLMs by introducing learnable tokens or templates, guiding the model to interpret new tasks more effectively without extensive parameter updates. CoOp (Zhou et al., 2022b) first replaces handcrafted textual templates with continuous, learnable prompt vectors. CoCoOp (Zhou et al., 2022a) conditions prompts on visual inputs, enabling instance-specific adaptation and improved robustness to domain shifts. Further ProDA (Lu et al., 2022) model prompt distributions for enhanced adaptability, while KgCoOp (Yao et al., 2023) preserves general textual knowledge through divergence minimization. MaPLe (Khattak et al., 2023a) couple visual and textual prompts for stronger cross-modal synergy. ProVP (Xu et al., 2025) leverages visual prompts aligned with CLIP's latent space. Regularization-based approaches like PromptSRC (Khattak et al., 2023b) and RPO (Lee et al., 2023) mitigate overfitting and representation drift, while MetaPrompt (Zhao et al., 2024) and TCP (Yao et al., 2024) introduce meta-learning and class-aware strategies to enhance prompt generalization.

Adapter-Based Learning methods offer another efficient alternative to full fine-tuning by introducing lightweight trainable modules within frozen pretrained networks. In VLMs, CLIP-Adapter (Gao et al., 2024) integrates a two-layer MLP after the image encoder to refine feature representations, whereas Tip-Adapter (Zhang et al., 2022b) leverages cached training features to accelerate inference and improve similarity computation. More advanced frameworks, such as MMA (Yang et al., 2024) introduces a new class of multimodal adapters to address the modality gap by jointly attending to intermediate representations from both modalities in each adapter block.

2.2 Similarity Measure in Deep Learning & CKA

A wide range of similarity measures has been proposed to compare representations in deep neural networks. Early approaches based on neuron alignment and correlation (Li et al., 2015; Wang et al., 2018a) are not permutation invariant and fail to capture distributed representations. Later, CCA-based methods such as SVCCA (Raghu et al., 2017) and PWCCA (Morcos et al., 2018) compare subspaces but are invariant to arbitrary invertible linear transformations, which can make them unable to distinguish representations when the dimensionality exceeds the number of samples. Kernel-based methods such as HSIC (Gretton et al., 2005), which measure dependence between representations via kernelized covariance, address some of these issues but are not invariant to isotropic scaling, making comparisons across layers difficult. To overcome this, CKA (Kornblith et al., 2019) normalizes HSIC, yielding a scale-invariant and orthogonally invariant similarity measure. Importantly, CKA can reliably identify layer correspondences across different architectures and initializations, whereas prior methods often fail.

Recent works (Davari et al., 2022; Cloos et al., 2024; Sevetlidis & Pavlidis, 2026) highlight limitations of CKA, such as sensitivity to certain transformations and outliers, and propose geometry-aware alternatives. Nevertheless, CKA is widely used for its simplicity, scalability, and strong empirical reliability, making it a standard for representation similarity analysis.

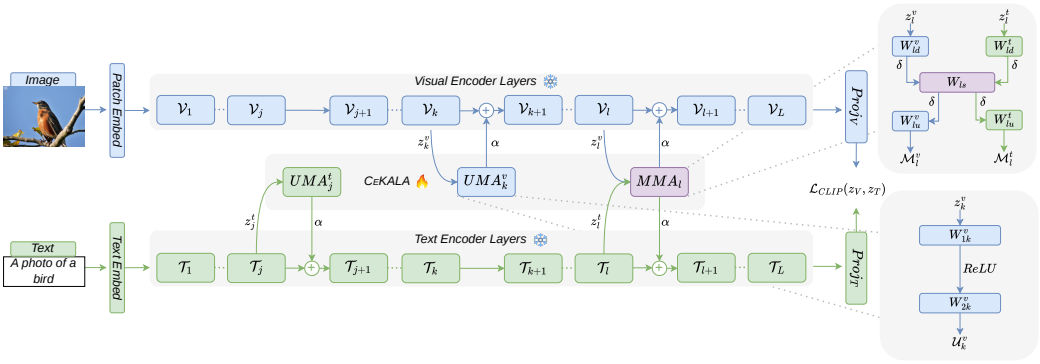


Figure 1: Model architecture of our proposed CEKALA.

3 Methodology

Following most of existing studies (Khattak et al., 2023a; Zhou et al., 2022b;a; Yao et al., 2024; Yang et al., 2024; Guo & Gu, 2025), we base on the pre-trained transformer-based CLIP models (Radford et al., 2021). The preliminaries about CLIP, CKA score and why we choose CKA over Cosine-Sim & CCA are described in Appendix A.

Our proposed method, CEKALA, is specifically designed to address the critical challenge of selectively applying adapters within VLMs. Rather than blindly applying adapters in a group of layers, we apply our algorithm to select the most important layers in the VLMs that need to be fine-tuned based on CKA. The overview of our method is provided in Fig 1.

3.1 CKA-based Layer Selection for Adapter Placement

Let $\mathcal{E}$ denote an encoder (either image encoder $\mathcal{E}_{img}$ or text encoder $\mathcal{E}_{txt}$) with L layers. Given a dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$, we compute the hidden representations at each layer:

$$\{H_1, H_2, \ldots, H_L\}, \quad H_i \in \mathbb{R}^{n \times d_i},$$

where each H_i corresponds to the activations of the i-th layer across all samples.

To quantify the uniqueness of each layer, we measure its similarity with all other layers using CKA. For a given layer i, we define its *diversity score* as:

$$\text{Score}_i = \frac{1}{L-1} \sum_{j=1, j\neq i}^{L} \left(1 - \text{CKA}(H_i, H_j)\right).$$

Intuitively, $\text{CKA}(H_i, H_j)$ measures how similar two layers are. Therefore, $1 - \text{CKA}(H_i, H_j)$ captures their dissimilarity. A high Score_i indicates that layer i is, on average, less similar to other layers, and thus encodes more distinct information.

Top-k Layer Selection. We rank all layers according to their diversity scores and select the top-k layers:

$$\mathbf{L} = \text{TopK}\left(\{\text{Score}_i\}_{i=1}^{L}, k\right).$$

This procedure is applied independently to the image and text encoders, yielding two sets of informative layers: $\tilde{\mathbf{L}_{img}}, \tilde{\mathbf{L}_{txt}}$.

Multimodal and Modality-Specific Decomposition. A key insight of our approach is that some layers may be important in *both* modalities, indicating potential cross-modal

alignment. To capture this, we partition the selected layers into three groups:

$$\begin{aligned}\mathbf{L}_{mm} &= \tilde{\mathbf{L}_{img}} \cap \tilde{\mathbf{L}_{txt}} \quad \text{(multimodal layers)},\\ \mathbf{L}_{img} &= \tilde{\mathbf{L}_{img}} \setminus \mathbf{L}_{mm} \quad \text{(image-specific layers)},\\ \mathbf{L}_{txt} &= \tilde{\mathbf{L}_{txt}} \setminus \mathbf{L}_{mm} \quad \text{(text-specific layers)}.\end{aligned}$$

The multimodal set $\mathbf{L}_{mm}$ captures layers that are simultaneously informative in both encoders, making them natural candidates for cross-modal adaptation. The remaining layers correspond to modality-specific processing stages. The overall algorithm for CKA based layer selection in VLMs is framed in Algorithm 1. Layer-wise CKA score for each dataset is provided in the Appendix F.

3.2 Design Choice of the Adapters

Multimodal Adapters. The most influential layers that are common in both modality requires mutlimodal adapter techniques as shared representation learning is necessary to facilitate effective cross-modal interaction and fusion, aligning the features derived from both vision and text encoders ($\mathbf{L}_{mm}$). For the multimodal adapters, we consider similar approach used in MMA (Yang et al., 2024). For the layers $l \in \mathbf{L}_{mm}$, we apply multimodal adapter as follows:

$$\begin{aligned}\mathcal{M}_l^v(z_l^v) &= W_{lu}^v \cdot \delta(W_{ls} \cdot \delta(W_{ld}^v \cdot z_l^v)) \quad \text{here } z_l^v = [c_l, \mathbf{E}_l]\\ \mathcal{M}_l^t(z_l^t) &= W_{lu}^t \cdot \delta(W_{ls} \cdot \delta(W_{ld}^t \cdot z_l^t)) \quad \text{here } z_l^t = [\mathbf{t}_l]_{j=1}^N\end{aligned}$$

Each modality branch first projects its input into a unified feature space using a branch-specific layer. The shared layer W_{ls} bridges the visual and textual pathways, enabling cross-modal gradient flow and improved feature alignment. Then it is followed by another modality-specific layer that aligns the output dimensions of each branch.

Unimodal Adapters. The layers in $\mathbf{L}_m$ where $m \in \{img, text\}$ clearly indicates that those layers are influential for adaptation to gain more modality specific fine-grained information. For those layers, we apply simple but yet unimodal adapters architecture inspiring from CLIP-Adapter (Gao et al., 2024) as follows:

$$\begin{aligned}\mathcal{U}_l^v(z_l^v) &= \text{ReLU}(\mathbf{V}_l^\top W_{1l}^v) W_{2l}^v \quad \text{here } z_l^v = [c_l, \mathbf{E}_l]\\ \mathcal{U}_l^t(z_l^t) &= \text{ReLU}(\mathbf{T}_l^\top W_{1l}^t) W_{2l}^t \quad \text{here } z_l^t = [\mathbf{t}_l]_{j=1}^N\end{aligned}$$

where W_{1l}^m and W_{2l}^m are two learnable weights in layer $l \in \mathbf{L}_m$ for $m-$ modality encoder. $\mathbf{V}_l$ and $\mathbf{T}_l$ is the frozen CLIP's representation at layer l described in Section A.

3.3 CEKALA Overall Framework

Finally, to obtain the adapted and enhanced multimodal representations, we fuse the adapter outputs with the original *CLIP frozen representation*. For fusion, we adopt the residual-like connection design inspired by MMA (Yang et al., 2024).

For the *vision encoder*, the layer-wise update mechanism for the class token (c_l) and patch tokens ($\mathbf{E}_l$) is defined as:

$$\begin{aligned}[c_l, \mathbf{E}_l] &= \mathcal{V}_l([c_{l-1}, \mathbf{E}_{l-1}]), \quad l \notin \tilde{\mathbf{L}_{img}}\\ [c_l, \mathbf{E}_l] &= \mathcal{V}_l([c_{l-1}, \mathbf{E}_{l-1}]) + \alpha \cdot \mathcal{U}_l^v([c_{l-1}, \mathbf{E}_{l-1}]), \quad l \in \mathbf{L}_{img}\\ [c_l, \mathbf{E}_l] &= \mathcal{V}_l([c_{l-1}, \mathbf{E}_{l-1}]) + \alpha \cdot \mathcal{M}_l^v([c_{l-1}, \mathbf{E}_{l-1}]), \quad l \in \mathbf{L}_{mm}\end{aligned}$$

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
ProDA	81.56	72.30	76.65	75.40	70.23	72.72	98.27	93.23	95.68	95.43	97.83	96.62
KgCoOp	80.73	73.60	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
MaPLe	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
LASP	82.70	74.90	78.61	76.20	70.95	73.48	98.10	94.24	96.16	95.90	97.93	96.90
RPO	81.13	75.00	77.78	76.60	**71.57**	74.00	97.97	94.37	96.03	94.63	97.50	96.05
PromptSRC	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
ProVP	85.20	73.22	78.76	75.82	69.21	72.36	98.92	94.21	96.51	**95.87**	97.65	96.75
MetaPrompt	83.65	75.48	79.09	77.52	70.83	74.02	98.13	94.58	96.32	95.53	97.00	96.26
TCP	84.13	75.36	79.51	77.27	69.87	73.38	98.23	94.67	96.42	94.67	97.20	95.92
MMA	83.20	76.80	79.87	77.31	71.00	74.02	**98.40**	94.00	96.15	95.40	98.07	96.72
CEKALA	83.37	77.65	80.06	77.83	70.48	73.68	**98.40**	**95.74**	**97.05**	95.46	98.11	**96.77**
CEKALA+ UMA	**83.63**	**78.11**	**80.45**	**78.10**	71.55	**74.40**	98.28	94.65	96.43	95.18	**98.19**	96.67

Method	StanfordCars			Flowers102			Food101			FGVC-Aircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	63.37	74.89	68.65	72.08	**77.80**	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
ProDA	74.70	71.20	72.91	97.70	68.68	80.66	90.30	88.57	89.43	36.90	34.13	35.46
KgCoOp	71.76	75.04	73.36	95.00	74.73	83.65	90.50	91.70	91.09	36.21	33.55	34.83
MaPLe	72.94	74.00	73.47	95.92	72.46	82.56	90.71	**92.05**	91.38	37.44	35.61	36.50
LASP	75.17	71.60	73.34	97.00	74.00	83.95	**91.20**	91.70	**91.44**	34.53	30.57	32.43
RPO	73.87	75.53	74.69	94.13	76.67	84.50	90.33	90.83	90.58	37.33	34.20	35.70
PromptSRC	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	**40.15**
ProVP	80.43	67.96	73.67	98.42	72.06	83.20	90.32	90.91	90.61	**47.08**	29.87	36.55
MetaPrompt	76.34	75.01	75.48	97.66	74.49	84.52	90.74	91.85	91.29	40.14	36.51	38.24
TCP	**80.80**	74.13	77.32	97.73	75.57	85.23	90.57	91.37	90.97	41.97	34.43	37.83
MMA	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
CEKALA	80.13	**76.27**	**77.60**	98.01	77.70	86.72	90.48	91.92	91.19	40.47	37.59	38.36
CEKALA+ UMA	79.35	75.11	76.62	**98.15**	**77.92**	**86.91**	90.39	91.57	90.98	41.73	**38.47**	39.41

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
ProDA	78.67	76.93	77.79	80.67	56.48	66.44	83.90	66.00	73.88	85.23	71.97	78.04
KgCoOp	80.29	76.53	78.36	77.55	54.99	64.35	85.64	64.34	73.48	82.89	76.67	79.65
MaPLe	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
LASP	80.70	78.60	79.63	81.40	58.60	68.14	94.60	77.78	85.36	84.77	78.03	81.26
RPO	80.60	77.80	79.18	76.70	62.13	68.61	86.63	68.97	76.79	83.67	75.43	79.34
PromptSRC	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
ProVP	80.67	76.11	78.32	83.95	59.06	69.34	**97.12**	72.91	83.29	**88.56**	75.55	81.54
MetaPrompt	82.26	79.04	80.62	83.10	58.05	68.35	93.53	75.21	83.38	85.33	77.72	81.35
TCP	**82.63**	78.20	80.35	82.77	58.07	68.25	91.63	74.73	82.32	87.13	80.87	83.83
MMA	82.27	78.57	80.38	83.20	65.63	73.38	85.46	**82.34**	83.87	86.23	80.03	82.20
CEKALA	82.41	**79.10**	80.72	83.20	63.21	71.40	84.47	**81.35**	82.87	86.18	82.63	84.32
CEKALA+ UMA	82.58	79.05	**80.77**	**83.78**	**66.85**	**73.92**	86.26	82.03	**84.00**	86.09	**83.77**	**84.88**

Table 1: Comparison with PEFT methods on Base-to-Novel Generalization setting. The base and novel class classification accuracies and their harmonic mean (HM) have been provided. HM quantifies the trade-off between adaptation and generalization.

Similarly, for the *text encoder*, the overall framework defining the layer-wise update for the sequence of text tokens ($\mathbf{t}_l$) is:

$$
\begin{aligned}
[\mathbf{t}_l]_{j=1}^{N} &= \mathcal{T}_l([\mathbf{t}_{l-1}]_{j=1}^{N}), \quad l \notin \mathbf{L}_{\tilde{text}} \\
[\mathbf{t}_l]_{j=1}^{N} &= \mathcal{T}_l([\mathbf{t}_{l-1}]_{j=1}^{N}) + \alpha \cdot \mathcal{U}_l^t([\mathbf{t}_{l-1}]_{j=1}^{N}), \quad l \in \mathbf{L}_{text} \\
[\mathbf{t}_l]_{j=1}^{N} &= \mathcal{T}_l([\mathbf{t}_{l-1}]_{j=1}^{N}) + \alpha \cdot \mathcal{M}_l^t([\mathbf{t}_{l-1}]_{j=1}^{N}), \quad l \in \mathbf{L}_{mm}
\end{aligned}
$$

Here, $\mathcal{V}_l$ and $\mathcal{T}_l$ denote the l-th frozen layers of the vision and text encoders, respectively. $\mathcal{U}$ and $\mathcal{M}$ correspond to the unimodal and multimodal adapters, while α is a learnable scaling factor that controls the contribution of the adapter outputs. In our setup, **CEKALA** refers to integrating only the multimodal adapters into the selected layers of the *CLIP* model,

whereas **CEKALA+ UMA** denotes the incorporation of both unimodal and multimodal adapters within those layers.

4 Results

Details on implementation, datasets, and computational cost are provided in the **Appendix** C and D.

Methods	ImageNet	**Average**	Caltech	OxfordPets	Stan Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101
CoOp	71.51	63.88	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
CoCoOp	71.02	65.74	**94.43**	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21
MaPLe	70.72	66.30	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69
PromptSRC	71.27	65.81	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75
TCP	71.40	66.29	93.97	91.25	64.69	71.21	86.69	23.45	67.15	44.35	**51.45**	68.73
MMA	71.00	66.61	93.80	90.30	66.13	72.07	86.12	25.33	**68.17**	46.57	49.24	68.32
CEKALA	72.48	66.20	92.66	91.04	65.23	72.77	86.14	24.93	66.93	46.80	47.15	68.34
CEKALA+ UMA	**73.46**	**67.41**	93.75	**92.79**	**66.82**	**73.09**	**86.92**	**25.61**	67.73	**47.34**	51.06	**69.03**

Table 2: Comparison with state-of-the-art methods on the Cross-Dataset Evaluation setting. Here, ImageNet is the source dataset, and the rightmost 10 columns are the target datasets. The average is taken across the target datasets.

Method	Source	Target			
	ImageNet	-V2	-S	-A	-R
CLIP	66.73	60.83	46.15	47.77	73.96
CoOp	71.51	64.20	47.99	49.71	75.21
CoCoOp	71.02	64.07	48.75	50.63	76.18
MaPLe	70.72	64.07	49.15	50.90	76.98
PromptSRC	71.27	64.35	49.55	50.90	77.80
MMA	71.00	64.33	49.13	**51.12**	77.32
CEKALA	72.48	63.52	48.87	46.80	76.85
CEKALA+ UMA	**73.46**	**65.12**	**51.18**	47.63	**78.07**

Table 3: Comparison SOTA PEFT methods for CLIP on domain generalization across 4 ImageNet variants: V2, Sketch (S), A, and R.

SigLIP	Base	Novel	HM
Base	74.28	70.46	72.32
MMA	78.62	72.70	75.54
CEKALA	80.05	73.25	76.50
CEKALA+ UMA	81.24	73.87	77.38

Table 4: CEKALA on SigLIP Model on ImageNet Dataset

Metrics	Base	Novel	HM
Cosine	82.79	75.66	78.88
CCA	82.53	75.20	78.49
Linear CKA	**83.37**	**77.65**	**80.06**
RBF CKA	**83.37**	**77.65**	**80.06**

Table 5: Ablation of Metrics

4.1 Base-to-Novel Generalization

Table 1 presents the results of CEKALA on the base-to-novel generalization task, where the model is trained on a set of base classes and tested on both base and novel classes, following the setup in prior works (Khattak et al., 2023a; Zhou et al., 2022b;a). We evaluate our method on 11 diverse image classification datasets: ImageNet (Deng et al., 2009b) and Caltech101 (Fei-Fei et al., 2004) for general object recognition; OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), and FGVC-Aircraft (Maji et al., 2013) for fine-grained classification; SUN397 (Oliva et al., 2010) for scene recognition; DTD (Cimpoi et al., 2014) for texture classification; EuroSAT (Helber et al., 2019) for satellite image recognition; and UCF101 (Soomro et al., 2012) for action recognition. This task allows us to assess CEKALA's transfer learning effectiveness on base classes and its ability to preserve the inherent generalization and zero-shot capabilities of pre-trained VLMs on novel classes. CEKALA is compared against several strong baselines (1) where it consistently demonstrates superior performance, evidenced by the following observations:

Generalization and Overall Performance: CEKALA achieves the highest average harmonic mean (HM) of 80.06% across all 11 datasets, surpassing prior methods such as MMA, which reports an HM of 79.87%. Moreover, incorporating the unimodal adapter (CEKALA+

UMA) further improves the overall performance, reaching an HM of 80.45%. In terms of generalization, CEKALA attains an average novel accuracy of 77.65%, while CEKALA+ UMA further raises it to 78.11%, indicating stronger generalization to unseen (novel) classes.

Base Class Performance: On seen classes, CEKALA maintains competitive performance with an average base accuracy of 83.37%, slightly exceeding MMA (83.20%). The addition of UMA further improves this to 83.63%. Notably, CEKALA achieves top performance on datasets such as Caltech101 (98.40%) while remaining competitive on others like ImageNet (77.83%), indicating that the proposed approach preserves strong representation learning for base categories.

Novel Class Performance: The primary strength of the CEKALA model lies in its ability to generalize to novel classes, which is crucial for effective transfer learning. CEKALA achieves the highest average novel class accuracy of 77.65% across all 11 datasets, surpassing strong baselines such as MMA (76.80%). With UMA, this further increases to 78.11%, demonstrating additional gains in unseen-class recognition. This superiority confirms that CEKALA's underlying mechanism for multimodal adaptation and prompt learning effectively preserves and enhances the VLM's capacity to generalize robustly to the new classes, avoiding catastrophic forgetting.

Dataset-Specific Highlights: CEKALA demonstrates strong performance across individual datasets, achieving the highest harmonic mean on a majority of the benchmarks. Notable gains are observed on Caltech101 (97.05%), UCF101 (84.32%), and StanfordCars (77.60%), along with consistent performance on challenging datasets such as SUN397 and EuroSAT. Importantly, incorporating UMA further boosts performance on datasets like EuroSAT (82.87% $\rightarrow$ 84.00%) and UCF101 (84.32% $\rightarrow$ 84.88%), highlighting its effectiveness across diverse domains. These results indicate that CEKALA effectively adapts across fine-grained, scene, and action recognition tasks.

Performance Nuances: Despite the strong overall performance of CEKALA, certain datasets remain challenging. In particular, on DTD, CEKALA achieves an HM of 71.40%, which is approximately 2.5% lower than the best-performing variant (73.92%). This can be attributed to the texture-centric nature of DTD, which requires capturing fine-grained local patterns that are less aligned with global semantic representations. Similarly, datasets like Caltech101 involve high inter-class similarity, making discrimination more subtle. These observations suggest that while CEKALA generalizes well overall but room for improvement in handling highly fine-grained and texture-dominated scenarios.

(a) Hidden Layers Dim

Dims	Base	Novel	HM
8	81.36	75.66	78.41
16	82.82	76.87	79.73
32	82.77	77.97	79.58
64	**83.63**	**78.11**	**80.06**
128	83.81	77.72	79.93

(b) Scaling Factor α

α	Base	Novel	HM
0.005	78.64	75.26	76.91
0.01	82.73	76.79	79.80
0.05	**83.63**	**78.11**	**80.06**
0.1	82.67	77.12	79.65
0.5	83.58	76.92	80.11

(c) Top k layers Ablation

k	Base	Novel	HM
2	81.25	75.36	78.19
4	82.56	76.74	79.54
6	83.63	**78.11**	**80.06**
8	**83.91**	76.53	80.04
10	82.92	77.25	79.98

Table 6: Ablations over CEKALA modules from §3 on the 11 datasets used in the Base-to-Novel Generalization setting.

4.2 Cross-dataset Evaluation

To evaluate the transferability of CEKALA to entirely new domains, we conduct cross-dataset evaluation following CoCoOp (Zhou et al., 2022a), where models are trained on ImageNet in a few-shot setting and directly evaluated on 10 unseen datasets without adaptation. As shown in Table 2, CEKALA achieves an average accuracy of 66.20%, while CEKALA+ UMA improves this to 67.41%, outperforming all competing methods. The improvements are consistent across diverse domains, with CEKALA+ UMA achieving top performance on several datasets, including OxfordPets (92.79%), StanfordCars (66.82%), Flowers102 (73.09%), Food101 (86.92%), Aircraft (25.61%), DTD (47.34%), and UCF101

(69.03%). These results demonstrate the strong cross-domain generalization ability of CEKALA and highlight the effectiveness of the unimodal adapter in enhancing transferability without requiring additional fine-tuning.

4.3 Domain Generalization

We evaluate the resilience of models to domain shifts and generalization to out-of-distribution (OOD) data. Following CoCoOp (Zhou et al., 2022a), the models, initially trained on ImageNet, are evaluated directly on four ImageNet variants, ImageNet-V2 (Shankar et al., 2021), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a), each introducing a distinct domain variation. Table **??** summarizes the results, where CEKALA+ UMA achieves the best performance on three out of four datasets, including ImageNet-V2 (65.12%), ImageNet-Sketch (51.18%), and ImageNet-R (78.07%), while remaining competitive on ImageNet-A. These results highlight its robustness and adaptability to shifts in the distribution of the training data.

4.4 Ablation Study

We conduct an extensive ablation study (summarized in Table 6) across the 11 datasets used in the base-to-novel generalization setting to validate the necessity and optimal configuration of CEKALA's core components. We consider CEKALA (without **UMA**)for this ablation study.

Adapter Hidden Dimensions (Table 6a). We analyze the impact of the bottleneck dimension of our adapters. We find that the optimal dimension is 64, achieving the highest 80.06% HM. Smaller dimensions (e.g., 8 or 16) underfit the complex multimodal relationships, limiting performance. While increasing the dimension beyond 64 (to 128) provides marginal benefit to Base class accuracy (83.81%) it slightly degrades the critical Novel class performance (77.72%).

Scaling Factor α (Table 6b). The scaling factor α is a hyper parameter designed to balance the contribution of the adapter's output (task-specific features) with the frozen CLIP backbone's features (task-agnostic knowledge). We find the optimal performance is achieved at $\alpha = 0.05$ (HM 80.06%). Extremely small values like 0.005 undermine the adapter, leading to lower performance, while larger values like 0.5 drastically reduce generalization by allowing the adapter to override the valuable pre-trained knowledge of the frozen VLM features.

Top k Layers Ablation (Table 6c). This ablation investigates the impact of the number of top-influential layers (k) selected for adaptation. We observe a clear trend: increasing k generally improves performance up to $k = 6$, which yields the maximum 80.06% HM. We also get almost same result when $k = 8$ with the best Base score (83.91%). Increasing k further to 10 slightly harms the model performance and also increase the trainable parameters.

Ablation of Similarity Metrics (Table 5). We also measure the influence of various similarity metrics like Cosine Similarity, Canonical Correlation Analysis (CCA) and two form of CKA Kernel (Linear & RBF) for the layer selection. We found that CKA performs better than Cosine & CCA and Linear-CKA & RBF-CKA both perform identical.

Performance of CEKALA on SigLIP (Table 5. We further evaluate the performance of CEKALA across additional models, including CLIP. In particular, we conduct an ablation study on SigLIP (Zhai et al., 2023), where we observe a consistent trend: MMA outperforms the base SigLIP model, while CEKALA further surpasses MMA, achieving improvements of 1.27% and 5.78% over the base model. Moreover, the combination of CEKALA with UMA also exceeds MMA, yielding gains of 2.44% and 7% compared to the baseline.

5 Conclusion

We present CEKALAthat advances vision-language model adaptation through three key innovations: (i) CKA score-based layer selection that identifies which layers most critically

impact downstream performance, (ii) a principled separation of unimodal and multimodal adaptation pathways, and (iii) specialized adapter designs tailored to each pathway's distinct role. By quantifying layer-wise influence patterns, CEKALA strategically allocates adaptation capacity where it is most needed, achieving the best performance across 13 (10 regular and 3 domain generalization datasets) out of 15 total datasets evaluated, significantly outperforming prior baselines. Our results demonstrate that CKA-aware adaptation rather than uniform layer treatment, is key to effectively balancing task-specific discrimination with zero-shot generalization.

References

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 446–461, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 839–847, 2017.

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

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Nathan Cloos, Moufan Li, Markus Siegel, Scott L. Brincat, Earl K. Miller, Guangyu Robert Yang, and Christopher J. Cueva. Differentiable optimization of similarity scores between models and brains, 2024. URL https://arxiv.org/abs/2407.07059.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of cka as a similarity measure in deep learning. *arXiv preprint arXiv:2210.16156*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009a.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009b.

Li Fei-Fei, R. Fergus, and P. 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, 2004. doi: 10.1109/CVPR.2004.383.

Peng Gao, Renjie Zhou, Junjie Ma, Rui Li, and Hongyuan Yu. Clip-adapter: Better vision-language models with feature adapters. In *International Conference on Machine Learning (ICML)*, pp. 3445–3455, 2021.

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.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita (eds.), *Algorithmic Learning Theory*, pp. 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31696-1.

Lin Gui, Haoming Li, Xueting Wang, Yiyi Wang, Xuan Zhang, and Yizhou Wang. Kat: Knowledge-augmented transformer for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11880–11889, 2021.

Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25015–25025, 2025.

Kai He, Xin Zhang, Yuxin Yu, Chen Chen, and Zi Liu. Unified parameter-efficient fine-tuning for vision tasks. In *International Conference on Computer Vision (ICCV)*, pp. 1–11, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

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. doi: 10.1109/JSTARS.2019.2918242.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, 2021a. doi: 10.1109/ICCV48922.2021.00823.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15257–15266, 2021b. doi: 10.1109/CVPR46437.2021.01501.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Shizhe Hu, Jianfeng Lu, Hui Liu, Junnan Li, and Xu Zhou. Scaling vision-language models for visual question answering. In *European Conference on Computer Vision (ECCV)*, pp. 477–494, 2022.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

Zifeng Huang, Zihang Zhang, Yezhen Cui, Junyi Wang, Ya Zhang, Kai Ma, Xiaolu Zheng, Florin Ghesu, Rencheng Wang, and Jianbo Li. Gloria: A multimodal global-local representation learning for medical images and reports. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 413–423, 2021.

Chenfei Jia, Ye Yang, Yin Xia, Yi Chen, Teng Zhang, Shari Ren, Scott E Oord, Huaping Li, and Hua Feng. Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.

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

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, 2023a. doi: 10.1109/CVPR52729.2023.01832.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15190–15200, 2023b.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, 2012.

Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1401–1411, 2023.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Junnan Li, Dongxu Li, Santiago Savarese, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 12297–12313, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C H Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pp. 12888–12900, 2022.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar (eds.), *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pp. 196–212, Montreal, Canada, 11 Dec 2015. PMLR. URL https://proceedings.mlr.press/v44/li15convergent.html.

Haoming Liu, Meng Chen, Yifei Wang, and Zi Liu. Few-shot adaptation for diffusion models with clip-conditioned latent space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 31090–31102, 2022.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Ionescu, Scott E Oord, Satheesh Shrivastava, and Vittorio Ferrari. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, pp. 21–37, 2016.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5206–5215, 2022.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. URL https://arxiv.org/abs/1306.5151.

Ron Mokady, Amir Hertz, Idan Hertz, Ofir Tsafrir, Igal Drori, and Naftali Tishby. Clipcap: Clip prefix for image captioning. In *International Conference on Computer Vision (ICCV)*, pp. 1–11, 2021.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a7a3d70c6d17a73140918996d03c014f-Paper.pdf.

Nicolas Muller, Yi Chen, Chen Lu, Sotirios A Tsaftaris, and Alejandro F Frangi. Joint image-text representation learning for medical report generation. *IEEE Transactions on Medical Imaging*, 41(8):1939–1950, 2022.

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, 2008. doi: 10.1109/ICVGIP.2008.47.

Aude Oliva, Krista A. Ehinger, Antonio Torralba, James Hays, and Jianxiong Xiao. SUN database: Large-scale scene recognition from abbey to zoo . In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, Los Alamitos, CA, USA, June 2010. IEEE Computer Society. doi: 10.1109/CVPR.2010.5539970. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2010.5539970.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *International Conference on Machine Learning (ICML)*, pp. 5192–5200, 2018.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

Jonas Pfeiffer and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6898–6909, 2020.

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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Vasileios Sevetlidis and George Pavlidis. Gauge-invariant representation holonomy. *arXiv preprint arXiv:2601.21653*, 2026.

Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9641–9649, 2021. doi: 10.1109/ICCV48922.2021.00952.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. URL https://arxiv.org/abs/1212.0402.

Weijie Su, Xizhou Zhu, Yinpeng Lu, Bin Ye, Xiaoguang Mo, Junfeng Lu, Hongsheng Hu, Yong Li, and Jian Sun. Vl-bert: Pre-training of vision-language transformers with anker points. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Yifan Sun, Liang Zheng, Weijian Deng, Qixiang Ye, Shengjin Liu, and Daochang Jiao. Beyond part models: Deep local features for person re-identification. In *European Conference on Computer Vision (ECCV)*, pp. 480–496, 2018.

Kyungho Sung, Haeyun Kim, Myunghyun Kim, and Junmo Jo. Lst-adapter: Low-rank subspace transfer adapter for vision-language models. In *European Conference on Computer Vision (ECCV)*, pp. 192–208, 2022.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. *Learning robust global representations by penalizing local predictive power*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5fc34ed307aac159a30d81181c99847e-Paper.pdf.

Peng Wang, Lin Yan, Junping Yan, Sheng Cheng, and Kai Ma. Medclip: Contrastive learning of medical image and report from limited labeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 29881–29892, 2022.

Weizhe Wang, Yanpei Zhao, Yandong Li, Rui Li, Jinglong Zhang, Xiaojun Su, Dongyan Zhu, Haifeng Wang, and Shijie Qiao. Simvlm: Simple visual language model pretraining with uniformed masking and prefix tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 25011–25022, 2021.

Xiangping Wang, Yifan Sun, Min Liu, Jiaxin He, and Weijian Deng. Learning discriminative features with cross-batch triplet loss for person re-identification. *IEEE Transactions on Image Processing*, 27(10):4750–4762, 2018b.

Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation. *International Journal of Computer Vision*, 133(2):511–526, 2025.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23826–23837, 2024.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6757–6767, 2023.

Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23438–23448, 2024.

Sergey Zagoruyko and Nikolaos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

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.

Renrui Zhang, Yi Zhang, Yi Chen, Teng Zhang, Shuyang Sun, Jianbo Li, and Yu Qiao. Tip-adapter: Training-free clip-adapter for vision-language models. In *International Conference on Computer Vision (ICCV)*, pp. 10271–10280, 2021.

Renrui Zhang, Ziyu Li, Yi Zhang, Kensho Yang, Shangzhe Lin, Wenqi Chen, Siyuan Wang, Dongsheng Zhao, Jieming Wu, and Yu Qiao. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13110–13120, 2022a.

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.

Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 33:1348–1360, 2024.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890, 2017.

Zhedong Zheng, Liang Zheng, and Yi Yang. Scalable person re-identification with deep metric learning. In *International Conference on Machine Learning (ICML)*, pp. 2209–2218, 2015.

Zhun Zhong, Liang Zheng, Donglin Cao, Weijian Deng, Jun Li, Shengjin Liu, and Hui Wang. Re-ranking person re-identification with local query expansion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1510–1519, 2017.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 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.

Ziyu Zhu, Junnan Li, Jieming Wu, Ying Zhang, Yan Yan, Haonan Su, Jianwei Wu, and Yong Liu. Not all frames are equal: Weakly supervised video grounding with multi-modal prompts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15104–15114, 2023.

A Preliminaries

A.1 CLIP

The CLIP (Contrastive Language–Image Pre-training) model (Radford et al., 2021) has gained substantial attention in both natural language processing and computer vision due

to its ability to bridge the semantic gap between textual descriptions and visual content. It adopts a *dual-encoder* (dual-tower) architecture comprising a text encoder f^T and an image encoder f^I, which are jointly pre-trained on large-scale image–text pairs using a *contrastive learning* objective. The key idea is to align semantically related image–text pairs close together in a shared vision–language embedding space while pushing apart unrelated pairs.

Image Encoder. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the image encoder f^I (typically a Vision Transformer) first divides $\mathbf{I}$ into M fixed-size patches. Each patch is linearly projected into a d_v-dimensional embedding, producing a sequence of patch embeddings $\mathbf{E}_0 \in \mathbb{R}^{M\times d_v}$. A learnable *class token* $\mathbf{c}_0$ is prepended to this sequence, and positional encodings are added. The augmented sequence is then passed through L transformer layers $\{\mathcal{V}_l\}_{l=1}^{L}$:

$$\mathbf{V}_l = [\mathbf{c}_l, \mathbf{E}_l] = \mathcal{V}_l([\mathbf{c}_{l-1}, \mathbf{E}_{l-1}]), \quad l = 1, 2, \ldots, L.$$

After the final transformer block, the output of the class token $\mathbf{c}_L$ captures the global visual representation. A projection head P_v maps this representation into the shared vision–language space $\mathbf{v} = P_v(\mathbf{c}_L), \quad \mathbf{v} \in \mathbb{R}^d$.

Text Encoder. For each class c from a dataset with C categories, CLIP constructs a textual prompt using a predefined template such as $s_c =$ "a photo of a [c]". The text s_c is tokenized into N discrete tokens $\{t_j\}_{j=1}^{N}$ and embedded into $\mathbf{T}_0 \in \mathbb{R}^{N\times d_t}$. Beginning-of-text (BOT) and end-of-text (EOT) tokens are included to mark the sequence boundaries. The sequence is then processed through L transformer layers $\{\mathcal{T}_l\}_{l=1}^{L}$:

$$\mathbf{T}_l = [\mathbf{t}_l]_{j=1}^{N} = \mathcal{T}_l([\mathbf{t}_{l-1}]_{j=1}^{N}), \quad l = 1, 2, \ldots, L.$$

The final EOT token output $\mathbf{t}_L^{(N)}$ serves as the global text representation, which is projected into the shared embedding space using a projection head P_t where $\mathbf{w}_c = P_t(\mathbf{t}_L^{(N)}), \quad \mathbf{w}_c \in \mathbb{R}^d$.

During pre-training, CLIP optimizes a symmetric contrastive loss to maximize the cosine similarity between matched image–text pairs and minimize it for mismatched ones. For zero-shot classification, given an image $\mathbf{I}$ and a set of C textual class descriptions $\{\mathbf{w}_c\}_{c=1}^{C}$, CLIP computes the class probability as:

$$P(y = c \mid \mathbf{I}) = \frac{\exp(\mathrm{sim}(\mathbf{v}, \mathbf{w}_c)/\tau)}{\sum_{i=1}^{C} \exp(\mathrm{sim}(\mathbf{v}, \mathbf{w}_i)/\tau)}$$

where τ is a temperature parameter controlling the sharpness of the distribution. The predicted class corresponds to the text representation with the highest similarity score.

A.2 Centered Kernel Alignment (CKA)

Centered Kernel Alignment (CKA) is a similarity measure used to quantify the correspondence between representations learned by different models or different layers of the same model. It is particularly well-suited for analyzing high-dimensional feature spaces, where naive similarity measures may fail to capture meaningful structural relationships. CKA builds upon the *Hilbert-Schmidt Independence Criterion* (HSIC), a kernel-based measure of statistical dependence between two sets of variables.

Hilbert-Schmidt Independence Criterion (HSIC). Let $\mathbf{X} \in \mathbb{R}^{n\times d_x}$ and $\mathbf{Y} \in \mathbb{R}^{n\times d_y}$ denote two sets of representations over n samples. Assuming that $\mathbf{X}$ and $\mathbf{Y}$ are centered, the squared

Frobenius norm of their cross-covariance matrix can be written as:

$$\frac{1}{(n-1)^2}\operatorname{tr}(\mathbf{X}\mathbf{X}^\top\mathbf{Y}\mathbf{Y}^\top) = \|\operatorname{cov}(\mathbf{X}^\top, \mathbf{Y}^\top)\|_F^2.$$

HSIC generalizes this formulation by replacing inner products with kernel functions. Let $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ be kernels defined on $\mathbf{X}$ and $\mathbf{Y}$, respectively. Denote the corresponding kernel matrices by $\mathbf{K}, \mathbf{L} \in \mathbb{R}^{n\times n}$, where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $L_{ij} = l(\mathbf{y}_i, \mathbf{y}_j)$. The empirical HSIC estimator is given by:

$$\text{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2}\operatorname{tr}(\mathbf{KHLH}),$$

where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. HSIC measures statistical dependence between $\mathbf{X}$ and $\mathbf{Y}$: larger values indicate stronger dependence. When universal kernels are used, HSIC = 0 implies independence. However, HSIC is not invariant to isotropic scaling of the representations.

Centered Kernel Alignment (CKA). To address the scale sensitivity of HSIC, CKA normalizes it to obtain a scale-invariant similarity measure:

$$\text{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\text{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\text{HSIC}(\mathbf{K}, \mathbf{K})\ \text{HSIC}(\mathbf{L}, \mathbf{L})}}.$$

This normalization makes CKA invariant to isotropic scaling and orthogonal transformations, which is crucial when comparing neural representations that may differ by rotations or rescalings.

Linear CKA. A widely used special case is *linear CKA*, obtained when both kernels are linear: $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top\mathbf{x}'$ and $l(\mathbf{y}, \mathbf{y}') = \mathbf{y}^\top\mathbf{y}'$. In this case, $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{L} = \mathbf{Y}\mathbf{Y}^\top$, and CKA simplifies to:

$$\text{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\|\mathbf{Y}^\top\mathbf{X}\|_F^2}{\|\mathbf{X}^\top\mathbf{X}\|_F\,\|\mathbf{Y}^\top\mathbf{Y}\|_F}.$$

In practice, both linear and RBF kernels can be used within the CKA framework. Empirical evidence from the original work shows that RBF and linear kernels produce qualitatively similar results across a wide range of experiments; consequently, linear CKA is typically preferred due to its simplicity and computational efficiency.

Why CKA over Cosine Similarity and CCA?

Cosine Similarity. Cosine similarity operates on vectorized representations and measures only the alignment between individual feature vectors. When extended to matrices (e.g., by flattening), it ignores the internal structure of representations across samples. Moreover, cosine similarity is not invariant to general linear transformations and cannot capture relationships between subspaces. In contrast, CKA operates on Gram matrices, effectively comparing pairwise similarities between samples, which allows it to capture higher-order structural information in the representations.

Canonical Correlation Analysis (CCA). CCA seeks linear transformations of $\mathbf{X}$ and $\mathbf{Y}$ that maximize their correlation. While powerful, CCA suffers from several limitations in high-dimensional settings. In particular, when d_x or d_y exceeds n, CCA can become ill-posed and may yield degenerate solutions. Additionally, CCA is invariant to arbitrary invertible linear transformations, which can be undesirable: two representations with very different geometries may still achieve high CCA similarity.

CKA, on the other hand, restricts invariance to orthogonal transformations and isotropic scaling, preserving meaningful geometric structure. Theoretically, linear CKA can be interpreted as a normalized Frobenius inner product between covariance operators, which avoids the degeneracies of CCA while remaining sensitive to representational geometry.

Algorithm 1 CKA-based Layer Selection for Adapter Placement

Require: Encoder $\mathcal{E}_{img}$ and $\mathcal{E}_{txt}$ with L layers, datasets $\mathcal{D}$, selection size k
Ensure: Multimodal layers $\mathbf{L}_{mm}$, image layers $\mathbf{L}_{img}$, text layers $\mathbf{L}_{txt}$

function SELECTTOPKLAYERS($\mathcal{E}, \mathcal{D}, k$)
  Compute representations $\{H_1, \ldots, H_L\}$ using $\mathcal{E}$ on $\mathcal{D}$
  for $i = 1$ to L **do**
    $s_i \leftarrow 0$
    for $j = 1$ to L **do**
      if $i \neq j$ **then**
        $s_i \leftarrow s_i + (1 - \text{CKA}(H_i, H_j))$
      end if
    end for
    $Score_i \leftarrow \frac{s_i}{L-1}$
  end for
  return Top-k layers ranked by $Score_i$
end function
$\mathbf{L}_{\tilde{img}} \leftarrow$ SELECTTOPKLAYERS($\mathcal{E}_{img}, \mathcal{D}, k$)
$\mathbf{L}_{\tilde{txt}} \leftarrow$ SELECTTOPKLAYERS($\mathcal{E}_{txt}, \mathcal{D}, k$)
$\mathbf{L}_{mm} \leftarrow \mathbf{L}_{img} \cap \mathbf{L}_{txt}$
$\mathbf{L}_{img} \leftarrow \mathbf{L}_{\tilde{img}} \setminus \mathbf{L}_{mm}$
$\mathbf{L}_{txt} \leftarrow \mathbf{L}_{\tilde{txt}} \setminus \mathbf{L}_{mm}$
return $\mathbf{L}_{mm}, \mathbf{L}_{img}, \mathbf{L}_{txt}$

Method	Mod	#Trainable Params	Train Time *ms/image*	Train Time *min/all*	FPS *100 BS*	HM
MaPLe	V-L	3.555M	39.5	26.4	1757.6	78.55
PromptSRC	V, L	0.046M	40.0	106.8	**1764.2**	79.97
ProVP	V	0.147M	4.4	107.2	928.9	78.76
MetaPrompt	V, L	**0.031M**	30.7	32.8	659.8	79.09
TCP	L	0.332M	5.3	17.7	950.6	79.51
MMA	V-L	0.675M	2.2	1.8	688.5	79.87
CEKALA	V-L	0.253M	**1.4**	**1.2**	1064.6	80.06
CEKALA+ UMA	V-L	0.336M	2.0	1.5	798.8	**80.45**

Table 7: Computation cost comparison of different methods using MMRL's setup (Guo & Gu, 2025) on ImageNet. 'V-L' denotes vision-language interaction, 'V, L' indicates separate fine-tuning, and 'L' represents textual-only fine-tuning. The training time is given for each image and for the whole dataset (16 shots). FPS represents frames per second at a 100 batch size during inference.

B Dataset Description

To thoroughly assess our method, we follow prior studies and conduct experiments on 14 datasets that span a broad spectrum of visual recognition tasks, including generic object classification, fine-grained categorization, scene recognition, texture analysis, satellite image understanding, and action recognition. This collection consists of 11 primary datasets along with 3 robustness-oriented variants of ImageNet. For general object recognition, we employ datasets such as ImageNet, Caltech101, and SUN397, each paired with simple prompts like "a photo of a [CLASS]." Fine-grained benchmarks—OxfordPets, StanfordCars, Flowers102, Food101, and FGVCAircraft—focus on more specialized object categories and therefore use domain-tailored prompt structures (e.g., "a photo of a [CLASS], a type of flower"). DTD serves as the benchmark for texture classification using prompts such as "[CLASS] texture," while EuroSAT contains satellite imagery and uses templates like "a centered satellite photo of [CLASS]." For action recognition, we utilize UCF101, where prompts describe dynamic activities (e.g., "a photo of a person doing [CLASS]").

Dataset	Classes	Train	Val	Test	Description	Prompt
ImageNet	1000	1.28M	∼	50000	Recognition of generic objects	"a photo of a [CLASS]."
Caltech101	101	4128	1649	2465	Recognition of generic objects	"a photo of a [CLASS]."
OxfordPets	37	2944	736	3669	Fine-grained classification of pets	"a photo of a [CLASS], a type of pet."
StanfordCars	196	6509	1635	8041	Fine-grained classification of cars	"a photo of a [CLASS]."
Flowers102	102	4093	1633	2463	Fine-grained classification of flowers	"a photo of a [CLASS], a type of flower."
Food101	101	50500	20200	30300	Fine-grained classification of foods	"a photo of [CLASS], a type of food."
FGVCAircraft	100	3334	3333	3333	Fine-grained classification of aircraft	"a photo of a [CLASS], a type of aircraft."
SUN397	397	15880	3970	19850	Scene classification	"a photo of a [CLASS]."
DTD	47	2820	1128	1692	Texture classification	"[CLASS] texture."
EuroSAT	10	13500	5400	8100	Land use & cover classification with satellite images	"a centered satellite photo of [CLASS]."
UCF101	101	7639	1898	3783	Action recognition	"a photo of a person doing [CLASS]."
ImageNetV2	1000	∼	∼	10,000	New test data for ImageNet	"a photo of a [CLASS]."
ImageNet-Sketch	1,000	∼	∼	50,889	Sketch-style images of ImageNet classes	"a photo of a [CLASS]."
ImageNet-A	200	∼	∼	7,500	Natural adversarial examples of 200 ImageNet classes	"a photo of a [CLASS]."
ImageNet-R	200	∼	∼	30,000	Renditions of 200 ImageNet classes	"a photo of a [CLASS]."

Table 8: Summary of all 14 datasets used in this work, including 11 distinct datasets and 3 variants of ImageNet.

Beyond these core datasets, we evaluate robustness and generalization using ImageNetV2, ImageNet-Sketch, and ImageNet-A/R. ImageNetV2 provides a re-collected test set aligned with the original ImageNet taxonomy. ImageNet-Sketch includes sketch-style renditions of ImageNet categories, whereas ImageNet-A and ImageNet-R contain natural adversarial examples and artistic reinterpretations, respectively. All prompt templates follow standard formulations established in earlier work, ensuring consistency across tasks. Detailed dataset statistics, including the number of classes and data splits, are provided in 8.

C Implementation Details

Experimental Setup. Following standard in adapter tuning research (Zhou et al., 2022a;b; Yao et al., 2024; Khattak et al., 2023a; Yang et al., 2024; Guo & Gu, 2025), we adopt CLIP with ViT-B/16 architecture (Radford et al., 2021) as our visual backbone across all experiments. Text prompts are manually designed following standard practices (Radford et al., 2021; Zhou et al., 2022b;a), with templates provided in 8. We use the same setting for the ablation with SigLIP (Zhai et al., 2023).

Training Configuration. We employ the AdamW optimizer with a learning rate of 0.001 and mixed-precision training for computational efficiency. Dataset-specific batch sizes are used: 32 for ImageNet and 4 for the remaining datasets. Training epochs vary by task: 5 epochs for ImageNet base-to-novel evaluation, 1 epoch for cross-dataset and domain generalization on ImageNet, 5 epochs for few-shot learning on ImageNet, and 50 epochs for few-shot learning on the remaining datasets. All results represent averages over three independent runs. All the experiments were run on a single NVIDIA RTX 4090 GPU.

Hyperparameter Settings. Based on the ablation study in 6, we configure CEKALAwith hidden dimension 64, $\alpha = 0.05$, $k = 6$. For fairness in comparison, the hyperparameters remain fixed across all datasets.

D Computational Cost

Table 7 presents a comprehensive comparison of computational efficiency and performance across several state-of-the-art prompt-learning and adapter-based methods under the standardized MMRL training setup on ImageNet. The table reports the number of trainable parameters, per-image and total training time (for the 16-shot setting), inference throughput measured in FPS with a batch size of 100, and final classification performance measured by harmonic mean (HM). Methods differ in their modality configuration, including vision–language interaction (V-L), separate vision and language tuning (V, L), and language-only tuning (L). As shown, CEKALAachieves the best overall trade-off, offering the fastest training speed and highest HM score while maintaining a relatively small parameter footprint, outperforming all baselines in both efficiency and accuracy.

Table 9: Comparison of CEKALA with previous state-of-the-art methods on few-shot learning across 11 datasets.

Dataset	Method	1 shot	2 shots	4 shots	8 shots	16 shots
Average	Linear probe CLIP	45.83	57.98	68.01	74.47	78.79
	CoOp	67.56	70.65	74.02	76.98	79.89
	CoCoOp	66.79	67.65	71.21	72.96	74.90
	MaPLe	69.27	72.58	75.37	78.89	81.79
	PromptSRC	72.32	75.29	78.35	80.69	82.87
	MMA	69.28	72.08	76.38	79.57	82.76
	CEKALA(Ours)	73.77	74.60	78.80	79.81	83.14
ImageNet	Linear probe CLIP	32.13	44.88	54.85	62.23	67.31
	CoOp	66.33	67.07	68.73	70.63	71.87
	CoCoOp	69.43	69.78	70.39	70.63	70.83
	MaPLe	62.67	65.10	67.70	70.30	72.33
	PromptSRC	68.13	69.77	71.07	72.33	73.17
	MMA	69.17	70.37	71.00	71.77	73.13
	CEKALA(Ours)	69.57	70.72	71.21	72.77	73.54
Caltech101	Linear probe CLIP	79.88	89.01	92.05	93.41	95.43
	CoOp	92.60	93.07	94.40	94.37	95.57
	CoCoOp	93.83	94.82	94.98	95.04	95.16
	MaPLe	92.57	93.97	94.43	95.20	96.00
	PromptSRC	93.67	94.53	95.27	95.67	96.07
	MMA	92.90	94.00	94.33	95.37	96.33
	CEKALA(Ours)	93.90	94.78	96.39	95.73	97.04
OxfordPets	Linear probe CLIP	44.06	58.37	71.17	78.36	85.34
	CoOp	90.37	89.80	92.57	91.27	91.87
	CoCoOp	91.27	92.64	92.81	93.45	93.34
	MaPLe	89.10	90.87	91.90	92.57	92.83
	PromptSRC	92.00	92.50	93.43	93.50	93.67
	MMA	91.23	91.97	92.23	92.77	93.23
	CEKALA(Ours)	92.10	92.37	92.81	92.83	93.48
StanfordCars	Linear probe CLIP	35.66	50.28	63.38	73.67	80.44
	CoOp	67.43	70.50	74.47	79.30	83.07
	CoCoOp	67.22	68.37	69.39	70.44	71.57
	MaPLe	66.60	71.60	75.30	79.47	83.57
	PromptSRC	69.40	73.40	77.13	80.97	83.83
	MMA	67.87	71.77	76.50	81.40	85.70
	CEKALA(Ours)	68.59	72.87	79.50	83.10	87.06
Flowers102	Linear probe CLIP	69.74	85.07	92.02	96.10	97.37
	CoOp	77.53	87.33	92.17	94.97	97.07
	CoCoOp	72.08	75.79	78.40	84.30	87.84
	MaPLe	83.30	88.93	92.67	95.80	97.00
	PromptSRC	85.93	91.17	93.87	96.27	97.60
	MMA	83.60	90.30	93.00	95.97	97.97
	CEKALA(Ours)	85.60	92.30	94.87	96.79	98.55

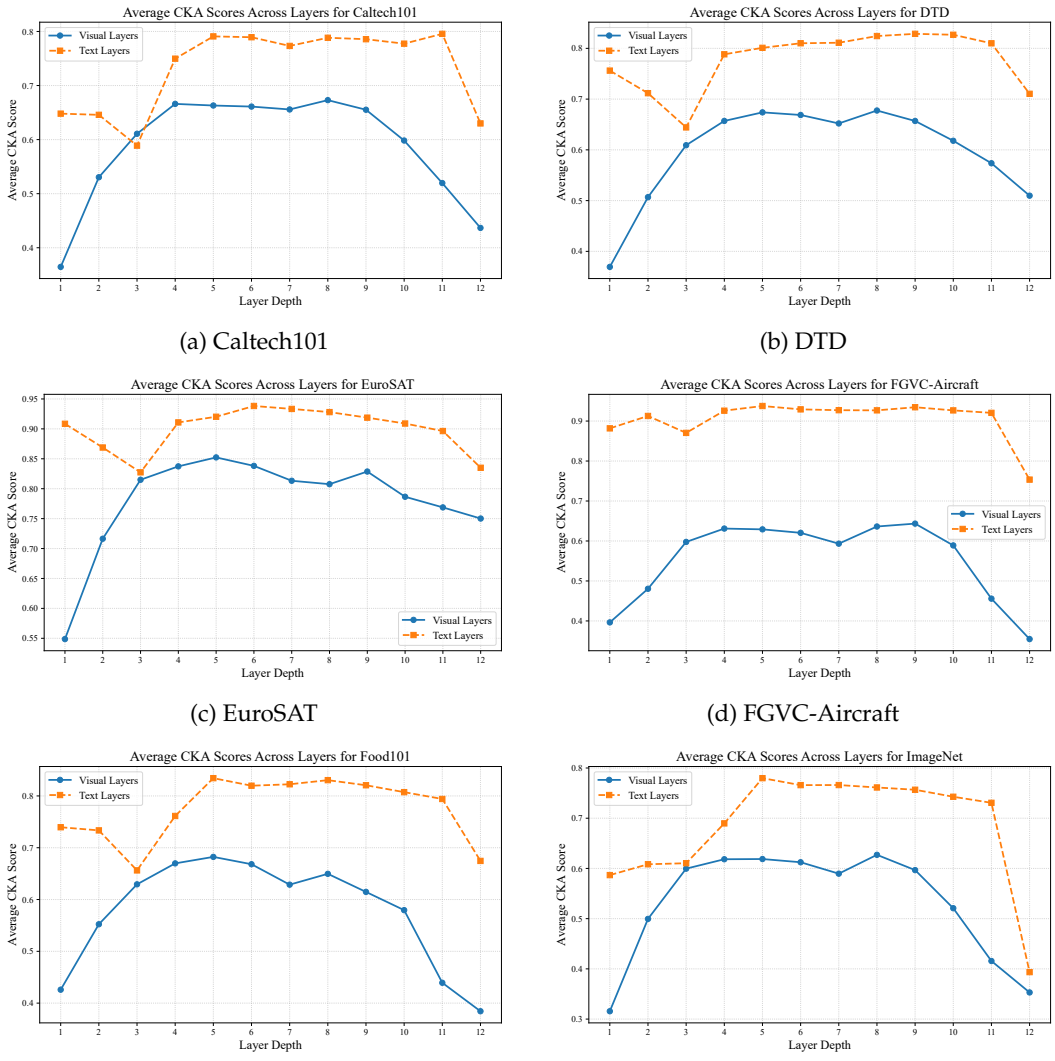
E Few-Shot Learning

Tables 9 and 10 present an extensive comparison of CEKALA with previous state-of-the-art approaches on few-shot classification across 11 datasets. CEKALA consistently delivers the best average performance across all shot configurations. The table is followed by MMRL(Guo & Gu, 2025) few-shot learning experimentation.

Table 10: Comparison of CEKALA with previous state-of-the-art methods on few-shot learning across various datasets (Table 11 continuation).

Dataset	Method	1 shot	2 shots	4 shots	8 shots	16 shots
Food101	Linear probe CLIP	43.96	61.51	73.19	79.79	82.90
	CoOp	84.33	84.40	84.47	82.67	84.20
	CoCoOp	85.65	86.22	86.88	86.97	87.25
	MaPLe	80.50	81.47	81.77	83.60	85.33
	PromptSRC	84.87	85.70	86.17	86.90	87.50
	MMA	83.03	82.50	82.13	83.00	84.57
	CEKALA(Ours)	85.24	85.48	86.33	87.17	87.92
FGVCAircraft	Linear probe CLIP	19.61	26.41	32.33	39.35	45.36
	CoOp	21.37	26.20	30.83	39.00	43.40
	CoCoOp	12.68	15.06	24.79	26.61	31.21
	MaPLe	26.73	30.90	34.87	42.00	48.40
	PromptSRC	27.67	31.70	37.47	43.27	50.83
	MMA	28.73	31.90	37.57	44.83	52.70
	CEKALA(Ours)	28.51	31.62	37.21	45.92	52.84
SUN397	Linear probe CLIP	41.58	53.70	63.00	69.08	73.28
	CoOp	66.77	66.53	69.97	71.53	74.67
	CoCoOp	68.33	69.03	70.21	70.84	72.15
	MaPLe	64.77	67.10	70.67	73.23	75.53
	PromptSRC	69.67	71.60	74.00	75.73	77.23
	MMA	64.00	67.17	69.97	72.30	74.63
	CEKALA(Ours)	65.25	70.82	74.23	76.37	77.85
DTD	Linear probe CLIP	34.59	40.76	55.71	63.46	69.96
	CoOp	50.23	53.60	58.70	64.77	69.87
	CoCoOp	48.54	52.17	55.04	58.89	63.04
	MaPLe	52.13	55.50	61.00	66.50	71.33
	PromptSRC	56.23	59.97	65.53	69.87	72.73
	MMA	52.27	56.90	63.93	67.97	73.47
	CEKALA(Ours)	56.82	61.78	68.09	72.48	76.01
EuroSAT	Linear probe CLIP	49.23	61.98	77.09	84.43	87.21
	CoOp	54.93	65.17	70.80	78.07	84.93
	CoCoOp	55.33	46.74	65.56	68.21	73.32
	MaPLe	71.80	78.30	84.50	87.73	92.33
	PromptSRC	73.13	79.37	86.30	88.80	92.43
	MMA	55.07	59.80	79.40	86.47	92.37
	CEKALA(Ours)	74.91	80.74	86.50	88.94	92.97
UCF101	Linear probe CLIP	53.66	65.78	73.28	79.34	82.11
	CoOp	71.23	73.43	77.10	80.20	82.23
	CoCoOp	70.30	73.51	74.82	77.14	78.14
	MaPLe	71.83	74.60	78.47	81.37	85.03
	PromptSRC	74.80	78.50	81.57	84.30	86.47
	MMA	74.17	76.17	80.10	83.43	86.30
	CEKALA(Ours)	75.29	77.83	81.14	84.22	87.28

F Layer-wise CKA Score for the Dataset

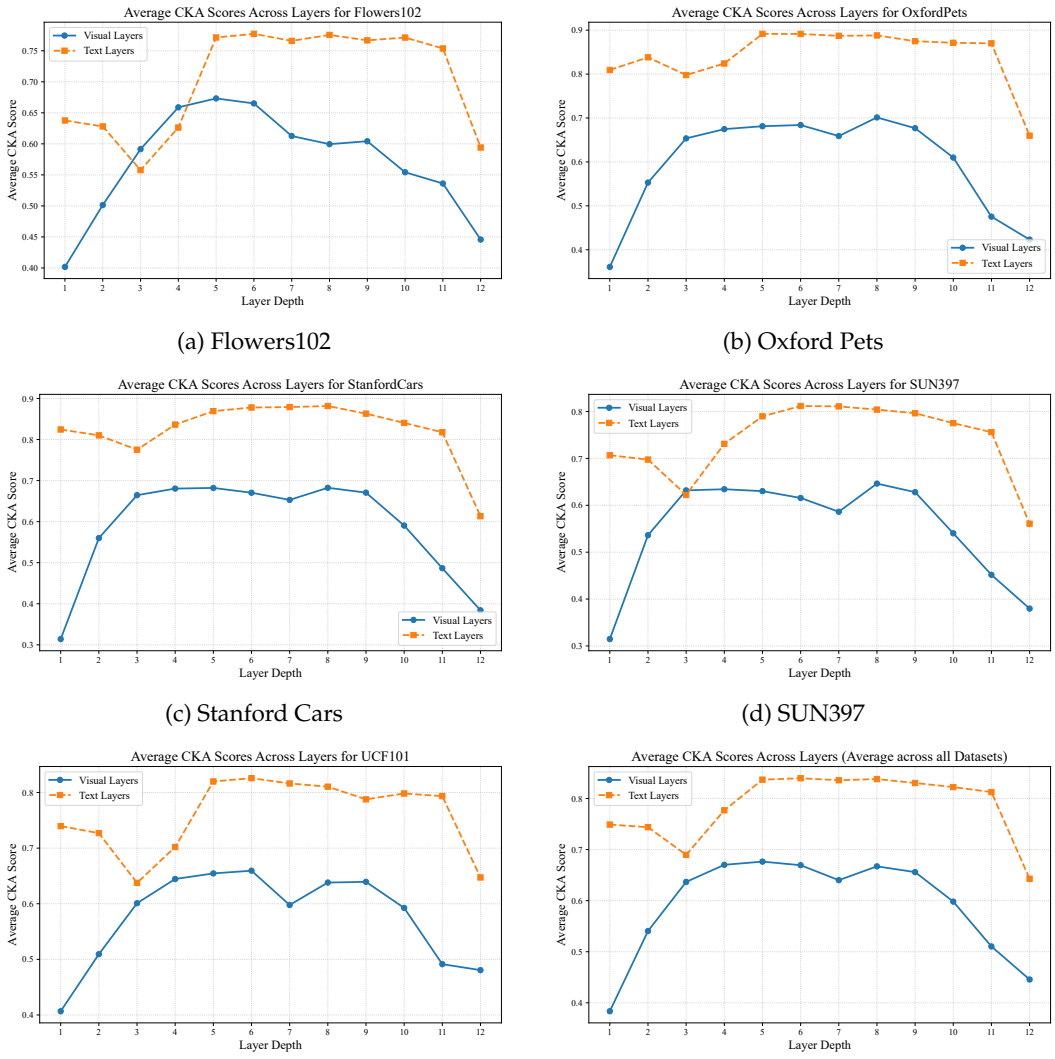


(a) Caltech101 (b) DTD

(c) EuroSAT (d) FGVC-Aircraft

(e) Food101 (f) ImageNet

Figure 2: CKA score plots for Caltech101, DTD, EuroSAT, FGVC-Aircraft, Food101, and ImageNet.



(a) Flowers102 (b) Oxford Pets

(c) Stanford Cars (d) SUN397

(e) UCF101 (f) Average

Figure 3: CKA score plots for Flowers102, Oxford Pets, Stanford Cars, SUN397, UCF101, and the average across datasets.