Supplemental 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)
- Analysis for understanding multi-modal prompts (Appendix B)
- Analysis for alternate prompting design choices (Appendix C)
- Prompting complexity (Appendix D)
- Comparison of MaPLe with heavier Co-CoOp (Appendix E)

A Additional Implementation details

In this section, we provide further hyper-parameter details of the proposed approaches presented in the main paper. Table 6 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.

Table 6: Hyper-parameter settings for deep prompting variants.

Method	Prompt Depth (K)	Visual tokens $\tilde{P}$	Textual tokens P
Language prompting	12	0	4
Vision prompting	12	4	0
Independent V-L prompting	9	2	2

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

B Analysis for Understanding Multi-modal Prompts

Our experimental results in Section 4.4 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 over Co-CoOp. To better understand when MaPLe is most effective, we dissect the individual dataset performances and perform an exhaustive per-class analysis. 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 (Bahng et al., 2022). 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 7, we highlight category-wise comparison between MaPLe and Co-CoOp for some selected datasets.

Table 7: 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.

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

C Analysis for 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).

Table 8: Ablation on prompt initialization. In general, the performance of MaPLe is affected by the choice of prompt initialization. Best results are achieved when only first layer prompts are initialized with the prompt '`a photo of a <category>`'.

Method	Base Acc.	Novel Acc.	Harmonic Mean (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**

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 $d_v > d_l$.

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

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**

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. We apply this approach to shallow versions (prompt depth $J = 1$) of independent V-L prompting (row-1) and MaPLe (row-2). To analyze whether independent V-L prompting and MaPLe provide complementary gains, we 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.

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.

Method	Base Acc.	Novel Acc.	Harmonic Mean (HM)
1: Independent V-L prompting + Progressive prompting	81.20	74.92	77.93
2: MaPLe + Progressive prompting	81.45	75.04	78.11
3: MaPLe + Independent V-L prompting	82.27	74.05	77.94
4: MaPLe	**82.28**	**75.14**	**78.55**

D PROMPTING COMPLEXITY

Table 11 shows the computational complexity of MaPLe in comparison with other approaches. Although MaPLe utilizes multi-modal prompts, its overall FLOPS (Floating Point Operations) exceeds only by 0.1% over CoOp and Co-CoOp. The independent V-L prompting also provides comparable FLOP count. In terms of inference speed, Co-CoOp is significantly slower and the FPS (Frames Per Second) remains constant as the batch size increases. In contrast, MaPLe has no such overhead and provides much better inference and training speeds. Further, MaPLe provides better convergence as it requires only half training epochs as compared to Co-CoOp (5 vs 10 epochs).

Table 11: Comparison of computational complexity among different prompting methods.

Method	GFLOPS	% GFLOPS w.r.t Co-CoOp	FPS BS=1	FPS BS=4	FPS BS=100	BS Overhead
CoOp	166.8	0.0	13.8	55.3	1353.0	No
Co-CoOp	166.8	0.0	13.9	14.9	15.1	Yes
Independent V-L prompting	167.1	0.2	14.1	56.7	1350.0	No
MaPLe	167.0	0.1	14.1	56.3	1365.0	No

E 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. For a fair comparison, we retrain a heavier Co-CoOp that matches the parameter count of MaPLe by stacking multiple additional layers in its Meta-Net block. Table 12 indicates the effectiveness of multi-modal prompting in MaPLe over the heavier Co-CoOp. This shows that the difference in the number of parameters is not the cause of gain in our case and the proposed design choices make a difference.

Table 12: 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.

Method	Base Acc.	Novel Acc.	Harmonic Mean (HM)
Co-CoOp	80.47	71.69	75.83
Heavier Co-CoOp	80.14	72.02	75.86
MaPLe	**82.28**	**75.14**	**78.55**