# MMRL: Multi-Modal Representation Learning for Vision-Language Models

## Supplementary Material

## A. Implementation Details

We follow prior studies[17, 47–49, 55–57] and adopt a 16-shot learning setting across all experiments, except for the few-shot learning tasks. The ViT-B/16[8] variant of the CLIP model serves as the visual backbone for all experimental setups. Hand-crafted text prompts from prior methods[34, 52, 56] are utilized and described in detail in Tab. 7. Optimization is performed using the AdamW optimizer with an initial learning rate of 0.001. All our models are trained with mix-precision for speeding up. For the larger ImageNet dataset, we employ a batch size of 32, while a batch size of 4 is used for all other datasets. Training on ImageNet for the base-to-novel generalization task spans 5 epochs, whereas training on the remaining datasets is conducted over 10 epochs. For cross-dataset evaluation and domain generalization tasks, we perform training for a single epoch on ImageNet. In the few-shot learning tasks, training is carried out for 5 epochs on ImageNet and 50 epochs for other datasets. The average accuracy is reported over three independent runs, with all experiments executed on a single NVIDIA RTX 4090 GPU.

Representation tokens are initialized from a zero-mean Gaussian distribution with a standard deviation of 0.02. We set $J = 6$, integrating the representation tokens beginning at the 6-th transformer layer. The dimension of the representation space, $d_r$, is set to 2048 for EuroSAT and 512 for all other datasets. Note that since the $d_r$ setting for EuroSAT differs from other datasets, in the $d_r$ ablation experiments we fix $d_r$ for EuroSAT to 2048 while adjusting $d_r$ on the other datasets. The number of representation tokens, $K$, is configured to 5. The parameter $\alpha$ is fixed at 0.7, and the details regarding the configuration of $\lambda$ are provided in Tab. 8.

## B. Dataset Details

Details of 14 datasets are shown in Tab. 7.

## C. Computational Cost

Table 6 summarizes the learnable parameters, training time per image, total training duration, inference speed (measured in frames per second, FPS, with a batch size of 100), and the final HM metric for each approach. Our proposed model, MMRL, demonstrates a compelling balance of computational efficiency and performance. The key observations are as follows:

- Models incorporating multimodal interaction mechanisms (e.g., MaPLe, MMA, and MMRL) generally involve a higher parameter count compared to models without such mechanisms.
- Both MMRL and the prior MMA approach exhibit significantly faster training speed, thereby reducing overall computational costs. While MaPLe and PromptSRC achieve higher inference speeds, their training durations are relatively longer. Notably, MMRL offers faster inference compared to MMA and MetaPrompt.
- To assess the performance of MMRL under constrained computational resources, we reduced the dimensionality of the representation space from 512 to 32. In this configuration, MMRL achieves a parameter count comparable to that of MMA, while still significantly outperforming the previous state-of-the-art model.

Table 6. All methods were trained on a single NVIDIA RTX 4090 GPU using the ImageNet dataset. Each model was implemented with publicly available code and default configurations as described in their respective papers [17, 18, 46, 47, 49, 53]. 'V-L' denotes vision-language interaction, indicating that efficient fine-tuning incorporates interactions between visual and textual modalities before prediction. 'V, L' signifies separate fine-tuning of each modality without inter-modal interaction before prediction, while 'L' refers to fine-tuning limited to the textual modality alone. 'Train time' is reported as both time per image and the total duration for training the full dataset(16-shots), while 'FPS (100 BS)' indicates frames per second with a batch size of 100 during inference.

| Method | Modality | Params (learnable) | Train time (ms/image) | Train time (minute/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 | 1.5 | 688.5 | 79.87 |
| MMRL | V-L | 4.992M | 5.3 | 3.6 | 762.4 | 81.20 |
| MMRL* | V-L | 0.689M | 5.3 | 3.6 | 767.8 | 80.84 |

## D. Ablation Analysis on $\lambda$

As shown in Tab. 8, increasing the value of $\lambda$ generally improves performance, with the optimal or near-optimal results typically observed when $\lambda$ is set between 4 and 6 across most datasets. Notably, as $\lambda$ continues to increase, its impact on model performance within the same dataset diminishes, indicating reduced sensitivity to variations in $\lambda$. This trend suggests that the model becomes more robust and less reliant on precise tuning of $\lambda$ at higher values.

Table 7. Summary of the 14 datasets.

| Dataset | Classes | Train | Val | Test | Description | Prompt |
|---|---|---|---|---|---|---|
| ImageNet | 1000 | 1.28M | ∼ | 50000 | Recognition of generic objects | "a photo of a [CLASS]." |
| Caltech101 | 100 | 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 aircrafts | "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 | 1,000 | ∼ | ∼ | 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. Ablation on $\lambda$ across 11 datasets, with results evaluated using the harmonic mean (HM) metric.

| $\alpha$ | ImageNet | Caltech101 | OxfordPets | StanfordCars | Flowers102 | Food101 | FGVCAircraft | SUN397 | DTD | EuroSAT | UCF101 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 74.01 | 95.97 | 96.35 | 76.00 | 84.42 | 90.10 | 38.52 | 79.67 | 68.21 | 82.65 | 81.63 |
| 0.01 | 74.07 | 96.12 | 96.39 | 75.95 | 84.82 | 90.23 | 37.87 | 79.85 | 67.73 | **87.21** | 82.11 |
| 0.1 | 74.23 | 96.25 | 96.49 | 76.32 | 84.81 | 90.53 | 38.66 | 80.23 | 69.79 | 83.21 | 82.91 |
| 0.2 | 74.38 | 96.40 | **96.74** | 76.67 | 85.31 | 90.61 | 39.27 | 80.25 | 70.58 | 82.68 | 82.70 |
| 0.5 | **74.45** | **96.68** | 96.54 | 77.09 | 85.74 | 90.86 | 40.37 | 80.61 | 72.67 | 82.87 | 83.05 |
| 3.0 | 74.09 | 96.59 | 96.51 | 77.72 | 86.65 | 90.98 | 40.48 | 81.10 | 73.54 | 77.95 | **83.89** |
| 4.0 | 74.04 | 96.62 | 96.55 | 77.73 | **86.78** | 90.98 | 40.66 | 81.14 | 73.75 | 77.27 | 83.45 |
| 5.0 | 73.93 | 96.62 | 96.60 | 77.86 | 86.42 | **91.03** | 40.42 | 81.07 | 73.69 | 78.05 | 83.84 |
| 6.0 | 73.83 | 96.61 | 96.66 | 78.05 | 86.48 | 91.00 | **41.15** | **81.20** | **73.82** | 75.23 | 83.68 |
| 7.0 | 73.78 | 96.62 | 96.58 | **78.06** | 86.53 | 90.95 | 40.88 | 81.10 | 73.65 | 75.85 | 83.55 |
| 10.0 | 73.68 | 96.64 | 96.56 | 77.86 | 86.46 | 91.00 | 41.01 | 80.93 | 73.68 | 77.61 | 83.38 |

Table 9. Ablation on different regularization strategies.

| Regularization | Base | Novel | HM |
|---|---|---|---|
| Cosine | **85.68** | **77.16** | **81.20** |
| L1 | 85.46 | 76.03 | 80.47 |
| MSE | 85.13 | 74.62 | 79.53 |

## E. Ablation Analysis on Regularization Strategies

We investigate the impact of various regularization strategies aimed at maximizing the similarity between class token features and frozen CLIP features to retain pre-trained knowledge. The results, summarized in Tab. 9, indicate that cosine regularization achieves the best performance. In contrast, both L1 and MSE losses lead to performance degradation, with MSE causing a significant decline. This result can be attributed to the more relaxed and flexible constraints of cosine regularization, enabling the class token to preserve generalizability while effectively capturing task-specific knowledge.

## F. Few-Shot Learning

Tabs. 10 and 11 provide detailed comparisons of MMRL and prior state-of-the-art methods on few-shot learning across 11 datasets. MMRL achieves the highest average performance across all shots. Note that the MMA results are reproduced from the open-source code, as the original paper does not report results for this experiment.

Table 10. Comparison of MMRL 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 |
| | MMRL (Ours) | **72.67** | **75.90** | **79.20** | **81.47** | **84.34** |
| 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 |
| | MMRL (Ours) | 69.00 | 70.30 | **71.40** | **72.33** | **73.40** |
| 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 |
| | MMRL (Ours) | **94.17** | **94.83** | **96.03** | **96.27** | **97.13** |
| 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 |
| | MMRL (Ours) | 90.87 | 91.57 | 92.57 | 93.03 | **93.83** |
| 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 |
| | MMRL (Ours) | 68.70 | 72.93 | **78.17** | **82.57** | **86.43** |
| 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 |
| | MMRL (Ours) | **85.97** | **91.20** | **94.60** | **96.60** | **98.40** |

Table 11. Comparison of MMRL 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 |
|---|---|---|---|---|---|---|
| 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 |
| | MMRL (Ours) | 84.87 | 85.53 | 85.77 | 86.33 | 87.03 |
| 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 |
| | MMRL (Ours) | 28.53 | **34.23** | **40.47** | **48.07** | **57.60** |
| 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 |
| | MMRL (Ours) | 68.90 | 71.53 | 73.93 | **76.00** | **77.70** |
| 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 |
| | MMRL (Ours) | **56.37** | **61.37** | **67.87** | **71.60** | **75.30** |
| 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 |
| | MMRL (Ours) | **76.00** | **82.87** | **87.67** | 88.73 | **93.37** |
| 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 |
| | MMRL (Ours) | **75.97** | **78.50** | **82.67** | **84.67** | **87.60** |