Supplementary Material for Dynamic Logits Adjustment and Exploration for Test-Time Adaptation in Vision-Language Models

Supplementary Material

Algorithm A1 Dynamic Logits Adjustment (DLA)

Require: Test stream $\{x_t\}_{t=1}^{T}$, image encoder $f(\cdot)$, text prototypes $\{\mathbf{t}_c\}_{c=1}^{C}$, temperature τ, hyper-parameter α

Initialize: $n_c \leftarrow 1$, $\mu_c \leftarrow 0$ for all $c = 1, \dots, C$
for $t = 1$ **to** T **do**
  1) CLIP prediction
  $v_t \leftarrow f(x_t)$
  $s^c_{\text{clip}} \leftarrow \dfrac{v_t^\top \mathbf{t}_c}{\tau}$ for all $c = 1, \dots, C$
  $p^c_{\text{clip}} \leftarrow \text{softmax}\big(s_{\text{clip}}\big)^c$ for all $c = 1, \dots, C$
  $\hat{y}_{\text{clip}} \leftarrow \arg\max_c p^c_{\text{clip}}$; $\quad P_{\text{clip}} \leftarrow p^{\hat{y}_{\text{clip}}}_{\text{clip}}$
  2) Empirical class prior
  $N \leftarrow \sum_{c=1}^{C} n_c$
  $\hat{p}(c) \leftarrow \dfrac{n_c}{N}$ for all $c = 1, \dots, C$
  3) Logit adjustment
  $d \leftarrow P_{\text{clip}} - \mu_{\hat{y}_{\text{clip}}}$
  for $c = 1$ **to** C **do**
    $B(c) \leftarrow \exp\big(-\alpha \cdot \hat{p}(c) \cdot (1 - d)\big)$
    $s^c_{\text{DLA}} \leftarrow s^c_{\text{clip}} \cdot B(c)$
  end for
  $p^c_{\text{DLA}} \leftarrow \text{softmax}\big(s_{\text{DLA}}\big)^c$ for all $c = 1, \dots, C$
  $\hat{y}_{\text{DLA}} \leftarrow \arg\max_c p^c_{\text{DLA}}$
  4) Update statistics
  $P_{\text{DLA}} \leftarrow p^{\hat{y}_{\text{DLA}}}_{\text{DLA}}$
  $\mu_{\hat{y}_{\text{DLA}}} \leftarrow \dfrac{\mu_{\hat{y}_{\text{DLA}}}\big(n_{\hat{y}_{\text{DLA}}} - 1\big) + P_{\text{DLA}}}{n_{\hat{y}_{\text{DLA}}}}$
  $n_{\hat{y}_{\text{DLA}}} \leftarrow n_{\hat{y}_{\text{DLA}}} + 1$
  $h_{\text{DLA}} \leftarrow -\sum_{c=1}^{C} p^c_{\text{DLA}} \log p^c_{\text{DLA}}$
end for

Algorithm A2 Consistency-Guided Exploratory Cache (CGEC)

Require: Sample tuple $(\mathbf{f}_v, h_{\text{DLA}}, i_{\text{now}})$, predictions $\hat{y}_{\text{clip}}, \hat{y}_{\text{DLA}}$, text prototypes $\{\mathbf{t}_c[0], \mathbf{t}_c[\text{now}]\}_{c=1}^{C}$, class-specific caches $\{\mathcal{P}_c\}_{c=1}^{C}$ (each stores $(\mathbf{f}'_v, h', i')$), capacity K, hyper-parameters β, η

$c \leftarrow \hat{y}_{\text{DLA}}, \quad h \leftarrow h_{\text{DLA}}$
1) Flip check + Semantic Consistency Filter (SCF)
if $\hat{y}_{\text{clip}} \neq \hat{y}_{\text{DLA}}$ **then**
  $h \leftarrow h \cdot \exp\Big(-\beta \cos\big(\mathbf{t}_{\hat{y}_{\text{clip}}}[\text{now}], \mathbf{t}_{\hat{y}_{\text{DLA}}}[\text{now}]\big)\Big)$
end if
2) Temporal Consistency Filter (TCF) on cached entries
for each $(\mathbf{f}'_v, h', i') \in \mathcal{P}_c$ **do**
  if $\mathbf{t}_c^\top[0]\, \mathbf{f}'_v > \mathbf{t}_c^\top[\text{now}]\, \mathbf{f}'_v$ **then**
    $h' \leftarrow h' \cdot \exp\big(\eta(i_{\text{now}} - i')\big)$
    $i' \leftarrow i_{\text{now}}$
  end if
end for
3) Priority queue (capacity K)
Define a new candidate triple $(\mathbf{f}'_v, h', i') \leftarrow (\mathbf{f}_v, h, i_{\text{now}})$
if $|\mathcal{P}_c| < K$ **then**
  Insert $(\mathbf{f}'_v, h', i')$ into $\mathcal{P}_c$
else
  $e_{\text{worst}} \leftarrow \arg\max_{(\tilde{\mathbf{f}}_v, \tilde{h}, \tilde{i}) \in \mathcal{P}_c} \tilde{h}$
  if $h' < h(e_{\text{worst}})$ **then**
    Replace e_{worst} with $(\mathbf{f}'_v, h', i')$
  end if
end if

A. Detailed Algorithms

We provide the complete test-time pseudo-code of our framework. Alg. A1 presents Dynamic Logits Adjustment (DLA), which maintains lightweight per-class statistics to recalibrate logits on-the-fly. Alg. A2 presents Consistency-Guided Exploratory Cache (CGEC), which updates per-class caches using semantic and temporal consistency. All symbols follow the definitions in the main paper (e.g., $\mathbf{t}_c[0]$ and $\mathbf{t}_c[\text{now}]$).

A. Plug-and-Play with Advanced Retrieval-Based Methods

We demonstrate that DLAE can be plugged into retrieval-based TTA methods that aggregate multiple candidates (e.g., ReTA [2], MCP++ [1]) without modifying their architectures or optimization pipelines.

A.1. Plug-and-Play Integration

For retrieval-based TTA methods such as ReTA [2] and MCP++ [1], we augment their text-image prediction and cache mechanisms with our DLAE, while keeping their original architectures and training pipelines completely unchanged. DLAE recalibrates the text-image logits during both training and inference via Dynamic Logits Adjustment, and employs the Consistency-Guided Exploratory Cache to store informative boundary samples that yield an additional loss $\mathcal{L}_{\text{conf}}^{\text{DLAE}}$ for optimizing textual representations. All other components of the base method remain unchanged.

Table A1. Comprehensive evaluation on cross-dataset generalization with advanced retrieval-based methods that do not rely solely on top-1 outputs. Top: standalone performance of ReTA [2] and MCP++ [1]. Bottom: plug-and-play enhancement after applying our framework (**Ours**). Higher is better (%).

Method	Caltech	DTD	Cars	EuroSAT	Aircraft	Flowers	Pets	UCF101	Food101	SUN397	Average
CLIP-RN50	85.88	40.37	55.70	23.69	15.66	61.75	83.57	58.84	73.97	58.80	55.82
ReTA [2]	90.35	52.46	61.11	39.64	22.62	**70.12**	86.90	66.18	77.83	65.11	63.23
ReTA + Ours	**90.83**	**54.26**	**61.83**	**45.00**	**23.18**	70.03	**87.23**	**67.00**	**78.02**	**65.21**	**64.26**
MCP++ [1]	91.13	53.61	61.76	55.74	23.40	**69.96**	87.49	67.86	78.44	65.55	65.49
MCP++ + Ours	**91.16**	**55.04**	**62.08**	**57.77**	**23.58**	69.83	**88.20**	**68.00**	**78.82**	**65.65**	**66.01**
CLIP-ViT-B/16	93.35	44.27	65.48	42.01	23.67	67.44	88.25	65.13	83.65	62.59	63.58
ReTA [2]	95.29	57.39	69.11	58.26	31.86	77.55	92.37	74.52	86.69	70.70	71.37
ReTA + Ours	**95.35**	**58.92**	**69.86**	**59.10**	**32.23**	**77.60**	**92.45**	**75.03**	**87.10**	**70.93**	**71.85**
MCP++ [1]	95.50	56.97	**70.13**	68.69	31.06	77.55	**92.40**	75.44	87.20	71.17	72.61
MCP++ + Ours	**95.74**	**57.35**	69.98	**69.34**	**32.04**	**78.13**	92.15	**77.04**	**87.21**	71.17	**73.02**

Table A2. Runtime and accuracy on ImageNet with ViT-B/16, evaluated on a single 24GB NVIDIA RTX 3090 GPU.

Method	Testing Time	Acc. (%)	Gain (%)
CLIP	9 min	66.73	–
TPT	10 h	68.98	+2.25
DPE	3 h 42 min	71.91	+5.18
DLAE (Ours)	**4 h 03 min**	**72.41**	**+5.68**

Table A3. Mis-caching rate of CGEC-admitted (consistency-passing) samples vs. DPE-cached samples. Lower is better.

Dataset	**CGEC**	**DPE**	**Gain**
DTD	31.72%	35.26%	3.54%
UCF101	2.31%	21.40%	19.09%
FGVC	20.65%	65.58%	44.93%

A.2. Effectiveness

As shown in Tab. A1, plugging DLAE into ReTA [2] and MCP++ [1] consistently improves cross-dataset generalization across backbones, indicating that DLAE complements retrieval-based aggregation without architectural modification.

A. Runtime Analysis

Compared with cache-based methods that require backpropagation (e.g., DPE), our method introduces only minor overhead, mainly from extra logit computation and lightweight cache-admission checks. Since CLIP inference and test-time backpropagation dominate the overall runtime, the additional cost is modest and the wall-clock time remains close to DPE, while achieving higher accuracy (Tab. A2).

As discussed in the main paper, our CGEC admission rule is designed to favor consistent samples that are also more likely to be correct. Concretely, SCF filters out samples with large semantic gaps, while TCF removes samples whose temporal evolution conflicts with the text-guided adaptation direction. To validate these motivations, we report the mis-caching rate in Tab. A3, defined as the fraction of incorrect samples among those admitted into the cache (i.e., passing both SCF and TCF) and thus mistakenly retained for subsequent learning. Compared with a standard cache baseline (DPE), CGEC-admitted (consistency-passing) samples consistently yield a lower mis-caching rate across datasets, confirming that our criteria reduce cache noise while preserving informative boundary evidence.

References

[1] Xinyu Chen, Haotian Zhai, Can Zhang, Xiupeng Shi, and Ruirui Li. Multi-cache enhanced prototype learning for test-time generalization of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2281–2291, 2025. 1, 2

[2] Yiwen Liang, Hui Chen, Yizhe Xiong, Zihan Zhou, Mengyao Lyu, Zijia Lin, Shuaicheng Niu, Sicheng Zhao, Jungong Han, and Guiguang Ding. Advancing reliable test-time adaptation of vision-language models under visual variations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4788–4797, 2025. 1, 2