

750 A Algorithm

Algorithm 1 Test-Time Prompt Tuning with Multi-Scale Visual Memory (M²TPT)

```

1: Input: Test samples  $\{\mathbf{X}^t\}_{t=0}^T$ ; initial prompt  $\mathbf{p}_{\text{init}}$ ; multi-scale memory  $\mathcal{M}^{\text{mm}}$ ; holistic visual
   memory  $\mathcal{M}^{\text{hol}}$ ; class-irrelevant memory  $\mathcal{M}^{\text{irr}}$ 
2: for  $t = 0$  to  $T$  do
3:   Randomly crop the current test image  $\mathbf{X}^t$  to  $\mathbf{X}_{[N]}^t$ .
   // Memory retrieval
4:   Use  $\mathbf{X}^t$  to query  $\mathcal{M}^{\text{mm}}$  and  $\mathcal{M}^{\text{hol}}$  by Eqs. 3 and 4
5:   Fetch the retrieved memory  $\mathcal{M}_{\hat{y}}^{\text{mm}}$  or  $\mathcal{M}_{\hat{y}}^{\text{hol}}$  based on similarity.
   // Prompt tuning
6:   if Retrieved memory is  $\mathcal{M}_{\hat{y}}^{\text{mm}}$  then
7:     Optimize the prompt with the losses defined in Eqs. 5 and 11 using  $\mathcal{M}_{\hat{y}}^{\text{mm}}$ .
8:   else
9:     Optimize the prompt with the loss defined in Eq. 5 using  $\mathcal{M}_{\hat{y}}^{\text{hol}}$ .
10:  end if
   // Memory update
11:  Update  $\mathcal{M}^{\text{mm}}$  and  $\mathcal{M}^{\text{hol}}$  by Eqs. 6 and 7
12:  Selectively update  $\mathcal{M}^{\text{irr}}$  by Eq. 10
   // Prediction
13:  Yield final prediction with the optimized prompt and updated visual memory by Eq. 8
14: end for

```

751 B Error Bar Analysis

752 We conduct three runs with different random seeds, each resulting in a distinct data sample order,
 753 across 10 downstream image classification datasets. We analyze the error bars for the predictions
 754 from the adapted prompt P_{pt} , the visual memory P_{memo} , and the combined final prediction P_{final} as
 755 defined in Eq. 8. As shown in Fig. 5, both P_{pt} and P_{memo} exhibit relatively low variance across runs,
 while the combined prediction shows a slightly higher standard deviation.

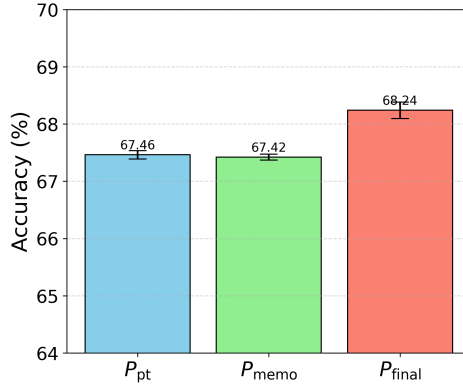


Figure 5: Results across 3 runs on 10 downstream classification datasets.

757 C Broader Impacts

758 This work contributes to the development of adaptive vision-language systems that are more re-
 759 sponsive to deployment-time data without requiring retraining or manual prompt engineering. By
 760 leveraging past visual information, our method enables more scalable adaptation in real-world sce-
 761 narios where prior knowledge of test data is unavailable. Moreover, by reducing dependence on

762 human-crafted or LLM-generated prompts, our approach lowers the barrier to deploying vision-
763 language models in new domains, benefiting users without expertise in prompt engineering. However,
764 deploying such memory-augmented models in safety-critical domains should involve strict monitoring
765 and fail-safes to prevent over-reliance on potentially noisy or outdated information.

766 **D Limitations**

767 A primary limitation of the proposed framework—shared with other prompt-tuning-based and prompt-
768 engineering-based adaptation methods for CLIP—is the assumption of a stable class vocabulary,
769 which requires that all class candidates are known in advance. In settings where new classes appear
770 dynamically, the memory structure and retrieval strategy may require further adaptation. Another
771 limitation is that memory retrieval and update depend on the quality of pseudo-labels. A high volume
772 of incorrect early predictions can lead to suboptimal memory updates, introducing significant noise
773 into the adaptation process.