TLAC: Two-stage LMM Augmented CLIP for Zero-Shot Classification

Ans Munir[1,4], Faisal Z. Qureshi[2], Muhammad Haris Khan[3], Mohsen Ali[1]

[1]Information Technology University, Lahore, Pakistan
[2]University of Ontario Institute of Technology, Oshawa, Canada
[3]Mohamed Bin Zayed University of AI, Abu Dhabi, UAE [4]MZASS AI

{msds20033, mohsen.ali}@itu.edu.pk, faisal.qureshi@ontariotechu.ca
muhammad.haris@mbzuai.ac.ae

Abstract

Contrastive Language-Image Pretraining (CLIP) has shown impressive zero-shot performance on image classification. However, state-of-the-art methods often rely on fine-tuning techniques like prompt learning and adapter-based tuning to optimize CLIP's performance. The necessity for fine-tuning significantly limits CLIP's adaptability to novel datasets and domains. This requirement mandates substantial time and computational resources for each new dataset. To overcome this limitation, we introduce simple yet effective training-free approaches, Single-stage LMM Augmented CLIP (SLAC) and Two-stage LMM Augmented CLIP (TLAC), that leverages powerful Large Multimodal Models (LMMs), such as Gemini, for image classification. The proposed methods leverages the capabilities of pre-trained LMMs, allowing for seamless adaptation to diverse datasets and domains without the need for additional training. Our approaches involve prompting the LMM to identify objects within an image. Subsequently, the CLIP text encoder determines the image class by identifying the dataset class with the highest semantic similarity to the LLM predicted object. Our models achieved superior accuracy on 9 of 11 base-to-novel datasets, including ImageNet, SUN397, and Caltech101, while maintaining a strictly training-free paradigm. Our TLAC model achieved an overall accuracy of 83.44%, surpassing the previous state-of-the-art few-shot methods by a margin of 6.75%. Compared to other training-free approaches, our TLAC method achieved 83.6% average accuracy across 13 datasets, a 9.7% improvement over the previous methods. Our Code is available at https://github.com/ans92/TLAC

1. Introduction

Vision-Language Models (VLMs) like CLIP (Contrastive Language-Image Pretraining) [29] have demonstrated impressive performance on a variety of downstream tasks, including few-shot learning [15, 43] and zero-shot learning [28]. Trained on a massive dataset of 400 million image-text pairs using a contrastive learning approach, CLIP exhibits strong generalization capabilities. However, CLIP still necessitates a certain level of task-specific knowledge (such as fine-tuning on downstream tasks) to attain optimal results. Fine-tuning the entire model on limited downstream data is often impractical due to the risk of overfitting. To address this challenge, two primary techniques have emerged to adapt CLIP to downstream tasks: Prompt Engineering [48, 49] and Adapter-based Fine-tuning [5, 11].

CLIP rely on textual prompts to guide its image understanding. When presented with an image, CLIP aims to identify the most relevant prompt from a vast pool of possibilities. For instance, given an image of a Red Car, CLIP might match it with the prompt "A photo of a red car." Previous work [49] has shown that changes in the structure of the CLIP prompt affect its accuracy. To address this challenge and to adapt CLIP to specific downstream tasks without extensive fine-tuning, prompt learning has emerged as a promising technique. Through training, the model learns to exploit the fixed elements within prompts such as "*A photo of a/an* [object]," allowing it to effectively apply this knowledge to downstream tasks. Another approach leverages lightweight, trainable adapter networks, leaving the larger CLIP model frozen. These techniques, in conjunction with CLIP, offer strong generalization capabilities and have achieved impressive results. However, a major limitation of these methods is the need for training on each new dataset. This hinders their potential as universal models capable of handling diverse domains and classes.

To mitigate these challenges, recent training-free approaches, such as CuPL [28] and MPVR [25], integrate Large Language Models (LLMs) with CLIP. Specifically, these methods first employ LLMs to generate textual descriptions for each class. Then, CLIP is used to determine

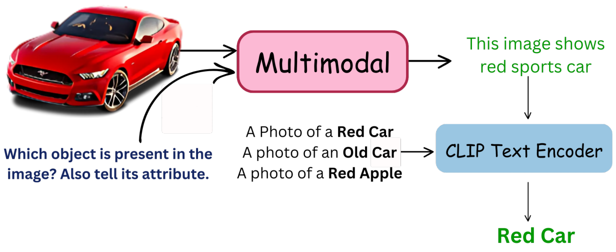

Figure 1. Diagram of SLAC model. Blue text is the LMM prompt. Red Car, Old Car and Red Apple are dataset class labels.

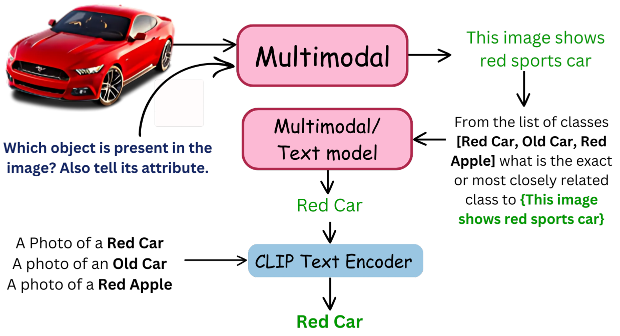
Multimodal
This image shows red sports car
Which object is present in the image? Also tell its attribute.
Multimodal/ Text model
From the list of classes [Red Car, Old Car, Red Apple] what is the exact or most closely related class to {This image shows red sports car}
Red Car
A Photo of a Red Car
A photo of an Old Car
A photo of a Red Apple
CLIP Text Encoder
Red Car

Figure 2. Diagram of our TLAC model. Blue text is the LMM prompt. Red Car, Old Car and Red Apple are dataset class labels.

the class description that best aligns with the visual representation of the image. While eliminating the need for training, these methods still necessitate the generation of class descriptions tailored to each new dataset.

To address this limitation, we propose a straightforward technique that leverages the capabilities of Large Multimodal Models (LMMs) [1] such as Gemini [35] eliminating the need for explicit class description generation. Our approach is training-free, allowing us to fully harness the power of LMMs, which benefit from extensive pre-training on diverse data. A key advantage of LMMs is their universality, enabling them to handle a wide range of datasets and domains. To the best of our knowledge, this is the first work to explore LMM for image classification and combined LMM, Gemini, with VLM, CLIP. The main contribution of this work is as follows:

- We propose a simple yet effective training-free approach to leverage Large Multimodal Models (LMMs) for image classification task.
- Specifically, this work introduces two models, Single-stage LMM Augmented CLIP (SLAC) and Two-stage LMM Augmented CLIP (TLAC), which combine the strengths of LMM and VLM to leverage their combined capabilities.
- Our experiments demonstrate that our approach achieves superior accuracy on a majority of evaluated datasets, including the large-scale ImageNet, all while remaining entirely training-free and requiring no training samples.

2. Method

Unlike previous work [15, 43, 48, 49] that relied on fine-tuning CLIP, we propose a training-free approach leveraging Large Multimodal Models (LMMs) for image classification. Our method offers a significant advantage: it's training-free, yet surpasses the performance of state-of-the-art fine-tuned models. Furthermore, our method requires no training samples, unlike some other training-free approaches [47]. We introduce two distinct models: SLAC and TLAC.

2.1. Single-stage LMM Augmented CLIP (SLAC)

In our SLAC model, we first provide an image and a prompt to the LMM, Gemini. For instance, the prompt might be *"Which object is present in the image? Also tell its attribute."* Gemini then generates a textual description, z, such as *"The image shows red sports car"* as illustrated in Figure 1. Subsequently, both z and all dataset class labels, y, are encoded into textual features using the CLIP text encoder. Finally, the similarity between z and each y is computed to determine the image's predicted class.

$$c^* = \underset{y \in C}{\arg\max} \Big(g(z).g(y) \Big) \tag{1}$$

where z is the answer predicted by Gemini LMM and y represents the class from the dataset classes C.

2.2. Two-stage LMM Augmented CLIP (TLAC)

Our SLAC model yielded promising results. However, we encountered instances where the model correctly identified the class present in the image, but the predicted label did not match the ground truth. For example, in flower datasets, some images were labeled with common names, while others used scientific names. Although our SLAC model, incorporating CLIP, mitigated this issue to some extent, it still struggled with scientific names. To address this limitation, we propose a TLAC model, as illustrated in Figure 2.

TLAC model employs a two-step approach. First, the LMM, Gemini, gives the answer to our question with respect to the image, similar to the SLAC model. However, instead of predicting the object present in the image, we ask the LMM to provide the object most relevant to the dataset

[1] In literature some works called them Multimodal Large Language Models (MLLMs)

	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	UCF101	ImageNetv2	ImageNet-R	ImageNet-S	Average
CLIP-S	73.5	94.3	93.1	76.9	76.2	90.3	30.0	67.6	52.5	73.8	60.9	74.0	46.2	69.9
CLIP-DS	75.5	93.7	93.5	78.1	79.5	90.9	31.8	69.0	54.8	76.2	61.9	77.7	48.8	71.6
CuPL	76.7	93.5	93.8	77.6	79.7	93.4	36.1	73.3	61.7	78.4	63.4	-	-	75.2
D-CLIP	75.1	**97.0**	93.0	75.1	79.5	91.1	31.8	69.6	56.1	76.2	62.2	76.5	48.9	71.7
Waffle	75.1	96.2	93.2	76.5	78.3	91.5	32.5	69.4	55.3	76.0	62.3	77.0	49.1	71.7
MPVR (Mix)	75.9	95.4	93.1	70.6	83.8	91.4	37.6	72.5	61.6	75.8	62.2	78.4	49.7	72.9
MPVR (GPT)	**76.8**	96.1	93.7	78.3	83.6	91.5	34.4	73.0	62.9	78.1	63.4	78.2	50.6	73.9
Ours (SLAC)	73.8	96.6	96.5	88.7	77.7	92.9	65.6	73.5	58.5	85.2	67.9	89.9	66.1	79.4
Ours (TLAC)	74.1	**97.0**	**97.1**	**90.2**	**85.7**	**94.4**	**79.4**	**79.0**	**72.6**	**89.5**	**69.2**	**90.8**	**68.2**	**83.6**

Table 1. Table compares the results of our models with those of previous training-free methods. Results of previous state-of-the-art models have been taken from [25].The best result is displayed in bold, while the second-highest result is shown in blue. Higher scores represent superior performance.

class, based on its initial prediction. If the initial prediction was accurate, the second step reinforces it. However, if the initial prediction was partially correct but used a different name (for example, *Gaillardia* instead of *Blanket Flower* as shown in Table 3), the second step corrects and improves the prediction, leading to higher accuracy. Figure 2 provides a visual overview of this TLAC model.

3. Experiments

3.1. Generalization from Base-to-Novel classes

Previous methods [15, 43, 48, 49] have been trained on base classes for 16 instances per class in a few-shot manner also known as 16-shot learning and then generalized the knowledge on novel classes. In contrast, our approach is training-free, making it applicable to new datasets and domains. Similar to prior work [15, 43, 48, 49], we evaluated our approach on 11 diverse image classification datasets such as ImageNet [8] and Caltech101 [10], 2 general object classification datasets; OxfordPets [27], StanfordCars [16], Flowers102 [26], Food101 [3], and FGVCAircraft [23], 5 fine-grained datasets; SUN397 [42], a scene understanding dataset; DTD, a satellite-image recognition dataset [7]; UCF101 [33], an action classification dataset.

3.2. Domain Generalization

In domain generalization, models are evaluated on out-of-distribution datasets to assess their robustness across different domains. [48] proposed testing ImageNet-trained models on ImageNet variants such as ImageNetV2 [31], ImageNet-Sketch [40], and ImageNet-R [13]. Results 4 demonstrate that our approach achieves superior results on these diverse domains, even without explicit training.

3.3. Implementation Details

For the task of image classification we employed two versions of Gemini: Gemini Flash 1.5-002 and Gemini Pro 1.5-002 as our Large Multimodal Model (LMM). In SLAC model, we mostly used Gemini Pro, while in second step of TLAC model, we used Gemini Flash. Consistent with prior work [15, 43], we utilized the CLIP-B/16 model as a VLM text encoder.

3.4. Main Results

Training-Free Methods

We compared the performance of our model against several training-free methods, including CLIP [29] (using both a simple prompt, CLIP-S, and dataset-specific prompts, CLIP-DS), CuPL [28], D-CLIP [24], Waffle [32], and MPVR [25]. Table 1 presents the results, showing that our model achieved superior accuracy on all datasets except ImageNet. Our SLAC and TLAC models achieved overall average accuracies of 79.4% and 83.6%, respectively, surpassing the previous highest accuracy of 73.9% reported by MPVR [25] with the margin of 9.7%.

On ImageNet, TLAC achieved 74.1%, slightly below the state-of-the-art results of MPVR (76.8%) and CuPL (76.7%). However, on Caltech101, TLAC reached 97.0%, matching the performance of D-CLIP. Notably, TLAC surpassed the previous best on OxfordPets, achieving 97.1%, a 3.3% improvement over CuPL's 93.8%. Significant accuracy gains were observed on StanfordCars (11.1%), FGVCAircrafts (41.8%), SUN397 (5.7%), and DTD (9.7%). Furthermore, TLAC demonstrated superior domain generalization, achieving improvements of 5.8% on ImageNetv2, 12.4% on ImageNet-R, and 17.6% on ImageNet-S.

	Overall Avg		ImageNet		Caltech101		OxfordPets		StanfordCars		Flowers102	
Method	Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel
CLIP	69.34	74.22	72.43	68.14	96.84	94.00	91.17	97.26	63.37	74.89	72.08	77.80
CoOp	82.69	63.22	76.47	67.88	98.00	89.81	93.67	95.29	78.12	60.40	97.60	59.67
Co-CoOp	80.47	71.69	75.98	70.43	97.96	93.81	95.20	97.69	70.49	73.59	94.87	71.75
ProDA	81.56	72.30	75.40	70.23	98.27	93.23	95.43	97.83	74.70	71.20	97.70	68.68
KgCoOp	80.73	73.60	75.83	69.96	97.72	94.39	94.65	97.76	71.76	75.04	95.00	74.73
MaPLe	82.28	75.14	76.66	70.54	97.74	94.36	95.43	97.76	72.94	74.00	95.92	72.46
LASP	82.70	74.90	76.20	70.95	98.10	94.24	95.90	97.93	75.17	71.60	97.00	74.00
RPO	81.13	75.00	76.60	71.57	97.97	94.37	94.63	97.50	73.87	75.53	94.13	76.67
MMA	83.20	76.80	77.31	71.00	98.40	94.00	95.40	**98.07**	78.50	73.10	97.77	75.93
Ours (SLAC)	74.43	78.69	79.90	73.78	94.19	96.62	93.62	96.48	74.79	88.73	77.21	77.73
Ours (TLAC)	76.81	**83.44**	80.10	**74.06**	91.67	**96.96**	94.26	97.09	74.99	**90.17**	79.01	**85.74**
		+6.75		+2.49		+2.57		-0.98		+14.64		+9.07

	Food101		FGVCAircraft		SUN397		DTD		EuroSAT		UCF101	
Method	Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel
CLIP	90.10	91.22	27.19	36.29	69.36	75.35	53.24	59.90	56.48	64.05	70.53	77.50
CoOp	88.33	82.26	40.44	22.30	80.60	65.89	79.44	41.18	92.19	54.74	84.69	56.05
Co-CoOp	90.70	91.29	33.41	23.71	79.74	76.86	77.01	56.00	87.49	60.04	82.33	73.45
ProDA	90.30	88.57	36.90	34.13	78.67	76.93	80.67	56.48	83.90	66.00	85.23	71.97
KgCoOp	90.50	91.70	36.21	33.55	80.29	76.53	77.55	54.99	85.64	64.34	82.89	76.67
MaPLe	90.71	92.05	37.44	35.61	80.82	78.70	80.36	59.18	94.07	73.23	83.00	78.66
LASP	91.20	91.70	34.53	30.57	80.70	78.60	81.40	58.60	94.60	77.78	84.77	78.03
RPO	90.33	90.83	37.33	34.20	80.60	77.80	76.70	62.13	86.63	68.97	83.67	75.43
MMA	90.13	91.30	40.57	36.33	82.27	78.57	83.20	65.63	85.46	**82.34**	86.23	80.03
Ours (SLAC)	92.51	92.87	56.42	65.60	69.89	73.45	52.08	58.45	48.78	56.72	79.37	85.18
Ours (TLAC)	88.46	**94.37**	62.79	**79.36**	73.32	**79.00**	68.06	**72.95**	49.10	58.67	83.13	**89.51**
		+2.32		+43.03		+0.30		+7.32		-23.67		+9.48

Table 2. Table shows the results of previous state-of-the-art few-shot methods on 11 base-to-novel datasets. Base accuracy is on training/seen classes while Novel accuracy is on new/unseen classes. Results of both of our models, SLAC and TLAC are also shown. Previous state-of-the-art results have been taken from [43]. Although not directly comparable to prior base accuracy due to our no-training approach, we still mention base accuracy to facilitate comparison between prior results that require training and our training-free method. Results show that our models demonstrate superior accuracy on novel classes as compared to fine-tuned models. The best result is displayed in bold, while the second-highest result is shown in blue. Higher results are better.

Base-to-Novel Generalization

In this experimental setup, we compared our approach with several state-of-the-art few-shot methods: CLIP [29], CoOp [49], Co-CoOp [48], ProDA [22], KgCoOp [44], MaPLe [15], LASP [4], RPO [17], and MMA [43]. These methods typically train on base classes in a 16-shot learning manner. Table 2 presents the results for base class accu-

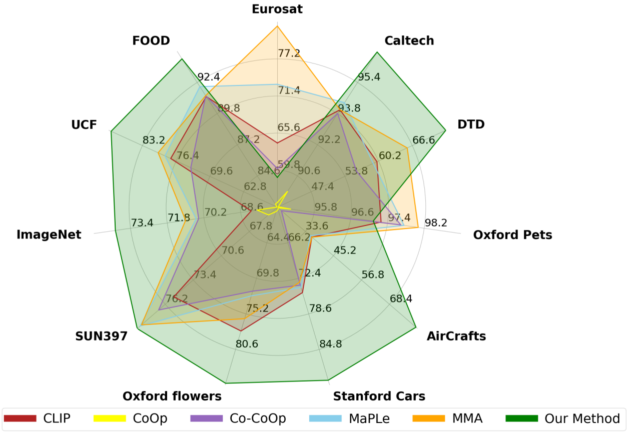

Figure 3. The figure illustrates the superior performance of our model compared to previous state-of-the-art few-shot methods on novel classes. Our approach has demonstrated significant performance gains across multiple challenging datasets, including Stanford Cars, Aircrafts, Oxford Flowers, ImageNet, and Caltech.

racy (**Base**) and novel class accuracy (**Novel**). Our approach, which does not involve training or fine-tuning, prevents direct comparison with the **Base** results of prior methods. Nevertheless, results 2 demonstrate that our model is competitive with fine-tuned models in terms of Base accuracy while achieving higher Novel accuracy. On three datasets-ImageNet, Food101, and FGVCAircraft-our model achieved higher Base accuracy than previous fine-tuned models, while remaining comparable on other datasets. This highlights the competitiveness of our model compared to other fine-tuned models, even on seen classes.

Our TLAC model achieved the highest overall accuracy of 83.44%, surpassing the previous state-of-the-art, MMA [43], by 6.75%. On general object recognition datasets, our approach outperformed state-of-the-art methods by 2.49% on ImageNet and 2.57% on Caltech101 in novel accuracy. For fine-grained image recognition, we achieved significant improvements of 14.64%, 9.07%, 2.32% and 43.03% on StanfordCars, Flowers102, Food101 and FGVCAircraft, respectively. However, we underperformed on OxfordPets by 0.98%. Additionally, we observed better performance on DTD and UCF101, with improvements of 7.32% and 1.04%, respectively. We also achieved better accuracy on scene understanding dataset, SUN397. On EuroSAT, our accuracy fell short of previous models. Figure 3 presents a comparison of the overall performance of our model with that of previous state-of-the-art models.

Consider the images in Table 3. Dataset labels are shown in green, and our model's predicted labels are in red. For the first two flower images, our SLAC model predicted their scientific names, "Gaillardia" and "Ruellia," instead of the common names "Blanket Flower" and "Mexican Petunia."

Blanket Flower	Mexican Petunia	Infant Bed	Infant Bed
Gaillardia	Ruellia	Cradle	Crib

Table 3. The tables shows the images with ground truth labels in green and predicted labels in red. Even though the predicted labels are sementically similar, SLAC model do not able to match them. In these situations TLAC model is effective.

Methods	ImageNetV2	ImageNet-S	ImageNet-R
CLIP	60.83	46.15	73.96
CoOp	64.20	47.99	75.21
Co-CoOp	64.07	48.75	76.18
Maple	64.07	49.15	76.98
MMA	64.33	49.13	77.32
Ours (SLAC)	67.93	66.11	89.91
Ours (TLAC)	**69.21**	**68.20**	**90.82**

Table 4. The table presents the results of our method compared to state-of-the-art few-shot methods. Higher scores better. The best result is displayed in bold, while the second-highest in blue.

Similarly, for the third and fourth images, the ground truth label was "Infant Bed", but our model predicted "Cradle" and "Crib". While these predictions are technically correct, our SLAC model struggled to identify the semantic equivalence between these terms. However, our TLAC model successfully corrected these mismatches, demonstrating the advantage of iterative refinement in certain scenarios.

Domain Generalization

Table 4 presents the performance of our approach compared to previous methods [15, 29, 49] on domain generalization datasets. While previous methods were trained on ImageNet, our approach was training-free. Our model achieved a 3.6% improvement over the state-of-the-art MMA model on ImageNetV2. Similarly, on ImageNet-S, and ImageNet-R, our approach yielded significant gains of 16.96%, and 12.59%, respectively.

4. Ablation Study

4.1. Effect of Prompt

Prior work [12, 15] has demonstrated the influence of prompt design on LLM/LMM accuracy, a phenomenon we also observed in our model. This effect is particularly pronounced for domain-specific datasets. The reason is LLMs/LMMs are highly sensitive to prompt wording as they rely on pattern recognition and context. A well-crafted

Prompt	Acc.
which object is present in the image.	94.43
which specific object is present in the image.	94.98
what is the specific type of object present in the image.	95.63

Table 5. Table showing the effect of different prompts on image classification accuracy of Caltech101 dataset. Acc. means Accuracy. Higher accuracy is better.

Prompt	Acc.
which object is present in the image.	84.55
which car is present in the image.	88.83
which specific car is present in the image.	90.20

Table 6. Table showing the effect of different prompts on image classification accuracy of StanfordCars dataset. Acc. means Accuracy. Higher accuracy is better.

prompt serves as a guide, offering the necessary context and examples to effectively "program" the model. This leads to more accurate outputs, especially when dealing with complex, domain-specific datasets that require nuanced understanding. To investigate this, we evaluated our model on a general-purpose dataset (Caltech101) and a domain-specific dataset (StanfordCars). Table 5 shows the impact of varying prompts on Caltech101. Asking the LMM "which object is present in the image" yielded lower accuracy compared to the more specific prompt "which specific object is present in the image." Further improvement was observed when prompting for a description: "what is the specific type of object present in the image," achieving an accuracy of 95.63%, compared to 94.43% and 94.98% with the previous prompts. While the accuracy differences are not significant, optimized prompting consistently improves performance.

On the domain-specific StanfordCars dataset, prompt engineering significantly impacts accuracy. As shown in Table 6, the generic prompt "which object is present in the image" yielded an accuracy of 84.55%, while the domain-specific prompt "which car is present in the image" improved accuracy to 88.83%. Further specifying the prompt to "which specific car is present in the image" resulted in an even higher accuracy of 90.20%. This demonstrates the importance of domain-specific prompts for domain-specific datasets.

4.2. Difference between Gemini Pro and Flash

Figure 4 compares the performance of our model using Gemini Pro and Gemini Flash LMM backbones in our SLAC model. While some datasets show slight performance variations between the two, the overall difference is not significant, and both achieve comparable results. Given the cost-effectiveness and speed advantages of Gemini Flash, it offers a compelling alternative to the more expensive and relatively slower Gemini Pro.

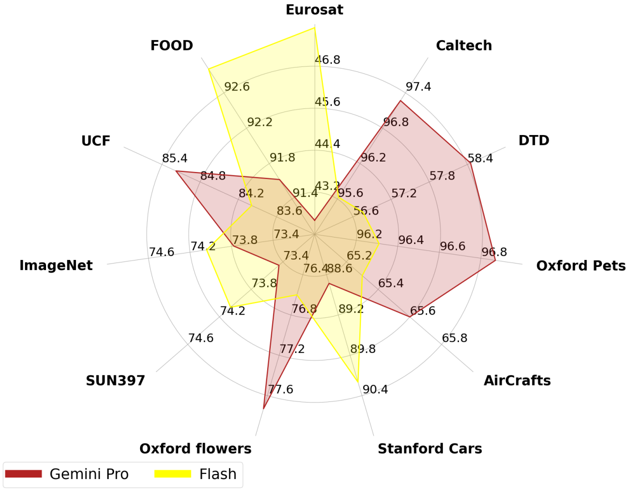

Figure 4. The figure illustrates the difference of accuracy between Gemini Pro and Gemini Flash LMMs in our model.

4.3. Effect of VLM

To evaluate the impact of the CLIP on our model architecture, we conducted experiments using the Caltech101 dataset and the Gemini Flash configuration. As illustrated in Figure 5, we observed a significant accuracy difference in the SLAC model. Without CLIP, utilizing only the LMM, Gemini Flash, resulted in an accuracy of 66.27%. Conversely, incorporating CLIP yielded a substantial improvement, reaching 96.62%. This disparity highlights the critical role of CLIP in the performance of SLAC model. In contrast, the TLAC model exhibited a minimal accuracy variation, achieving 96.18% without CLIP and 96.51% with CLIP, suggesting CLIP's reduced significance in this particular architecture.

4.4. Adding classes in prompt

While class names were not included in our standard Gemini prompts, our experiment demonstrates a performance boost when they are provided. For instance, accuracy of our SLAC model on the OxfordPets dataset improved from 96.48% to 98.27% with the inclusion of class names. To maintain a realistic evaluation, reflecting scenarios where class names are often unavailable, we did not incorporate them in our experiments.

5. Discussion and Future Direction

This work demonstrates the advantages of using pre-trained Large Multimodal Models (LMMs) for image classification. A key contribution is our approach that requires no train-

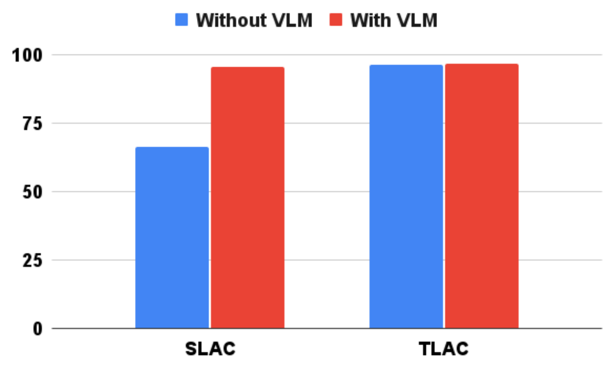

Figure 5. This figure demonstrates the impact of integrating a VLM, CLIP, on the accuracy of our proposed architecture.

ing or fine-tuning on new datasets and thus offering strong generalization capabilities. This generalization stems from the web-scale data used to train LMMs. While prior work has explored using CLIP, trained on 400 million image-text pairs [29], which in its original form often struggles to achieve high accuracy on various datasets, even with few-shot fine-tuning. In contrast, LMMs leverage significantly larger datasets. Consequently, our method achieves higher average accuracy on 13 datasets as compared to previous state-of-the-art training-free approaches. Similarly, our method also achieved better accuracy than previous few-shot methods on 9 of 11 standard image classification datasets and on all four domain generalization datasets, without any training or fine-tuning.

This work explores image classification with LMMs, achieving promising results. However, the potential for data contamination presents a critical challenge. Given the scale and diversity of LMM training data, it is possible that these models have encountered images similar to, if not portions of, commonly used benchmarks. To address this, future research should prioritize developing robust LMM evaluation methodologies, including the creation of large-scale, diverse, held-out datasets distinct from common training data. Such datasets are crucial for assessing true generalization ability of LMMs.

6. Limitations

While the aforementioned advantages are notable, our method also presents certain limitations. A primary concern pertains to the LMM model, specifically whether the observed accuracy is attributable to our unique model configuration or the inherent power of the Gemini architecture. Furthermore, the potential for data contamination exists, raising concerns that the LMM model may have been exposed to the datasets employed for evaluation. To mitigate the first limitation, future studies could investigate the performance of different LMM models to isolate the contribution of our specific configuration from that of the underlying backbone. Regarding the second limitation, the development of a novel benchmark dataset, previously unseen by the LMM, would be a valuable contribution. We recommend that future research endeavors focus on these areas.

Another limitation is that LMMs tend to show lower accuracy on highly specific datasets. For example, our model underperforms fine-tuned CLIP on OxfordPets because fine-tuning allows models to learn dataset-specific correlations, a significant advantage on specialized datasets like OxfordPets. Similarly, we observe this trend with EuroSAT, a low-resolution (64x64 pixel) aerial image dataset with 10 scene classes (e.g., River). Consequently, datasets with high specialization (e.g., OxfordPets) or low image resolution (e.g., EuroSAT) currently favor fine-tuned CLIP models over LMMs.

7. Literature Review

7.1. Vision Language Models

Recently, vision language models (VLMs), which are trained on both vision and text simultaneously, have gained significant popularity due to their impressive zero-shot and few-shot capabilities. Single modality models, such as vision or language models, are trained in isolation, which results in a modality gap between them. Recent VLMs include CLIP [29], ALIGN [14], FILIP [45], and Florence [46]. These VLMs have been trained on large amounts of data in a self-supervised manner. Despite being trained on large datasets, these VLMs still face challenges in achieving state-of-the-art results on downstream tasks.

7.2. Efficient Transfer Learning for VLMs

Efficient transfer learning methods, such as prompt learning and adapter-based approaches, have been introduced to adapt Vision-Language Models (VLMs) to downstream tasks. These methods add a small number of parameters to the pre-trained VLM, updating only these newly introduced parameters during fine-tuning while keeping the original VLM parameters frozen. Text prompts, consisting of sentence-based instructions can be either handcrafted or learned during training (a process known as “Prompt Learning”). Originally developed in Natural Language Processing (NLP) [18, 19, 21], prompt learning has been successfully adopted in VLMs [48–50]. Extending beyond text prompts, some methods [30, 38] have explored prompts within the VLM’s visual encoder. Maple [15] introduced multimodal prompt learning, training prompts in both the language and visual branches. CoOp [49] refines continuous prompt vectors within the language module to improve CLIP’s few-shot transfer learning performance. Address-ing CoOp's limitations in generalizing to unseen categories, Co-CoOp [48] conditions prompts on input images. Further enhancing prompt learning, [22] proposes optimizing multiple prompt sets by learning their underlying distribution.

In addition to prompt learning, an alternative approach involves integrating lightweight modules, termed adapters, into VLMs for adaptation to downstream tasks [11, 47]. CLIP-Adapter [11] fine-tunes vision-language models by incorporating feature adapters on either the visual or language branch. AdaptFormer [5], an approach for Vision Transformers (ViTs), introduces lightweight modules to enhance transferability without modifying pre-trained weights. Multi-Modal Adapter [43] aligns the representations learned by adapters in the vision and text branches by adding multimodal adapters.

7.3. Training-Free Methods

Tip-Adapter [47] is a training-free method that adapts CLIP for few-shot learning by adding a non-parametric adapter. SuS-X [39] adapts CLIP for training-free image classification by dynamically creating support sets from category names and using them within a zero-shot framework to guide predictions. These methods employ dataset class features for training-free inference. Alternative approaches utilize class descriptions. For examle, CuPL [28] combines a Large Language Model (LLM) with CLIP to generate customized prompts for image classification. By using an LLM, CuPL creates numerous category-specific prompts with detailed visual descriptions, enabling better discrimination between similar classes in a zero-shot manner. Meta-Prompting for Visual Recognition (MPVR) [25] uses a system prompt, task description, and a fixed in-context example to guide an LLM in generating visual style-infused query templates.

7.4. Large MultiModal Models (LMMs)

Following the rise of LLMs, Large Multimodal Models (LMMs) have become a prominent research area. Large multimodal models are developed in two stages: pre-training, where the model is trained on a massive dataset using self-supervised tasks like next-word prediction; and post-training, where it is fine-tuned to follow instructions. GPT-4 [1], developed by OpenAI, is a large multimodal model that can process both image and text inputs to generate text outputs. Claude 3 [2] is a multimodal model that accepts multiple input modalities, including text, tables, graphs, and photos. Developed by Google AI, Gemini [3] [34, 36] is a cutting-edge multimodal model capable of understanding and generating content across different modalities, including text, code, images, and other data types. Gemini comes in two versions: Flash and Pro. Flash is designed for speed and cost efficiency while Pro, is engineered for superior accuracy.

The large multimodal models mentioned above are closed-source, meaning their precise architecture and training data are not publicly disclosed. In contrast, open-source models make their architecture and trained weights publicly available. Pixtral-12B [2] is a 12-billion-parameter multimodal language model. Qwen2-VL [41] introduces a dynamic resolution mechanism for processing images. LLaVA (Large Language and Vision Assistant) [20] is an end-to-end trained LMM that combines CLIP with the Vicuna [6]. MiniGPT-4 [51] aligns a frozen visual encoder with a frozen large language model (Vicuna) to replicate GPT-4's advanced multimodal abilities. Llama [9] is a new family of language models. While Llama 3 [37] and 3.1 [9] were language models, Llama 3.2 introduced multimodal capabilities.

This paper proposes using a Large Multimodal Model (LMM) in a training-free fashion, eliminating the need for LLM fine-tuning due to its pre-training on extensive volume of data. Specifically, we utilize Gemini for inference due to its readily accessible API through the Google Cloud Platform (GCP). The initial $300 GCP credit provided at signup enabled us to use Gemini across all our datasets at virtually no cost. In contrast, open-source models like Llama require access to significant GPU resources for inference. Unfortunately we lack the compute resources to evaluate our scheme using Llama. Our work presents results using both Gemini Pro and Gemini Flash.

8. Conclusion

Standard pre-trained Vision-Language Models (VLMs) like CLIP often require fine-tuning to achieve high accuracy on downstream tasks such as image classification. Full model fine-tuning is computationally expensive, leading to the exploration of efficient alternatives like prompt learning (which introduces a small number of learnable parameters) and adapter layers (small multilayer perceptrons). However, both methods require fine-tuning for each new dataset, demanding significant computational resources. In contrast, we propose leveraging Large Multimodal Models (LMMs) such as Gemini for image classification. Trained on massive datasets, LMMs have demonstrated impressive language and vision understanding capabilities. Their extensive pre-training eliminates the need for further fine-tuning, allowing direct application to new datasets without additional training. Our results demonstrate that LMMs achieve significant improvements on large general datasets like ImageNet, as well as domain-specific datasets such as Flowers102, FGVCAircraft, and StanfordCars. Future works can explore other LMMs, both open-source and closed-source, and compare their performance with that of Gemini.

[2] https://claude.ai/

[3] https://gemini.google.com/app

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8

[2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 8

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 3

[4] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23232–23241, 2023. 4

[5] 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. 1, 8

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 8

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 3

[8] 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*, pages 248–255. Ieee, 2009. 3

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 8

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 3

[11] 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. 1, 8

[12] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*, 2024. 5

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 3

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 7

[15] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 1, 2, 3, 4, 5, 7

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

[17] 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*, pages 1401–1411, 2023. 4

[18] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 7

[19] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 7

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 8

[21] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 7

[22] 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*, pages 5206–5215, 2022. 4, 8

[23] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3

[24] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 3

[25] M Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Sivan Doveh, Jakub Micorek, Mateusz Kozinski, Hilde Kuehne, and Horst Possegger. Meta-prompting for automating zero-shot visual recognition with llms. In *European Conference on Computer Vision*, pages 370–387. Springer, 2025. 1, 3, 8

[26] 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*, pages 722–729. IEEE, 2008. 3

[27] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 3

[28] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 1, 3, 8

[29] 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*, pages 8748–8763. PMLR, 2021. 1, 3, 4, 5, 7

[30] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022. 7

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 3

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[33] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[34] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 8

[35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[36] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 8

[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 8

[38] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 7

[39] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2725–2736, 2023. 8

[40] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8

[42] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 3

[43] 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*, pages 23826–23837, 2024. 1, 2, 3, 4, 5, 8

[44] 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*, pages 6757–6767, 2023. 4

[45] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 7

[46] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 7

[47] 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*, pages 493–510. Springer, 2022. 2, 8

[48] 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*, pages 16816–16825, 2022. 1, 2, 3, 4, 7, 8

[49] 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, 2022. 1, 2, 3, 4, 5, 7

[50] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15659–15669, 2023. 7

[51] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 8