Supplementary Material

This section contains supplementary material that provides additional details for the main paper and further experimental analysis. This section follows the contents in the following order.

- Additional implementation details (Appendix A)
- Alternate prompting design choices (Appendix B)
- Understanding multi-modal prompts (Appendix C)
- Comparison of MaPLe with heavier Co-CoOp (Appendix D)

A. Additional Implementation details

In this section, we provide further hyper-parameter details of the proposed approaches presented in the main paper. Table 7 shows the hyper-parameters chosen for vision, language and independent V-L prompting techniques. We use a learning rate of 0.0025 for language and vision prompting, and 0.0035 for independent V-L prompting.

Method	Prompt Depth (K)	V-tokens ($\tilde{P}$)	T-tokens (P)
Language prompting	12	0	4
Vision prompting	12	4	0
I-V-L prompting	12	2	2

Table 7. Hyper-parameter settings for deep prompting variants. I-V-L refers to independent V-L prompting. Here K represents the depth of prompts. Number of prompt tokens in vision and language branches are denoted as $\tilde{P}$ and P respectively.

CoOp in Co-CoOp setting: The CoOp approach trained in Co-CoOp setting (denoted by CoOp†) uses training configurations of CoCoOp. Similar to Co-CoOp training, CoOp† trains the standard CoOp method for 10 epochs instead of default 200 epochs. We use a batch size of 4 with a learning rate of 0.0035.

B. Alternate Design Choices

Prompt Initialization: Table 8 shows the effect of prompt initialization on MaPLe. Best performance is achieved when the learnable prompts in the first layer are initialized with the prompt '`a photo of a` <`category`>' and rest of the layers are initialized randomly (row-3). Initializing prompts with a similar template in all layers leads to lower performance suggesting that this is redundant as these prompts learn hierarchically different contextual concepts in different layers (row-1). However, complete random initialization of prompts provides competitive performance (row-2). For implementation, if the number of learnable prompts $M = \#P$ are less than the total tokens of initial prompt template, we convert the former M word embeddings of template with learnable prompts and consider the rest of word embeddings of prompt template as fixed and use all token embeddings (learnable prompts + fixed word tokens) as input to text encoder.

Method	Base	Novel	HM
1: MaPLe: All layers: '`a photo of a`'	81.90	74.22	77.88
2: MaPLe: Random initialization	82.27	75.10	78.52
3: MaPLe: Only first layer: '`a photo of a`'	**82.28**	**75.14**	**78.55**

Table 8. Ablation on prompt initialization. In general, the performance of MaPLe is affected by the choice of prompt initialization.

Direction of prompt projection: As discussed in Section 3.2.3, MaPLe explicitly conditions the vision prompts $\tilde{P}$ on the language prompts P ($P \rightarrow \tilde{P}$) using a V-L coupling function $\mathcal{F}$. Here, we provide analysis for an alternative design choice where P is conditioned on $\tilde{P}$ ($\tilde{P} \rightarrow P$). Table 9 shows that our approach ($P \rightarrow \tilde{P}$) is a better choice which can be reasoned by the lower information loss in such a design since the dimension size d_v of vision prompts is greater than the dimension size d_l of language prompts.

Exploring other prompting designs: We provide analysis on other possible multi-modal prompting design choices in comparison to MaPLe. As learnable prompts in different transformer layers do not interact with each other, we explore a *progressive prompting* approach where the prompts at each block are conditioned on the prompts from the previous block via linear projection which are then added with the deep prompts initialized at every corresponding layer. We apply this approach to independent V-L prompting (row-1) and MaPLe (row-2). To analyze whether independent V-L prompting and MaPLe provide complementary gains, we also explore a design choice combining them together (row-3) in the same model. The results in Table 10 indicate that MaPLe provides best performance as compared to other design choices.

C. Understanding Multi-modal Prompts

Our experimental results in Section 4.3 indicates that the performance gains of MaPLe in comparison to Co-CoOp varies significantly across different datasets. For some datasets, like ImageNet and Caltech101, the gains are less than 1%, while on other datasets like EuroSAT, FGVCAircrafts and DTD, MaPLe shows significant improvements like +13% over Co-CoOp. To better understand at which cases MaPLe is most effective, we dissect the individual dataset performances and perform an exhaustive per-class analysis. Consistent with earlier work [1], we conjecture that CLIP pretraining dataset has been curated in a way that maximizes its zero-shot performance on ImageNet-1k and can be used as a proxy for CLIP pretraining dataset. Fur-

Prompt Proj.	Base	Novel	HM
$\tilde{P} \rightarrow P$	81.37	73.25	77.10
$P \rightarrow \tilde{P}$	**82.28**	**75.14**	**78.55**

Table 9. Projecting from P to $\tilde{P}$ provides the best results.

Method	Base	Novel	HM
1: I-V-L + Progressive prompting	81.20	74.92	77.93
2: MaPLe + Progressive prompting	81.45	75.04	78.11
3: MaPLe + I-V-L prompting	82.27	74.05	77.94
4: MaPLe	**82.28**	**75.14**	**78.55**

Table 10. Analysis on alternative design choices for V-L prompting. Overall, MaPLe proves to be the best variant among alternate prompting-related design choices.

ther, datasets like EuroSAT (satellite images) and DTD (texture dataset) has more distributional gap from ImageNet [1]. Fig. 5 shows per class analysis for selected datasets in the order of increasing diversity (distribution gap w.r.t CLIP pretraining dataset, *i.e.* generic objects). The overall trend indicates that MaPLe is more effective than Co-CoOp as the diversity of the dataset increases. We conjecture that this is because fine-tuning or prompting bridges the gap between the distribution of the downstream and the pretraining dataset and thus improves the performance. However, the effectiveness would therefore be less substantial for datasets with little distribution shifts. This intriguing property is also validated for visual prompting in literature [1]. MaPLe provides completeness in prompting by learning both vision and language prompts to effectively steer CLIP, this makes it more adaptive than Co-CoOp to improve on datasets with larger distribution shifts.

Additionally, we note that MaPLe benefits on categories which would have been rarely seen by CLIP during its pretraining (400 million image caption dataset, obtained from internet images). We observe that MaPLe provides significant gains over Co-CoOp for vision concepts that tend to be rare and less generic, *e.g.*, satellite images. In contrast, MaPLe performs competitively to Co-CoOp on frequent and more generic categories *e.g.*, forest, river, dog, *etc.* Multi-modal prompts allow MaPLe to better adapt CLIP for visual concepts that are rarely occurring as compared to existing uni-modal prompting techniques. In Table 12, we highlight category-wise comparison between MaPLe and Co-CoOp for some selected datasets.

Text embeddings analysis: As all samples within a category are represented using a single text embedding, we take a quantitative approach in Tab. 11 for analyzing the text embeddings of CoOp and MaPLe. We show the pairwise cosine similarity and normalized l_2 distance metrics averaged across text embeddings. We observe that MaPLe shows better separability among the categories.

	l_2 distance ↑			Cosine similarity ↓		
Method	DTD	UCF	EuroSAT	DTD	UCF	EuroSAT
CoOp	0.87	0.85	0.57	0.62	0.63	0.83
MaPLe	**0.93**	**0.87**	**0.78**	**0.57**	**0.62**	**0.69**

Table 11. Avg. cosine similarity and l_2 distance of text embeddings. MaPLe shows better separability among the text categories.

Dataset	MaPLe is better than Co-CoOp	Co-CoOp is better than MaPLe
Caltech101 (Generic Objects)	Crontosaurus, Gerenuk, Sea Horse	Elephant, Ceiling Fan, Cellphone
EuroSAT (Satellite Image)	Annual Crop Land, Permanent Crop Land	-
UCF101 (Action recognition)	Handstand Walking, Playing Daf	Walking With Dog, Horse Riding

Table 12. Analyzing the nature of categories where MaPLe performs better than Co-CoOp. Co-CoOp performs favourably well on generic categories, while MaPLe provides benefits on classes that are typically rare.

D. Comparing MaPLe with Heavier Co-CoOp

The multi-modal deep prompting architecture design of MaPLe along with its V-L coupling function $\mathcal{F}$ constitutes more learnable parameters as compared to CoOp and Co-CoOp. To verify that the performance gain is not due to increased parameter count, we compare Co-CoOp with MaPLe shallow ($J = 1$) that utilizes prompts only at the first layer of vision and language branch of CLIP. Further, we also experiment with a heavier Co-CoOp in which we retrain a version of Co-CoOp that matches the parameter count of MaPLe ($J = 9$) by stacking multiple additional layers in its Meta-Net block. Table 13 indicates the effectiveness of multi-modal prompting in MaPLe (for both $J = 1$ and $J = 9$) over the heavier Co-CoOp. In addition to that, we experiment with MaPLe†, which uses a unified V-L coupling function for all layer prompts. MaPLe† with about 9x lesser parameters than MaPLe also improves over existing methods. This shows that the difference in the number of parameters is not the cause of gain in our case and the proposed multi-modal prompting design choice makes a difference.

Method	Base	Novel	HM
Co-CoOp	80.47	71.69	75.83
Heavier Co-CoOp	80.14	72.02	75.86
MaPLe shallow ($J = 1$)	80.10	73.52	76.67
MaPLe† ($J = 9$)	82.29	74.34	78.11
MaPLe ($J = 9$)	**82.28**	**75.14**	**78.55**

Table 13. Comparison of MaPLe with a heavier Co-CoOp model. We retrain a heavier version of Co-CoOp which is comparable with MaPLe in terms of total parameter count. MaPLe† is a MaPLe version which utilizes a common V-L coupling function for all layers.