PromptSync: Bridging Domain Gaps in Vision-Language Models through Class-Aware Prototype Alignment and Discrimination

Supplementary Material

Method	Top-1 Average Accuracy(%)	Latency
MaPLe + TPT	58.08	0.41
PromptAlign	59.37	0.46
PromptSync*	61.88	0.49
PromptSync	61.92	0.65

Table 7. **Performance and Latency**: Performance and Latency comparison of PromptSync with state-of-the-art baselines and its variant which reuse the learned prompt tokens after prototype discrimination without learning them for each incoming test sample.

9. Benchmark Settings

Base-to-Novel Generalisation: Following MaPLe [21], we evaluate PromptSync on a zero-shot setting. We split the dataset into base and novel classes. The model is trained only on the base classes in a few-shot setting and evaluated on the base and novel classes.

Cross-dataset Transfer: We evaluate PromptSync on the ImageNet[11] pre-trained model on other datasets to determine the transfer performance. Following CoCoOp[46], our model is trained on all 1000 ImageNet classes in a few-shot manner.

Domain Generalisation: We evaluate PromptSync on out-of-distribution (OOD) datasets for domain generalizability. Similar to cross-dataset, we evaluate our ImageNet-trained model directly on OOD datasets, which are described in Section 4.

10. Performance and Latency

The experiments presented in the table 7 above involve a comparison of different methods, namely MaPLe + TPT, PromptAlign, PromptSync*, and PromptSync. In these experiments, we evaluated the top-1 average accuracy (%) and latency (in hours for a single prompt update) of each method. Specifically, we investigated PromptSync with and without saving the updated prompt obtained after prototype discrimination, with the variant denoted as PromptSync* indicating the adaptation of prompt tokens for test samples after restoring saved prompt tokens.

The results, as shown in Table 7, include latency measurements represented in hours for a single prompt update, and all evaluations are conducted on the ImageNet-A dataset. Notably, the PromptSync* variant demonstrates a faster processing time compared to the full PromptSync method, with only a marginal drop in performance. This outcome underscores the achieved generalization through prototype alignment. Furthermore, in comparison to previous methods such as MaPLe + TPT and PromptAlign, the PromptSync* variant exhibits only a slight increase in latency (0.03 hours) while still improving overall performance.

11. Sensitivity Comparison

We further performed the sensitivity comparison of our method as compared to other state-of-the-art baselines. Figure 2(a) shows the comparison of performance during test time adaptation as the number of views increases. All the results are on ImageNet-A dataset. In comparison to PromptAlign and MaPLe + TPT, their performance almost plateaus around 64 views with insignificant improvement further, while PromptSync shows a consistent improvement with the increase in views and insignificant improvement beyond 128. This proves the generalizability achieved by our method since it optimises base CLIP over a larger number of possible shifts in the dataset, resulting in better performance. Figure 2(b) shows the performance comparison as the number of prompt update steps increases. All the methods increase their performance with an increase in the number of steps; however, our method shows better adaptation to the test sample with more steps in comparison to PromptAlign and MaPLe + TPT. For apples-to-apples comparison we perform a single-step update (with 128 views) following TPT [35].

12. LAION400M Proxy Dataset Analysis

Given CLIP's impressive zero-shot performance on ImageNet, we opted for ImageNet as a viable proxy source dataset, aligning with prior research [33]. We worked with a subset of LAION400M, comprising 2.5 million images (2 times the size of ImageNet). Furthermore, we carried out an ablation study on the alignment strategy using LAION400M as the source dataset, a dataset known to mirror CLIP's training dataset [9]. The results for this ablation study is shown in Table 8. Notably, the performance impact remains consistent when utilizing this subset of LAION400M alongside ImageNet. Source class prototypes are computed on the proxy source data to derive the distribution for alignment during test time. As this proxy dataset aligns with the model's training set, this offline computation remains unchanged despite environmental shifts and only necessitates computation once.

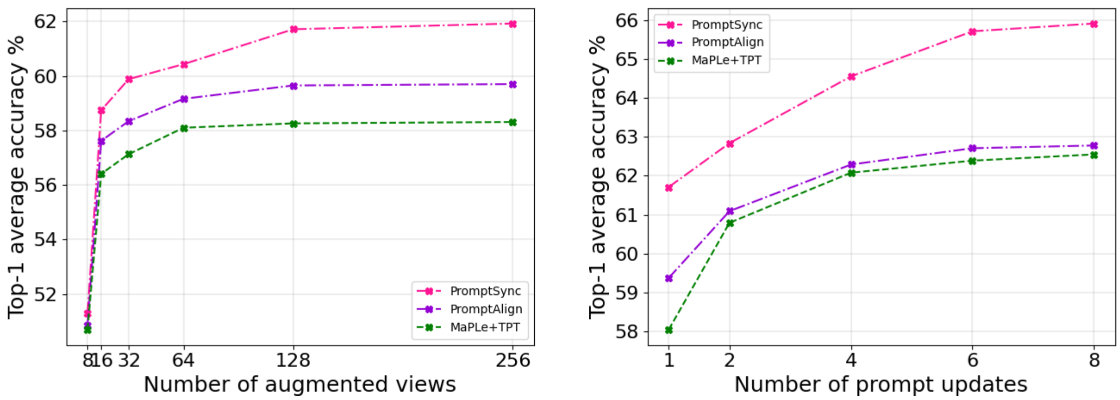

Figure 2. **Sensitivity Comparison**. (a) Top-1 accuracy improves with number of augmented views (b) Top-1 accuracy improves consistently with number of prompt update steps.

Method	Flowers	DTD	Pets	Cars	UCF	Caltech	Food	SUN	Aircraft	Eurosat	Avg
ImageNet	77.68	50.99	91.89	69.24	71.04	95.78	87.72	67.98	25.91	59.36	69.74
LAION	77.68	51.00	91.88	69.25	71.03	95.79	87.75	68.00	25.90	59.35	69.76

Table 8. Performance impact analysis using both ImageNet and LAION400M subset