Dual Memory Networks: A Versatile Adaptation Approach for Vision-Language Models

Supplementary Material

The following materials are provided in this supplementary file:

- Discussion with Test-time Adaptation.
- Full results of few-shot classification (*cf*. Section 4.2 in the main paper).
- More analyses (*cf*. Section 4.3 in the main paper).

A. Discussion with Test-time Adaptation (TTA)

Our approach, especially the DMN-ZS variant, shares some high-level ideas with TTA methods [12, 51] by updating the model (*e.g*., memory) at test time. However, there are some key distinctions. First, unlike [12, 51], we leverage all historical test samples (not just the current one), improving the results by 3.77% (cf. Tab. 3). Second, we avoid test-time optimization, maintaining fast test speed (cf. Tab. 5). Third, we integrate the utilization of test and training data via flexible memory networks, extending the applicability, *e.g*., few-shot classification (cf. Tab. 1).

B. Full Results of Few-shot Classification

The full results of training-free few-shot classification and traditional few-shot classification are presented in Figures A7, A8, and A9. Similar to the observations in the main paper, our DMN consistently surpasses competing approaches in terms of average accuracy across 11 datasets, maintaining superiority with different backbone architectures and varying numbers of training samples. On individual datasets, although our method occasionally lags behind other state-of-the-art methods in certain settings (*e.g*., the Food101 dataset), it achieves consistent gains on the acknowledged ImageNet dataset, affirming its effectiveness.

C. More Analyses

Classifier Weights. We fix $\alpha_1 = 1.0$ in Eq. (12) and search for the optimal α_2 and α_3 for each downstream task. The discrete search space for α_2 and α_3 is $\{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 300\}$. The searched optimal classifier weights are shown in Tab. A6. We can observe that the value of α_2 is typically larger than that of α_3, highlighting the importance of historical test knowledge. We also find that fixing $\alpha_2 = 1.0$ and $\alpha_3 = 0.3$ can generally lead to good results in different task settings, as presented in Fig. A10.

Non-linear Function $\varphi(\cdot)$. We compare the adopted non-linear function $\varphi(x) = \exp(-\beta(1-x))$ with the popular SoftMax function, *i.e*., $\mathrm{SoftMax}(\beta \mathrm{x})$. We also search for the optimal β for the SoftMax function. As shown in Fig. A11, our strategy typically outperforms the popular SoftMax function. The possible reason for this could be that the output of SoftMax is influenced by both the value of a single element and its relative size compared to other elements. Therefore, the output of SoftMax is directly related to the memory length. In our method, the effective memory length varies due to the different shot numbers and the online update of dynamic memory, which may affect the usage of SoftMax function. In contrast, the output of our adopted $\varphi(\cdot)$ only depends on the value of a single element, making it more suitable for our task setting.

Test Data Order. By managing test data order with random seeds, we observed slight performance variations. For instance, DMN-ZS scored 72.25$\pm$0.21% on ImageNet over 3 random runs.

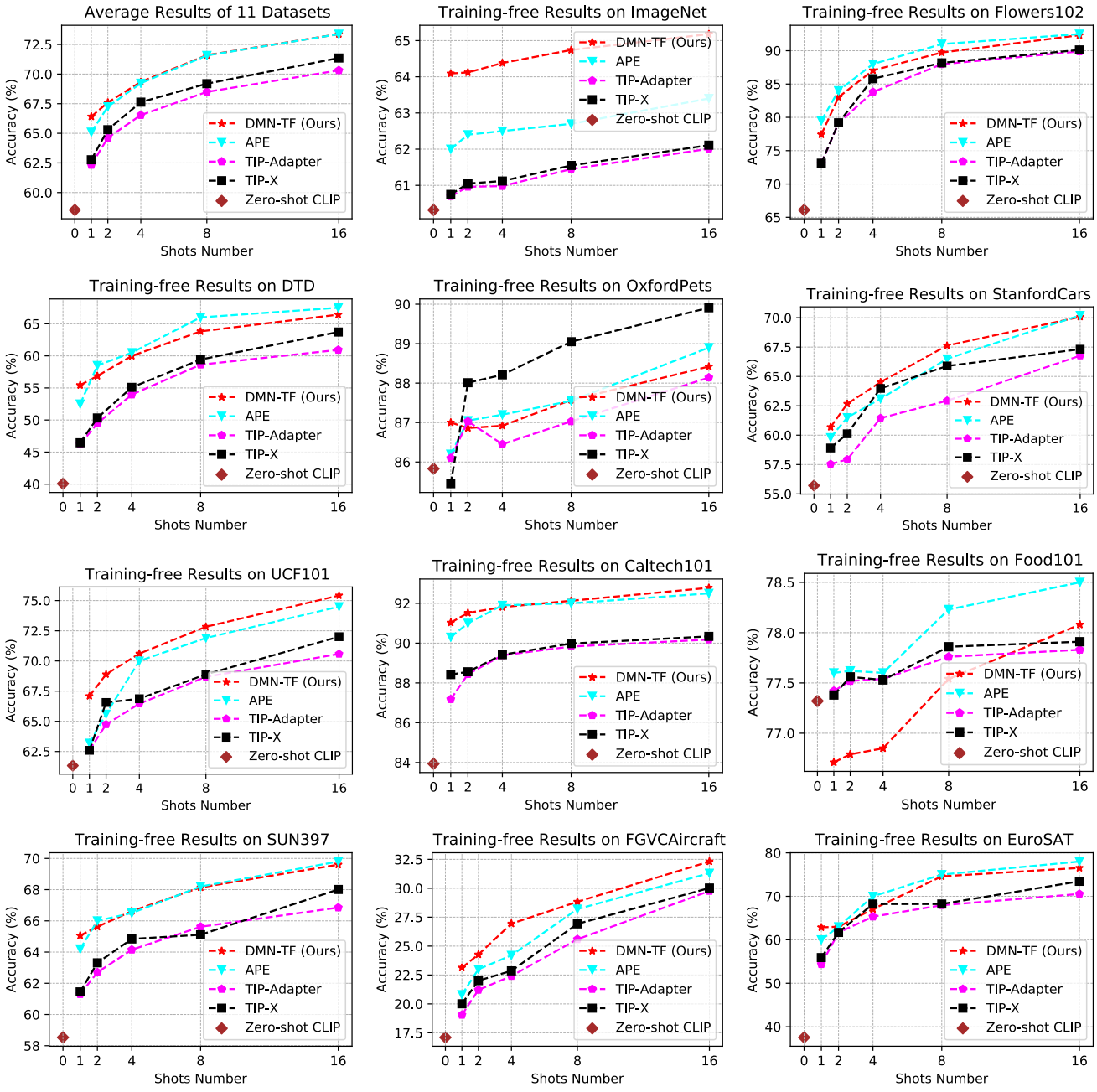

Figure A7. Training-free few-shot results of our DMN-TF and other methods on 11 classification datasets with the ResNet50 backbone.

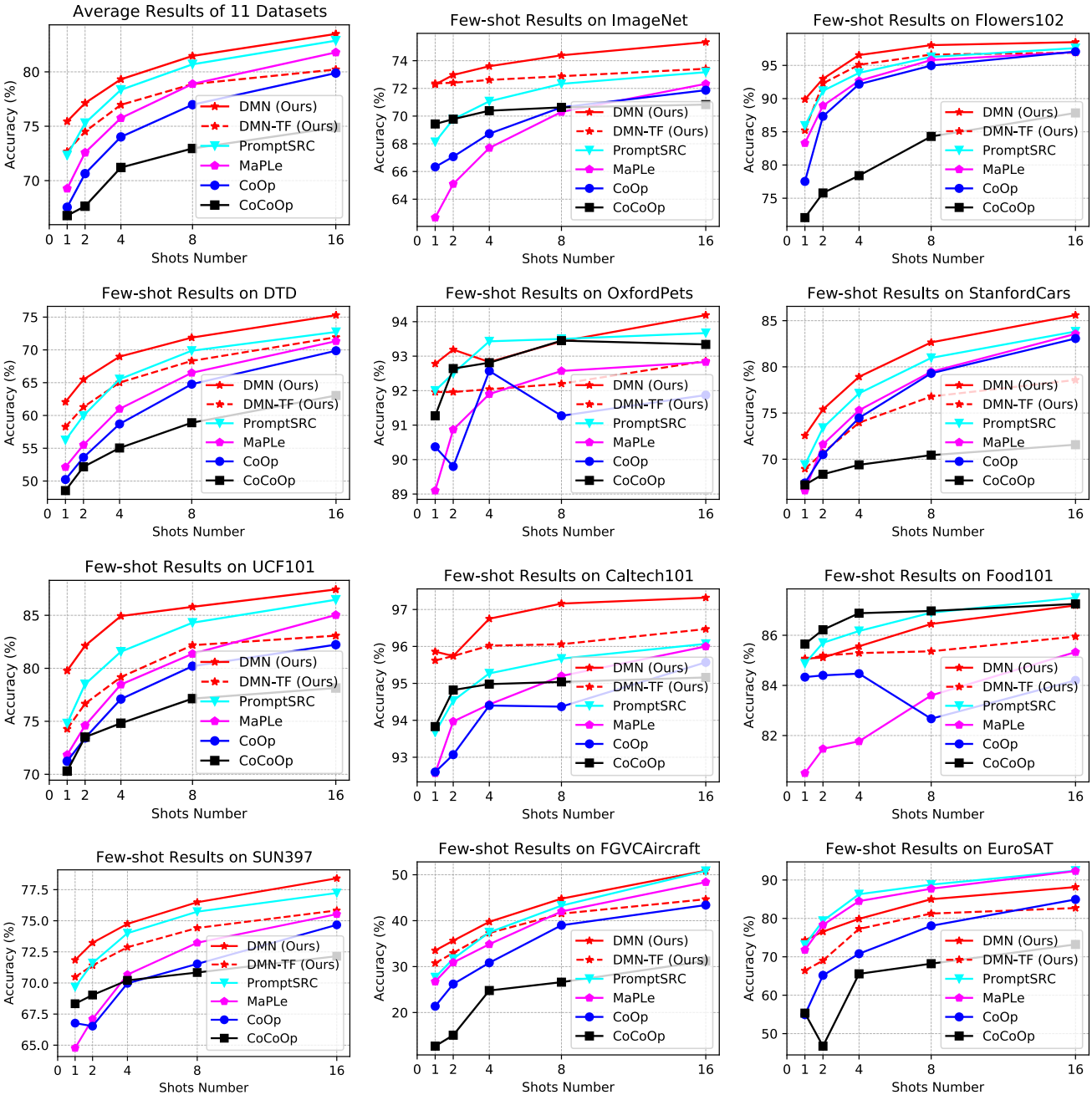

Figure A8. Few-shot results of our DMN and other methods on 11 classification datasets with the VITB/16 backbone.

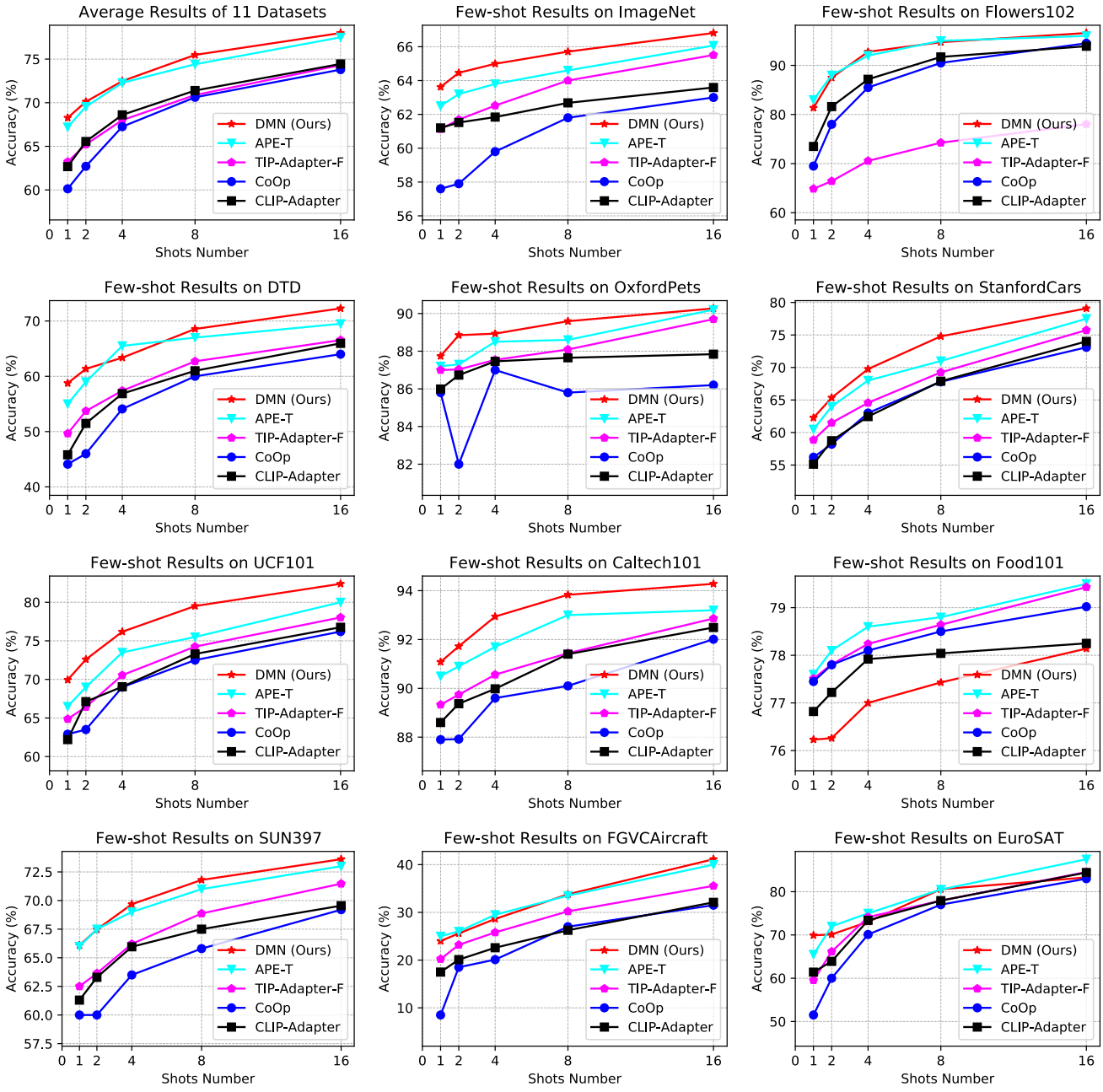

Figure A9. Few-shot results of our DMN and other methods on 11 classification datasets with the ResNet50 backbone.

Settings	Items	ImageNet	Flower	DTD	Pets	Cars	UCF	Caltech	Food	SUN	Aircraft	EuroSAT
1shot	α_2	1.0	0.3	0.3	1.0	1.0	100	0.3	0.3	3.0	0.3	1.0
	α_3	0.1	1.0	0.03	0.3	0.001	3.0	0.001	0.1	1.0	0.001	0.1
2shot	α_2	1.0	0.3	1.0	1.0	1.0	0.3	0.3	0.3	1.0	3.0	1.0
	α_3	0.3	1.0	1.0	0.001	0.03	0.03	0.3	0.001	0.3	0.3	1.0
4shot	α_2	1.0	0.3	0.3	1.0	1.0	3.0	1.0	0.3	0.3	1.0	1.0
	α_3	0.3	1.0	0.3	0.03	0.03	3.0	0.3	0.1	0.3	1.0	1.0
8shot	α_2	1.0	1.0	0.1	1.0	3.0	1.0	0.3	0.3	0.3	3.0	0.3
	α_3	0.3	1.0	0.1	0.3	0.001	0.3	0.3	0.03	0.001	3.0	1.0
16shot	α_2	1.0	3.0	0.3	1.0	3.0	1.0	1.0	0.3	1.0	0.3	0.1
	α_3	1.0	1.0	0.03	0.03	0.001	0.1	0.001	0.03	0.01	3.0	1.0

Table A6. Searched optimal classifier weights of DMN for different task settings and datasets with the VITB/16 backbone.

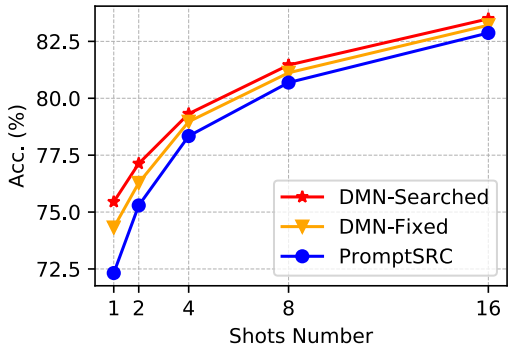

Figure A10. Average results of DMN on 11 datasets with the VITB/16 backbone. DMN-Searched and DMN-Fixed represent results with searched and fixed classifier weights, respectively. We also provide results of the recent PromptSRC method for reference.

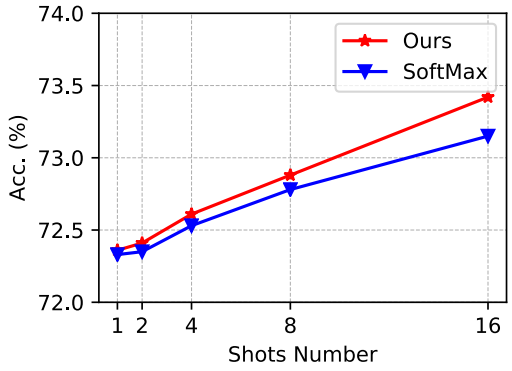

Figure A11. Results of DMN-TF with different non-linear functions on ImageNet dataset, where the VITB/16 backbone is adopted.