Prompt-OT: An Optimal Transport Regularization Paradigm for Knowledge Preservation in Vision-Language Model Adaptation (Supplementary Material)

Xiwen Chen[1,2*] Wenhui Zhu[3*] Peijie Qiu[4*] Hao Wang[2] Huayu Li [5]
Haiyu Wu[5] Aristeidis Sotiras[3] Yalin Wang[6] Abolfazl Razi[2]
[1] Morgan Stanley, [2] Clemson University, [3] Arizona State University,
[4] Washington University in St. Louis, [5] University of Arizona, [6] University of Notre Dame

A. Implementation Detail

We train PromptOT for 30 epochs for the Base-to-Novel Generalization benchmark and 2 epochs for the remaining three benchmark settings, respectively. The respective epochs are fixed across all datasets. All experiments are run on a node of the cluster with one V100 16GB NVIDIA GPU. Following a deep prompting strategy, we apply prompts to all transformer blocks for the first benchmark and the first three transformer blocks for the remaining benchmarks, and we use four text prompts and four vision prompts. Following previous works [9], in For Base-to-Novel Generalization benchmark, we set the first half of classes as base classes and the remaining classes as novel classes. Models are only trained on base classes. The accuracy is reported based on the evaluation of base classes and novel classes, respectively. The overview of this framework is shown in Fig. S1.

A.1. Datasets

We used the following datasets in our experiments:

Exps. 1 and 2. ImageNet [4], Caltech101 [5], OxfordPets [13], StanfordCars [10], Flowers102 [12], Food101 [1], and FGVCAircraft [11], SUN397 [17], UCF101 [15], DTD [2], and EuroSAT [6].

Exp. 3. ImageNet [4] as a source dataset and use ImageNet-A [8], ImageNet-R [7], ImageNet-Sketch [16] and ImageNetV2 [14].

A.2. Separate OT

The way of applying **separate OT** is similar to SRC loss. Suppose the adapted vision features are presented as $\boldsymbol{H} = \{\boldsymbol{h}^i\}_{i=1}^{n}$ and adapted text features are presented as $\boldsymbol{G} = \{\boldsymbol{g}^i\}_{i=1}^{C}$. Again, n and C denote the number of samples and number of classes, respectively. Similarly, we denote the zero-shot vision features and text features as $\boldsymbol{H}_{zs}$ and $\boldsymbol{H}_{zs}$, respectively. The Separate OT is denoted as:

$$\mathcal{L}_{\text{SOT}} = \mathcal{L}_{\text{OT}}(\boldsymbol{H}, \boldsymbol{H}_{zs}) + \mathcal{L}_{\text{OT}}(\boldsymbol{G}, \boldsymbol{G}_{zs}). \quad (1)$$

Table S1. Wilcoxon signed-rank test on Exp 1.

	PromptSRC VS Prompt-OT
P-value	0.016

A.3. Reproducibility

The code and model weights will be publicly released upon acceptance. **In the supplementary zip file, we provide the key code for the configuration and implementation of our proposed method.**

Our code is developed based on https://github.com/muzairkhattak/PromptSRC.

B. Wilcoxon signed-rank test

We perform a Wilcoxon signed-rank test [3] to compare our method with PromptSRC, the previous state-of-the-art, using harmonic mean accuracy as the evaluation metric. As shown in Table S1, our method achieves statistically significant improvements over PromptSRC (p-value ¡ 0.05).

C. Limitation

While our proposed PromptOT framework demonstrates strong performance across various benchmarks and introduces theoretically grounded improvements over prior prompt learning approaches, we acknowledge certain limitations that open avenues for future exploration:

First, the OT-based regularization introduces additional computational overhead compared to simpler constraints (e.g., point-wise L2 losses). Although we adopt mini-batch OT solvers to keep training efficient, the method still requires access to additional computational resources.

Second, while our approach avoids relying on external augmentations or auxiliary models (e.g., large LLMs), this

*These authors contributed equally to this paper

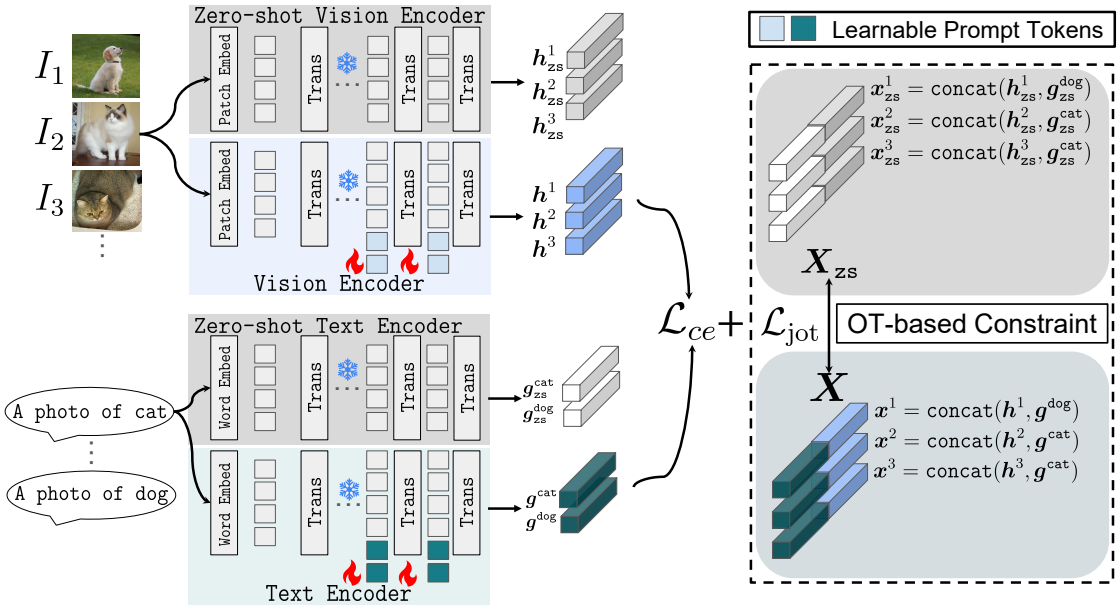

Figure S1. The overview of our proposed framework. Only the prompt tokens are trainable, and the rest of the weights in both the zero-shot encoders and the adapted encoders are frozen. Despite the cross-entropy $\mathcal{L}_{ce}$ adopted, we also adopt the proposed joint optimal transport loss $\mathcal{L}_{\text{jot}}$ between joint zero-short representation and adapted representation to constrain the model.

also means it may not fully leverage certain recent advances in generative or data-enhancement techniques. Integrating such components in a resource-efficient manner could further boost performance.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 1

[3] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, pages 178–178. IEEE, 2004. 1

[6] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *J-STARS*, 12(7):2217–2226, 2019. 1

[7] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. 1

[8] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 1

[9] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 1

[10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV*, pages 554–561, 2013. 1

[11] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE, 2008. 1

[13] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 1

[14] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 1

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[16] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 1

[17] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 1