

Supplementary Materials: Generating Prompts in Latent Space for Rehearsal-free Continual Learning

A THEORETICAL PROOF

A.1 Proof of ELBO

PROPOSITION A.1. Suppose that P, X, Y are the prompt, input data and its label respectively. The logarithmic likelihood function of classification results can be decomposed into two parts:

$$\log p(Y|X) = ELBO(q) + KL(q(P) \| p(P|X, Y)). \quad (1)$$

Proof. By expanding the logarithmic likelihood and using the conditional probability formula, we can obtain

$$\log p(Y|X) = \int_P q(P) \log p(Y|X) dP \quad (2)$$

$$= \int_P q(P) \log \left[\frac{p(P, Y|X)}{p(P|X, Y)} \right] dP \quad (3)$$

$$= \int_P q(P) \log p(P, Y|X) dP - \int_P q(P) \log p(P|X, Y) dP \quad (4)$$

$$= \int_P q(P) \log \left[\frac{p(P, Y|X)}{q(P)} \right] dP - \int_P q(P) \log \left[\frac{p(P|X, Y)}{q(P)} \right] dP \quad (5)$$

$$= \underbrace{\mathbb{E}_{q(P)} [\log p(P, Y|X)] - \mathbb{E}_{q(P)} [\log q(P)]}_{ELBO(q)} \quad (6)$$

$$- \underbrace{\{\mathbb{E}_{q(P)} [\log p(P|X, Y)] - \mathbb{E}_{q(P)} [\log q(P)]\}}_{KL(p(P|X, Y) \| q(P))} \quad (7)$$

Where Eq.(3) holds because $p(P, Y|X) = p(P|X, Y)p(Y|X)$.

A.2 Proof of Variational Loss

PROPOSITION A.2. Suppose that variational distribution $q(P)$ is a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with mean μ and covariance matrix Σ , namely

$$q(P) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp \left(-\frac{1}{2} (P - \mu)^T \Sigma^{-1} (P - \mu) \right). \quad (8)$$

Then, the expectation of logarithmic variational distribution can be expressed as

$$\begin{aligned} \mathbb{E}_{q(P)} [\log q(P)] &= -\frac{k}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| \\ &\quad - \frac{1}{2} \mathbb{E}_{q(P)} \left[(P - \mu)^T \Sigma^{-1} (P - \mu) \right]. \end{aligned} \quad (9)$$

Proof. We expand the expectation into an integral form, and obtain that

$$\mathbb{E}_{q(P)} [\log q(P)] = \int q(P) \log q(P) dP \quad (10)$$

$$= \int \mathcal{N}(\mu, \Sigma) \log \mathcal{N}(\mu, \Sigma) dP \quad (11)$$

We analyze the logarithmic terms of the normal distribution separately

$$\log \mathcal{N}(\mu, \Sigma) \quad (12)$$

$$= \log \left(\frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp \left(-\frac{1}{2} (P - \mu)^T \Sigma^{-1} (P - \mu) \right) \right) \quad (13)$$

$$= \log \left(\frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \right) + \left(-\frac{1}{2} (P - \mu)^T \Sigma^{-1} (P - \mu) \right) \quad (14)$$

$$= -\frac{k}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (P - \mu)^T \Sigma^{-1} (P - \mu) \quad (15)$$

Combing Eq.(15) and Eq.(11), we have

$$\mathbb{E}_{q(P)} [\log q(P)] \quad (16)$$

$$= \int \mathcal{N}(\mu, \Sigma) \left(-\frac{k}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma|) \right. \quad (17)$$

$$\left. - \frac{1}{2} (P - \mu)^T \Sigma^{-1} (P - \mu) \right) dP \quad (18)$$

$$= -\frac{k}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| \quad (19)$$

$$- \frac{1}{2} \mathbb{E}_{P \sim \mathcal{N}(\mu, \Sigma)} \left[(P - \mu)^T \Sigma^{-1} (P - \mu) \right] \quad (20)$$

A.3 Independent Situation

PROPOSITION A.3. If prompt P is assumed to be independent on each prompt length, namely the covariance matrix Σ of the variational distribution $q(P)$ is a diagonal matrix. Then, the expectation of logarithmic variational distribution is

$$\mathbb{E}_{q(P)} [\log q(P)] = -\frac{k}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} I. \quad (21)$$

proof. If the covariance matrix Σ is diagonal, then Σ^{-1} and $(P - \mu)$ are commutable. Thus, we have

$$\mathbb{E}_{P \sim \mathcal{N}(\mu, \Sigma)} \left[(P - \mu)^T \Sigma^{-1} (P - \mu) \right] \quad (22)$$

$$= \Sigma^{-1} \mathbb{E}_{P \sim \mathcal{N}(\mu, \Sigma)} \left[(P - \mu)^T (P - \mu) \right] \quad (23)$$

Since $\mathbb{E}_{P \sim \mathcal{N}(\mu, \Sigma)} \left[(P - \mu)^T (P - \mu) \right]$ is another form of covariance matrix, thus

$$\mathbb{E}_{P \sim \mathcal{N}(\mu, \Sigma)} \left[(P - \mu)^T \Sigma^{-1} (P - \mu) \right] = \Sigma^{-1} \Sigma = I. \quad (24)$$

Combining Eq.(9) and Eq.(24), we can obtain that

$$\mathbb{E}_{q(P)} [\log q(P)] = -\frac{k}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} I. \quad (25)$$

B INFORMATION FOR BENCHMARKS

In this paper, We conducted experiments on 5 datasets, including Split CIFAR-100 [3], Split DomainNet [6], Split Pets [5], Split EuroSAT [4] and Split CropDisease [1]. Table 1 summarizes detailed information of these datasets.

Table 1: Specifications of the various CL benchmarks evaluated in the main paper.

Dataset	Classes	Tasks	Training Set	Validation Set	Testing Set
Split CIFAR-100	100	10	4000	1000	1000
Split DomainNet	345	15	96724	24182	52041
Split Pets	35	7	2774	706	3469
Split EuroSAT	10	5	16200	5400	5400
Split CropDisease	35	7	34260	8566	10692

C ADDITIONAL EXPERIMENTS

C.1 Ablation Studies on Split DomainNet

In the main paper, we only provide the ablation experimental results on Split CIFAR-100 due to limited space. Here we provide additional results of ablation experiments on Split DomainNet in Table 2 and 3. When it comes to Split DomainNet, it is evident that removing \mathcal{L}_{nexp_Mah} and \mathcal{L}_{nlog_cov} results in the reductions of 0.97% and 0.66% on average accuracy, respectively. More importantly, the Avg Acc will decrease by 15.48% and 5.69% when Σ and ϵ are removed respectively.

Table 2: Ablation results (%) of \mathcal{L}_{nexp_Mah} and \mathcal{L}_{nlog_cov} on Split CIFAR-100 and Split DomainNet.

Method Method	Split CIFAR-100			Split DomainNet		
	Avg Acc (\uparrow)	Lrn Acc (\uparrow)	Forgetting (\downarrow)	Avg Acc (\uparrow)	Lrn Acc (\uparrow)	Forgetting (\downarrow)
GPLS	96.22 \pm 0.43	97.12 \pm 0.51	1.12 \pm 0.32	90.13 \pm 1.01	93.44 \pm 0.63	3.56 \pm 0.49
w/o \mathcal{L}_{nexp_Mah}	95.07 \pm 1.23	96.51 \pm 0.81	1.89 \pm 0.72	89.16 \pm 1.21	92.54 \pm 0.72	3.86 \pm 0.54
w/o \mathcal{L}_{nlog_cov}	95.56 \pm 1.08	96.83 \pm 0.62	1.67 \pm 0.63	89.47 \pm 1.08	92.89 \pm 0.64	3.62 \pm 0.50
w/o \mathcal{L}_{nexp_Mah} & \mathcal{L}_{nlog_cov}	94.98 \pm 1.26	96.32 \pm 2.25	2.11 \pm 1.02	88.56 \pm 1.31	92.12 \pm 0.83	4.02 \pm 0.62

Table 3: Ablation results (%) of variables for prompt generation on Split CIFAR-100 and Split DomainNet.

Method Method	Split CIFAR-100			Split DomainNet		
	Avg Acc (\uparrow)	Lrn Acc (\uparrow)	Forgetting (\downarrow)	Avg Acc (\uparrow)	Lrn Acc (\uparrow)	Forgetting (\downarrow)
GPLS	96.22 \pm 0.43	97.12 \pm 0.51	1.12 \pm 0.32	90.13 \pm 1.01	93.44 \pm 0.63	3.56 \pm 0.49
Ablate μ	95.29 \pm 0.63	96.73 \pm 0.42	1.62 \pm 0.28	89.82 \pm 0.57	92.75 \pm 0.68	3.81 \pm 0.63
Ablate Σ	68.02 \pm 4.15	81.93 \pm 3.37	15.58 \pm 2.03	74.65 \pm 3.24	85.69 \pm 2.28	11.86 \pm 1.89
Ablate ϵ	90.49 \pm 1.21	95.55 \pm 0.98	5.63 \pm 0.74	84.44 \pm 2.02	90.65 \pm 2.12	6.66 \pm 0.82

C.2 Analysis on the number of Encoder Layers

In the main paper, we adopted the one layer MLP encoder [2], which is a simple and lightweight structure. To explore the impact of encoder layers on overall performance, we also increase the number of encoder layers and the results are presented in Table 4. Experimental results show that increasing the number of layers in the encoder cannot further improve the generated prompts. On the contrary, increasing an arbitrary number of layers will result in a slight decrease in the final classification accuracy.

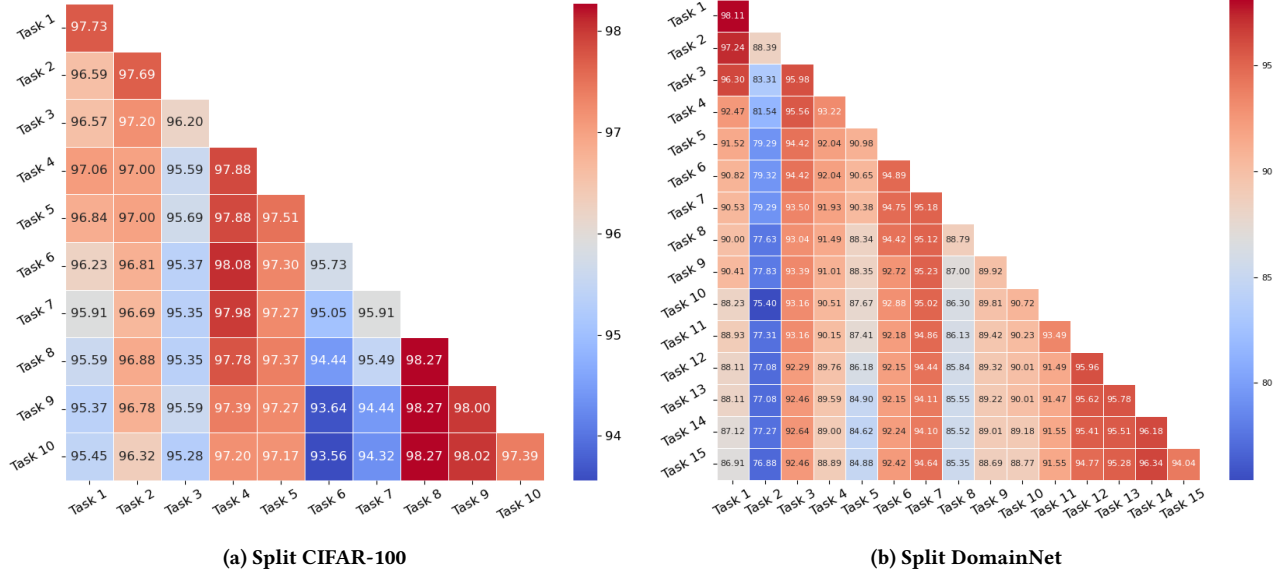
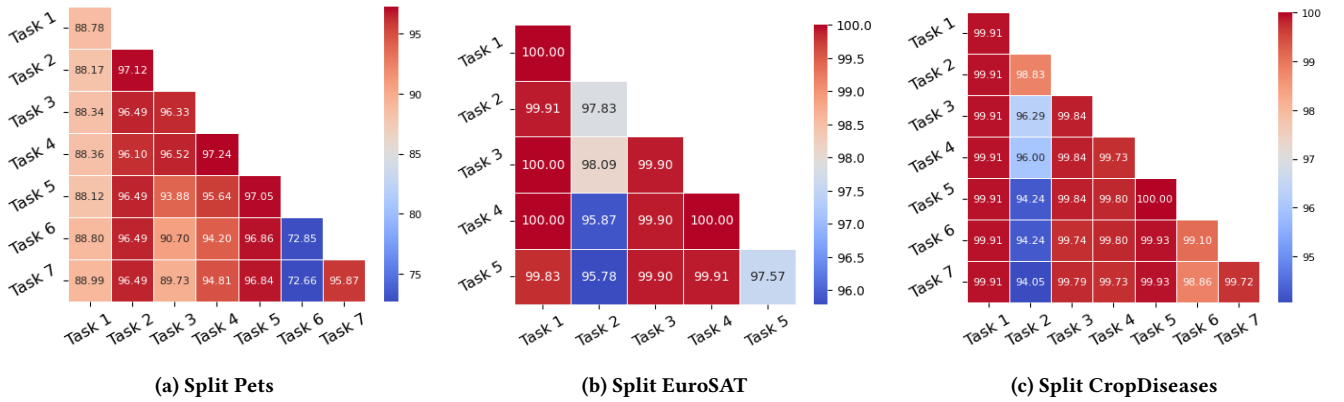
C.3 Analysis on Training Time

The training durations on a single-card A100 GPU for L2P, DualPrompt, S-Prompt++, CODA-Prompt, HiDe-Prompt, DAP and GPLS on Split CIFAR-100 dataset are 0.55, 2.00, 2.01, 2.08, 2.80, 2.14 and 2.23, respectively.

Table 4: Experimental results (%) on GPLS when adopting different encoder layers.

Method	Avg Acc (\uparrow)	Lrn Acc (\uparrow)	Forgetting (\downarrow)
GPLS	96.22 \pm 0.43	97.12 \pm 0.51	1.12 \pm 0.32
GPLS (2 layer Encoder)	95.89 \pm 0.63	96.95 \pm 0.63	1.31 \pm 0.34
GPLS (3 layer Encoder)	95.75 \pm 0.60	96.72 \pm 0.72	1.45 \pm 0.39
GPLS (4 layer Encoder)	95.58 \pm 0.57	96.64 \pm 0.68	1.63 \pm 0.55

D DETAILED ACCURACY FOR EACH TASK

**Figure 1: Detailed accuracy on Split CIFAR-100 and Split DomainNet.****Figure 2: Detailed accuracy on Split Pets, Split EuroSAT and Split CropDiseases.**

E T-SNE VISUALIZATION OF PROMPT FOR EACH TRANSFORMER LAYER

In the main paper, we present T-SNE visualization for the prompts on the first layer generated by GPLS and DAP. To comprehensively observe the differences between prompts generated by our GPLS and DAP, we provide the T-SNE visualization of prompts for all the transformer

layers (blocks) in Table 3. We have the following key observations. (1) The prompts generated by GPLS roughly converge to a straight line and exhibit approximate continuity when projected onto a low dimensional manifold. (2) The geometric characteristics of the prompt visualization under DAP closely resemble those of GPLS in the initial two layers. However, from the third layer onwards, the projection points of DAP under various tasks demonstrate a clustered distribution in the manifold space. (3) The visualization of DAP prompts reveals a disordered distribution pattern in layers 5, 7, 9, and 10.

The above observations indicate that we can obtain prompts with manifold features and better domain adaptability by utilizing the variational encoder and customized objective function .

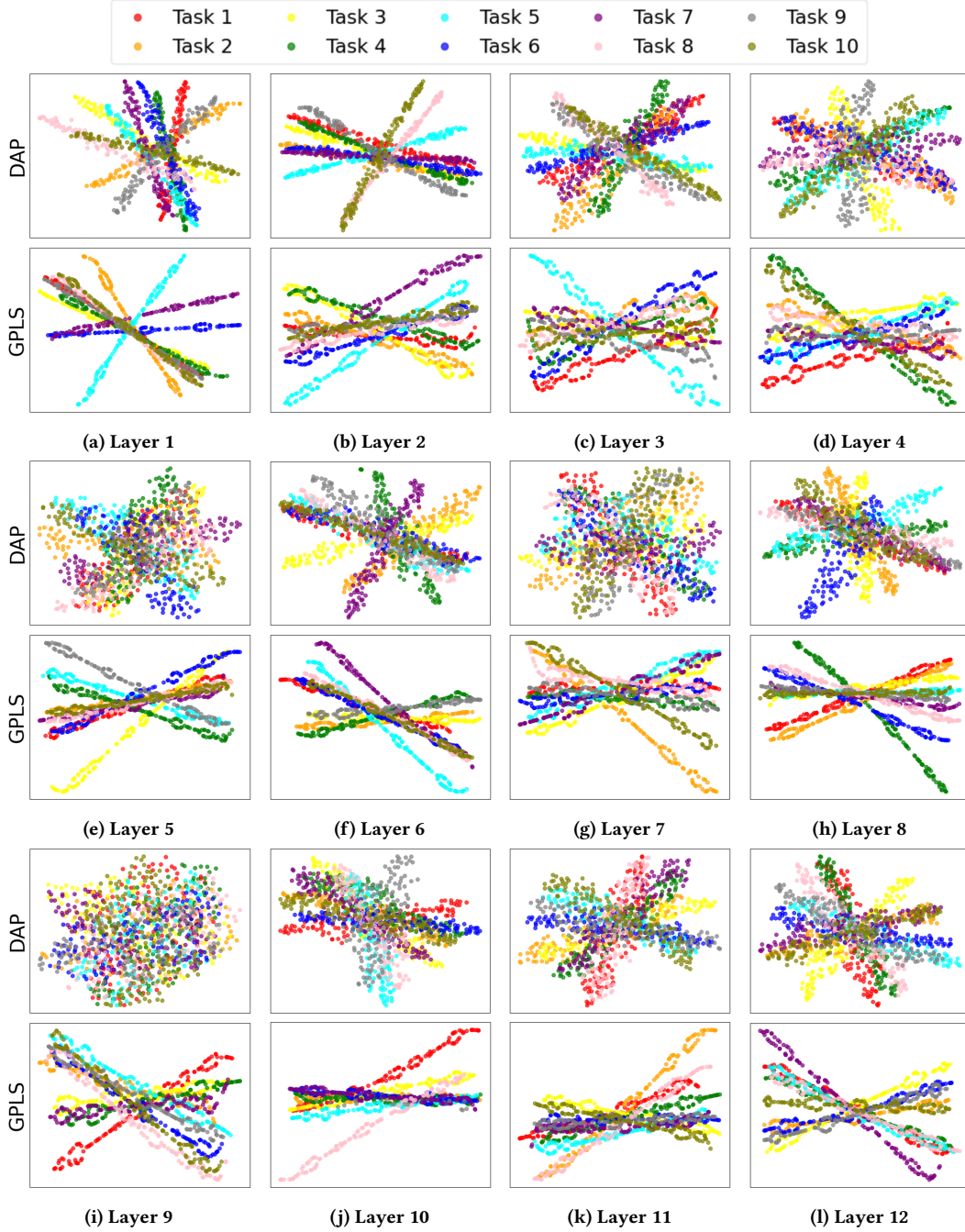


Figure 3: T-SNE visualizations of generated prompts for 10 tasks on Split CIFAR-100 for each layer.

F IMPLEMENTATION DETAILS

For Split CIFAR-100, Split Pets, Split EuroSAT and Split CropDisease, we run the experiments on one NVIDIA A100 GPU. When it comes to Split DomainNet, we run the experiments on one NVIDIA A800 GPU.

REFERENCES

- [1] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 7 (2019), 2217–2226.
- [2] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations (ICLR)*.
- [3] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. In *Technical report*. Citeseer.
- [4] Sharada P Mohanty, David P Hughes, and Marcel Salathé. 2016. Using deep learning for image-based plant disease detection. *Frontiers in plant science* 7 (2016), 215232.
- [5] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 3498–3505.
- [6] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment Matching for Multi-Source Domain Adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 1406–1415.