Supplementary Material

Self-regulating Prompts: Foundational Model Adaptation without Forgetting

The following section contains supplemental information and encompasses more implementation details, results comparison, and a thorough ablative analysis of PromptSRC. The contents are organized in the following order.

- Additional implementation details (Appendix A)
- Additional results comparison (Appendix B)
- Additional ablative analysis (Appendix C)

A. Additional Implementation details

Additional Training details: We use a publically available ViT-B/16 CLIP model with $d = 512$ and use a learning rate of 0.0025 which is fixed for all experiments in all benchmarks. We train PromptSRC for 50 epochs for few-shot settings and 20 epochs for the remaining three benchmark settings respectively. The respective epochs are fixed across all datasets. All models are trained using SGD optimizer and utilize a single NVIDIA A100 GPU.

Gaussian Weighted Prompt Aggregation (GPA): We note that the prompts learned in the initial training epochs are not mature and act as noise due to their random initialization. On the other hand, prompts learned in the last few epochs are task-specific and highly favors the supervised downstream task distribution. GPA strives to maintain a balance by assigning lower weights to initial prompts, higher weights to middle prompts, and relatively lower weights to final prompts, resulting in optimal prompt representations that improve generalization to downstream tasks. Gaussian distribution in GPA is defined over the epochs and its mean is dictated by the epoch number. We then sample weights ($w_i \sim \mathcal{N}(\mu, \sigma^2)$) for prompts of every epoch to get the final prompt aggregation. Hyper-parameters are set using validation splits Table 8 shows the hyper-parameter values chosen for the proposed GPA technique, which are kept fixed for respective base-to-novel generalization, cross-dataset and domain generalization setting. For few-shot setting, we use $\mu = 30$ and $\sigma^2 = 30$ for ImageNet, Caltech101, OxfordPets, Food101, UCF101 and SUN397. For datasets including StanfordCars, Flowers102, FGVCAircraft, DTD and EuroSAT, we use $\mu = 45$ and $\sigma^2 = 5$.

GPA parameter	Base-to-Novel	Cross dataset	D.G
μ	15	6	6
σ^2	1	10	10

Table 8: Hyper-parameters settings used in GPA technique for various benchmark settings. D.G refers to domain generalization.

Textual diversity: For the textual diversity technique, we randomly select 60 prompt templates from the complete template list provided in [35]. Specifically, our textual diversity component uses the following prompt templates.
"a photo of a {category}."
"a bad photo of a {category}."
"a photo of many {category}."
"a sculpture of a {category}."
"a photo of the hard to see {category}."
"a low resolution photo of the {category}."
"a rendering of a {category}."
"graffiti of a {category}."
"a bad photo of the {category}."
"a cropped photo of the {category}."
"a tattoo of a {category}."
"the embroidered {category}."
"a photo of a hard to see {category}."
"a bright photo of a {category}."
"a photo of a clean {category}."
"a photo of a dirty {category}."
"a dark photo of the {category}."
"a drawing of a {category}."
"a photo of my {category}."
"the plastic {category}."
"a photo of the cool {category}."
"a close-up photo of a {category}."
"a black and white photo of the {category}."
"a painting of the {category}."
"a painting of a {category}."
"a pixelated photo of the {category}."
"a sculpture of the {category}."
"a bright photo of the {category}."
"a cropped photo of a {category}."
"a plastic {category}."
"a photo of the dirty {category}."
"a jpeg corrupted photo of a {category}."
"a blurry photo of the {category}."
"a photo of the {category}."
"a good photo of the {category}."
"a rendering of the {category}."
"a {category} in a video game."
"a photo of one {category}."
"a doodle of a {category}."
"a close-up photo of the {category}."
"the origami {category}."
"the {category} in a video game."
"a sketch of a {category}."
"a doodle of the {category}."
"a origami {category}."
"a low resolution photo of a {category}."
"the toy {category}."
"a rendition of the {category}."
"a photo of the clean {category}."

"a photo of a large {category}."
"a rendition of a {category}."
"a photo of a nice {category}."
"a photo of a weird {category}."
"a blurry photo of a {category}."
"a cartoon {category}."
"art of a {category}."
"a sketch of the {category}."
"a embroidered {category}."
"a pixelated photo of a {category}."
"itap of the {category}."

Evaluation metrics: We report top-1 base-class and novel-class accuracy for each dataset in base-to-novel generalization setting. We also report harmonic mean (HM) between base and novel class accuracy which is the main metric that represents generalization performance.

For all shots ($K = 1, 2, 4, 8, 16$) in few-shot setting, we report top-1 accuracies obtained on the corresponding test-set of each dataset using the splits provided in CoOp [59].

Similar to few-shot setting, we report top-1 accuracies obtained on the test set of each dataset for cross dataset evaluation and domain generalization experiments respectively.
Algorithm: In algorithm 1, we show the pseudo-code implementation of our proposed PromptSRC framework.

B. Additional results comparison

In this section, we provide additional per-dataset results comparison and show the compatibility of PromptSRC for diverse tasks and recent VL models.
Generalization of PromptSRC towards video understanding tasks: We verify the applicability of our approach across new tasks and evaluate PromptSRC on a video action recognition generalization benchmark. Following the base-to-novel generalization setting of ViFi-CLIP [37], we employ PromptSRC on a Kinetics-400 pre-trained ViFi-CLIP [37] and learn prompts on UCF-101 video dataset. The results are shown in Table 9. In comparison with the naive IVLP method, PromptSRC shows favorable performance gains and even surpasses fully fine-tuned video-adapted CLIP models like ActionCLIP. This suggests that the proposed PromptSRC approach can generalize to other diverse modality downstream tasks including videos.
Compatibility of PromptSRC in recent foundational VL models: We have demonstrated the effectiveness of our approach on the CLIP Vision-Language (VL) model in the main manuscript. To assess how our approach scales with more recent foundational VL models, we conduct analysis using a newly introduced VL model, EVA-CLIP (CVPR'23) [9]. EVA-CLIP has been pre-trained using advanced self-supervision and optimization techniques. We employ the IVLP and PromptSRC prompting approaches to fine-tune the EVA-CLIP ViT-B/16 model in the base-to-novel generalization setting. The comparison of results is presented in Table 11. PromptSRC consistently improves the generalization performance on 10/11 datasets and provides an absolute average HM gain of +2.09% in comparison with the IVLP baseline approach.

Algorithm 1 Learning Self-regulating prompts

Input: Dataset $\mathcal{D} = \{X, y\}^N$, Model $\theta_{\text{CLIP}} = \{\theta_g, \theta_f\}$, Prompt vectors $\boldsymbol{P} = \{\boldsymbol{P_v}, \boldsymbol{P_t}\}$. No. of text templates = N. iteration (i) = 1.
Require: Initialize GPA prompt param. $\boldsymbol{P}^{\text{GPA}} = \{\boldsymbol{p_v}, \boldsymbol{p_t}\}^{GPA}$. Sample Gaussian weights for GPA $\{w_1, w_2, w_3, \cdot, w_T\}$. GPA is applied after every c iterations.
for i $\in [1, T]$ **do**
  sample data $\{X, y\} \subseteq \mathcal{D}$
  `// prompted features.`
  Using θ_{CLIP} and $\boldsymbol{P}$, obtain prompted visual and text features
  $\tilde{\boldsymbol{f}}_{\boldsymbol{p}} \leftarrow f(\tilde{\boldsymbol{x}}_{\boldsymbol{p}}, \theta_f)$, $\tilde{\boldsymbol{g}}_{\boldsymbol{p}} \leftarrow g(\tilde{\boldsymbol{y}}_{\boldsymbol{p}}, \theta_g)$
  `// normal CE supervision loss.`
  $\mathcal{L}_{\text{sup}} \leftarrow \mathcal{L}_{\text{CE}}(\text{sim}(\tilde{\boldsymbol{f}}_{\boldsymbol{p}}, \tilde{\boldsymbol{g}}_{\boldsymbol{p}}), y)$
  `// pre-trained features.`
  Obtain pre-trained visual and textual features using only θ_{CLIP}
  $\tilde{\boldsymbol{f}} \leftarrow f(\tilde{\boldsymbol{x}}, \theta_f)$, $\tilde{\boldsymbol{g}} \leftarrow \frac{1}{N}\sum_{i=1}^{N} g(\tilde{\boldsymbol{y}}^{\boldsymbol{i}}, \theta_g)$
  `// self-regularizing consistency losses.`
  $\mathcal{L}_{\text{SCL}} \leftarrow \lambda_1 \mathcal{L}_{\text{SCL-image}}(\tilde{\boldsymbol{f}}_{\boldsymbol{p}}, \tilde{\boldsymbol{f}}) + \lambda_2 \mathcal{L}_{\text{SCL-text}}(\tilde{\boldsymbol{g}}_{\boldsymbol{p}}, \tilde{\boldsymbol{g}}) + \mathcal{L}_{\text{SCL-logits}}(\text{sim}(\tilde{\boldsymbol{f}}_{\boldsymbol{p}}, \tilde{\boldsymbol{g}}_{\boldsymbol{p}}), \text{sim}(\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{g}}))$
  `// compute total loss.`
  $\mathcal{L}_{\text{final}} \leftarrow \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{SCL}}$
  `// update prompt vectors with combined loss.`
  $\boldsymbol{P} \leftarrow \boldsymbol{P} - \delta \nabla_{\boldsymbol{P}} \mathcal{L}_{\text{final}}$
  `// Gaussian prompt ensembling.`
  if mod(i, c) $== 0$ **then**
    $\boldsymbol{P}^{\text{GPA}} \leftarrow \boldsymbol{P}^{\text{GPA}} + w_{\text{i}}.\boldsymbol{P}$
  end if
end for

Method	Base Acc.	Novel Acc.	HM
Vanilla CLIP	78.50	63.60	70.30
ActionCLIP	85.60	75.30	80.10
XCLIP	95.40	74.00	83.40
A5	95.80	71.00	81.60
IVLP	95.90	74.10	83.60
PromptSRC	**96.43**	**76.79**	**85.50**

Table 9: Performance comparison in video action recognition generalization benchmark on UCF-101. We employ PromptSRC and IVLP on ViFi-CLIP and compare with the prior video approaches.

Method	Base Acc.	Novel Acc.	HM
Independent V-L prompting (IVLP)	84.21	71.79	77.51
PromptSRC with single prompt diversity	**84.32**	75.52	79.68
PromptSRC with ensembled prompt diversity	84.26	**76.10**	**79.97**

Table 10: Analysis on alternate design choices for the textual diversity in PromptSRC. Incorporating textual diversity by ensembling multiple text templates achieves better generalization.

Dataset		IVLP	PromptSRC	Δ
Average on 11 datasets	Base Acc.	86.31	**86.34**	+0.03
	Novel Acc.	74.96	**78.68**	+3.72
	HM	80.24	**82.33**	+2.09
ImageNet	Base Acc.	82.13	**82.40**	+0.27
	Novel Acc.	72.20	**76.03**	+3.83
	HM	76.85	**79.09**	+2.24
Caltech101	Base Acc.	**99.33**	98.97	-0.36
	Novel Acc.	96.47	**97.10**	+0.63
	HM	97.88	**98.03**	+0.15
OxfordPets	Base Acc.	95.17	95.63	+0.46
	Novel Acc.	98.43	98.43	+0.00
	HM	96.77	**97.01**	+0.24
Stanford Cars	Base Acc.	**85.90**	85.07	-0.83
	Novel Acc.	83.97	**86.40**	+2.43
	HM	84.92	**85.73**	+0.81
Flowers102	Base Acc.	99.47	99.47	+0.00
	Novel Acc.	77.43	**79.57**	+2.14
	HM	87.08	**88.41**	+1.34
Food101	Base Acc.	90.60	**91.37**	+0.77
	Novel Acc.	90.70	**91.97**	+1.27
	HM	90.65	**91.67**	+1.02
FGVC Aircraft	Base Acc.	**46.80**	46.40	-0.40
	Novel Acc.	**28.90**	28.80	-0.10
	HM	**35.73**	35.54	-0.19
SUN397	Base Acc.	83.30	**84.50**	+1.20
	Novel Acc.	76.93	**80.80**	+3.87
	HM	79.99	**82.61**	+2.62
DTD	Base Acc.	84.60	**86.27**	+1.67
	Novel Acc.	59.47	**63.53**	4.06
	HM	69.84	**73.17**	+3.33
EuroSAT	Base Acc.	**96.13**	93.43	-2.70
	Novel Acc.	62.90	**82.30**	+19.40
	HM	76.04	**87.51**	+11.47
UCF101	Base Acc.	86.00	**86.23**	+0.23
	Novel Acc.	77.20	**80.57**	+3.37
	HM	81.36	**83.30**	+1.94

Table 11: Compatibility of PromptSRC approach using a recent V-L model: EVA CLIP [9] in the Base-to-novel generalization setting. PromptSRC shows overall favourable performance on EVA CLIP. Absolute gains over IVLP method are shown in blue.

Results of individual components: In Table 12, we show the per-dataset results for each component of our PromptSRC framework in the base-to-novel generalization setting. Our results indicate that overall, the proposed regularization components are effective in improving performance in comparison with the naive IVLP prompt learning approach.

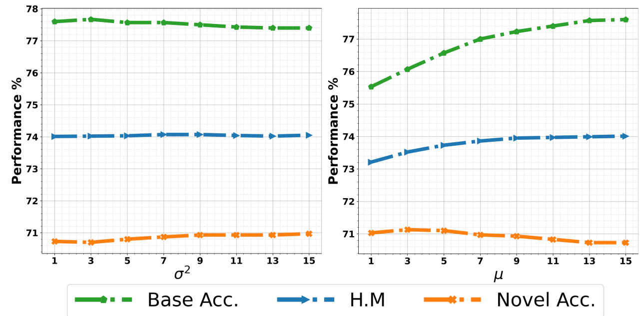

Figure 6: Ablation on GPA hyper-parameters on ImageNet.

C. Additional ablation study

On Variants of Textual diversity: Our proposed method for achieving textual diversity involves using an ensemble of frozen CLIP textual features obtained through multiple text augmentations. Here, we provide an analysis of an alternate approach for incorporating textual diversity. Instead of using an ensemble, we use a single prompt template chosen at random from N available templates to generate frozen CLIP textual features. The results averaged over 11 datasets, are shown in Table 10. However, we observe that PromptSRC with the ensembled textual diversity technique outperforms the alternate approach. This suggests that using an ensemble of frozen CLIP features encourages the learning of more diverse prompt representations.

Below, we conduct detailed ablation experiments on the ImageNet validation set to analyze the effect of GPA hyper-parameters on the final performance.

GPA hyper-parameters: We conduct ablation on μ and σ^2 hyper-parameters of GPA for the ImageNet dataset and show the results in Figure 6. Overall, varying σ^2 has a minute effect on performance. On the other hand, as we increase μ, GPA provides more weights to prompts learned in the latter epochs which increases the base class performance and slightly decreases the novel class performance.

Few-shot experiments: Table 13 shows the detailed per-dataset results of various methods in the few-shot setting. Overall, PromptSRC achieves consistant improvements over existing methods for all shots.

Dataset		IVLP	+ $\mathcal{L}_{\text{SCL}}$	+ GPA	+ Textual diversity	Δ
Average over 11 datasets	Base Acc.	84.21	84.21	84.16	84.26	+0.04
	Novel Acc.	71.79	75.38	75.69	76.10	+4.31
	H.M	77.51	79.55	79.70	79.97	+2.46
ImageNet	Base Acc.	77.00	77.53	77.47	77.60	+0.60
	Novel Acc.	66.50	69.77	70.03	70.73	+4.23
	H.M	71.37	73.45	73.56	74.01	+2.64
Caltech101	Base Acc.	98.30	98.03	97.97	98.10	-0.20
	Novel Acc.	93.20	94.37	94.67	94.03	+0.83
	H.M	95.68	96.17	96.29	96.02	+0.34
OxfordPets	Base Acc.	94.90	95.37	95.27	95.43	+0.43
	Novel Acc.	97.20	97.03	97.10	97.30	+0.10
	H.M	96.04	96.19	96.18	96.30	+0.27
StanfordCars	Base Acc.	79.53	78.87	78.03	78.27	-1.26
	Novel Acc.	71.47	74.60	74.87	74.97	+3.50
	H.M	75.28	76.68	76.42	76.58	+1.30
Flowers102	Base Acc.	97.97	97.97	98.00	98.07	+0.10
	Novel Acc.	72.10	76.90	77.10	76.50	+4.40
	H.M	83.07	86.17	86.30	85.95	+2.88
Food101	Base Acc.	89.37	90.37	90.57	90.67	+1.30
	Novel Acc.	90.30	91.23	91.47	91.53	+1.23
	H.M	89.83	90.80	91.02	91.10	+1.27
FGVCAircraft	Base Acc.	42.60	42.33	42.30	42.73	+0.13
	Novel Acc.	25.23	35.60	36.83	37.87	+12.6
	H.M	31.69	38.67	39.38	40.15	+8.46
SUN397	Base Acc.	81.60	82.53	82.57	82.67	+1.07
	Novel Acc.	75.50	78.70	78.83	78.47	+2.97
	H.M	78.43	80.57	80.66	80.52	+2.08
DTD	Base Acc.	82.40	83.13	82.97	83.37	+0.97
	Novel Acc.	56.20	61.90	62.00	62.97	+6.77
	H.M	66.82	70.96	70.97	71.75	+4.92
EuroSAT	Base Acc.	96.73	93.07	93.50	92.90	-3.83
	Novel Acc.	67.83	69.30	69.93	73.90	+6.07
	H.M	79.74	79.45	80.02	82.32	+2.58
UCF101	Base Acc.	85.93	87.10	87.07	87.10	+1.17
	Novel Acc.	74.17	79.73	79.80	78.80	+4.63
	H.M	79.62	83.25	83.28	82.74	+3.12

Table 12: Detailed performance comparison on individual datasets for showing effect of individual components in PromptSRC approach. Absolute gains of PromptSRC (IVLP + $\mathcal{L}_{\text{SCL}}$ + GPA + Textual diversity) over the IVLP are shown in blue.

Dataset	Method	1 shot	2 shots	4 shots	8 shots	16 shots
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 (Ours)	68.13	69.77	71.07	72.33	73.17
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 (Ours)	93.67	94.53	95.27	95.67	96.07
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 (Ours)	56.23	59.97	65.53	69.87	72.73
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 (Ours)	73.13	79.37	86.30	88.80	92.43
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 (Ours)	69.40	73.40	77.13	80.97	83.83
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 (Ours)	85.93	91.17	93.87	96.27	97.60
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 (Ours)	27.67	31.70	37.47	43.27	50.83
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 (Ours)	69.67	71.60	74.00	75.73	77.23
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 (Ours)	92.00	92.50	93.43	93.50	93.67
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 (Ours)	74.80	78.50	81.57	84.30	86.47
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 (Ours)	84.87	85.70	86.17	86.90	87.5
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 (Ours)	72.32	75.29	78.35	80.69	82.87

Table 13: Per-dataset performance comparison of PromptSRC with various methods in few-shot setting.