Complementary Cache Guidance with Gradient Disentanglement for Continuous Test-Time Adaptation

Fanchun Meng[1,2*] **Yuhang Pei**[1*] **Jiazhen Huang**[2] **Tao Ren**[1] **Yifan Wang**[3†]
Wei Ju[4] **Xiao Luo**[5]
[1]Northeastern University, China
[2]Tsinghua University, China
[3]University of International Business and Economics, China
[4]Peking University, China
[5]University of Wisconsin–Madison, USA

Abstract

Continuous test-time adaptation (CTTA) is crucial for deploying vision-language models (VLMs) in real-world environments where data distributions shift over time. Existing approaches typically utilize confidence-driven pseudo-labeling to enhance the performance on the target domain while retraining VLMs using historical information. Despite the progress, their performance is far from satisfactory due to error accumulation from noisy pseudo-labels and optimization interference between newly acquired and historical knowledge. Towards this end, we propose a novel approach named Complementary Cache Guidance with Gradient Disentanglement (CURE) for CTTA of VLMs. The core of our CURE is to balance instant knowledge acquisition and historical knowledge consolidation using both complementary cache systems and parameter space disentanglement. In particular, our CURE first constructs affinity structures among diverse text prompts, which guide majority voting to improve the quality of pseudo-labeling. More importantly, our CURE introduces a complementary cache system consisting of a short-term cache and a long-term cache with sample quality monitoring. The short-term cache stores recent reliable samples with low reweighted entropy, while the long-term cache preserves representative samples with high gradient consistency across iterations. Then, we extract prototypes from both caches for cross-modal alignment, enabling the model to acquire new knowledge while mitigating old knowledge forgetting. To further reduce the interference when learning from two caches, we perform low-rank decomposition of the gradient space, which facilitates historical knowledge consolidation in the null space of short-term signals. Extensive experiments on several benchmarks demonstrate the superiority of CURE over state-of-the-art baselines. Our source code is available at https://github.com/Fan-Chun-Meng/CURE.

1 Introduction

Recent advances in multi-modal learning have increasingly relied on pretrained vision-language models (VLMs) [21, 51], e.g., CLIP [39], which exploit extensive image-text paired data to acquire semantically rich representations that generalize effectively across a wide range of downstream tasks. However, when deployed in real-world applications, these models are often exposed to distribution shifts that were not present during training, which can significantly degrade their performance. To address this real-time challenge, previous efforts have utilized downstream data to align pretrained

*Fanchun Meng and Yuhang Pei contributed equally.
†Correspondence to: Yifan Wang (yifanwang@uibe.edu.cn)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

models with the target distribution, either through domain-specific fine-tuning [61, 13] or prompt tuning [66, 65, 2]. While effective in some scenarios, these methods often require labeled target data, limiting their flexibility and scalability in dynamic environments where data arrives continuously and labels are unavailable.

Test-time adaptation (TTA) [28], which enables model adaptation directly during inference with unlabeled test samples, has emerged as a promising paradigm for handling distribution shifts in VLMs. Existing TTA methods include self-training with pseudo-labels [29, 52, 43] and self-supervised methods that incorporate auxiliary tasks [30, 26], regularization [4, 48], batch normalization [27, 32] and entropy minimization [46, 57] strategies. For VLMs, prompt-based approaches learn domain-specific prompts using a self-supervised objective, i.e., entropy minimization [41, 11, 1]. As a result, these methods require backpropagation through the text encoder at each test-time update, which leads to substantial computational overhead and may induce catastrophic forgetting of previously acquired patterns. Instead, cache-based approaches maintain high-confidence test samples in the constructed cache for knowledge accumulation and computational efficiency [22, 56, 63, 31], enabling on-the-fly prediction calibration during inference with improved efficiency and adaptation stability.

Despite significant progress, we argue that adapting VLMs for continually changing test environments remains challenging due to the following key issues: ❶ *Noise and miscalibrated pseudo-labels under continually changing distributions.* Unlike static test distributions, real-world test sample distributions are dynamically changing in an online and non-stationary manner. As these distributions continually shift, the predicted pseudo-labels can become noisy and miscalibrated. Such errors are subsequently accumulated during the adaptation process, ultimately leading to degraded model performance [59]. ❷ *Parameter interference during the adaptation process.* As shown in Figure 1, when VLMs adapt to new distributions, updating model parameters to accommodate novel patterns may overlap or conflict with previously learned parameters, which will result in severe catastrophic forgetting.

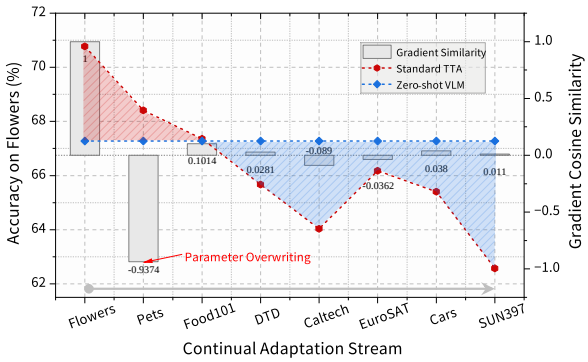

Figure 1: Catastrophic forgetting during continuous test-time adaptation. The curves show accuracy on the initial Flowers102 domain as adaptation proceeds across the domains on the x-axis. The blue dashed line represents frozen zero-shot CLIP-ViT-B/16, while the red dashed line represents a standard TTA baseline. Gray bars indicate gradient cosine similarity between updates from the current and initial domains.

Towards this end, in this paper, we propose a Complementary Cache Guidance with Gradient Disentanglement framework (termed CURE) for continuous test-time adaptation of VLMs, which aims to address the effects of noisy pseudo-labels and parameter interference under the continually changing environment. Specifically, when test samples arrive in a stream, we introduce a set of class-specific prompts and construct the local textual representation as the voting committee to identify reliable samples. The selected reliable samples are then maintained in complementary cache systems, where the short-term cache stores recently arrived samples, and the long-term cache retains reliable samples over a longer temporal horizon. To further prevent parameter interference during continuous adaptation, we introduce a disentangled low-rank space to track the gradient online. On the one hand, we constrain parameter updates to a low-rank principal subspace of the gradients to maintain robustness over short-term adaptation. On the other hand, for long-term historical knowledge consolidation, we introduce a null-space constraint that restricts parameter updates within a task-orthogonal subspace. Additionally, we introduce an adaptive optimization strategy to progressively relax the constraint during test-time adaptation.

In a nutshell, the contributions of this paper are as follows: ❶ ***New Perspective.*** We highlight the parameter interference for continuous test-time adaptation of VLMs and introduce a space disentanglement view to study the problem. ❷ ***Novel Methodology.*** We propose CURE, which comprises the short- and long-term cache systems to maintain reliable samples at different temporal scales and tracks online gradients within a disentangled low-rank subspace to enable distinct parameter updates guided by the two cache systems. ❸ ***State-of-the-art Performance.*** We conduct extensive

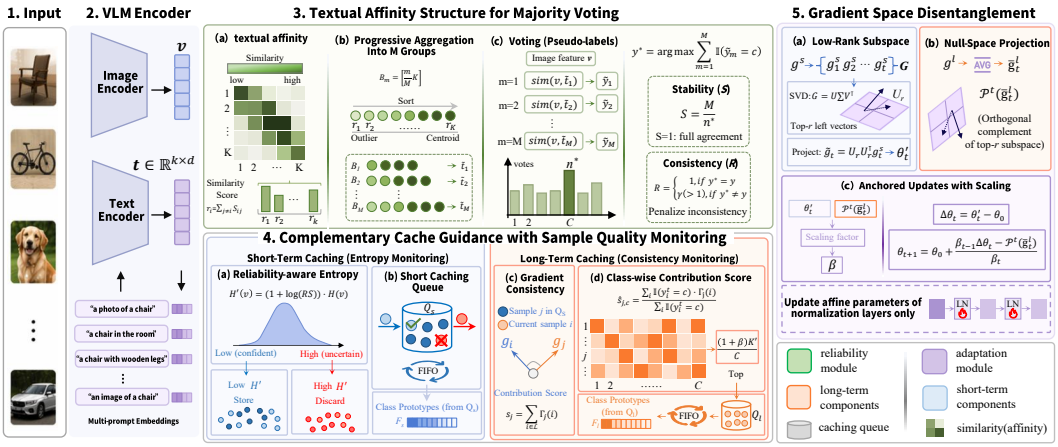

Figure 2: The overall framework of our CURE. Given the streaming test samples, we construct the affinity structure among text prompts as a voting committee to estimate the prediction reliability, then maintain a complementary cache system to capture recent adaptation dynamics and stable historical knowledge. The cached representations further guide low-rank subspace disentanglement.

experiments on a diverse set of benchmark datasets, and the results demonstrate that our CURE consistently achieves better performance compared with baselines.

2 Preliminaries & Problem Definition

Zero-Shot VLM. VLMs are pretrained on large-scale image-text pairs to align visual and textual representations in a shared semantic space [14]. A VLM comprises two parallel encoders, namely an image encoder $\mathcal{F}_v(\cdot)$ and a text encoder $\mathcal{F}_t(\cdot)$. For a classification task with C classes, each class $c \in \{1, \dots, C\}$ is represented by a text prompt $\boldsymbol{p}_c$ (e.g., "a photo of a [CLASS]."). For an input image , the VLM obtains zero-shot predictions by measuring the similarity between its visual representation and the text embeddings of all C candidate classes, as follows:

$$\boldsymbol{v} = \mathcal{F}_v(\boldsymbol{x}), \boldsymbol{t}_c = \mathcal{F}_t(\boldsymbol{p}_c), \mathbb{P}_{\text{VLM}}(y = c|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{v}, \boldsymbol{t}_c)/\tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(\boldsymbol{v}, \boldsymbol{t}_{c'})/\tau)}, \tag{1}$$

where $\boldsymbol{v}, \boldsymbol{t}_c \in \mathbb{R}^d$ are extracted features in a unified embedding space with dimension d, $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity and τ is the temperature parameter.

Test-Time Adaptation for VLMs. Given a pretrained VLM, the output prediction can be significantly distorted when test data undergo distribution shifts w.r.t. training domain [12]. Test-time adaptation aims to mitigate this issue by updating the model during inference using only unlabeled test samples [19, 68]. In existing VLM adaptation methods, a common strategy is to construct N augmented views $\{\mathcal{A}_n(\boldsymbol{x})\}_{n=1}^{N}$ for each test sample $\boldsymbol{x}$ and derive the confident prediction based on the entropy threshold τ as:

$$\mathbb{P}_{\text{TPT}}(\boldsymbol{x}) = \frac{1}{\rho N} \sum_{n=1}^{N} \mathbb{I}(H(\mathbb{P}_{\text{VLM}}(\mathcal{A}_n(\boldsymbol{x}))) \leq \tau) \mathbb{P}_{\text{VLM}}(\mathcal{A}_n(\boldsymbol{x})), \tag{2}$$

where $H(\cdot)$ denotes the self-entropy of the prediction. This will enhance the text prompt by minimizing the self-entropy of the top-ρ percentile confident prediction, i.e., $\min H(\mathbb{P}_{\text{TPT}}(\boldsymbol{x}))$. Recently, cache-based methods maintain high-confidence samples in a queue $\boldsymbol{Q} \in \mathbb{R}^{K' \times C \times d}$ to handle streaming test samples, where K' is the max number of samples per class. The test sample serves as a query to retrieve relevant auxiliary information from the cache to improve the prediction:

$$\mathbb{P}_Q(\boldsymbol{v}) = \mathcal{A}(\boldsymbol{v}\boldsymbol{Q}^{\text{T}})\boldsymbol{Y}_Q, \tag{3}$$

where $\mathcal{A}(\boldsymbol{z}) = \alpha \exp(-\beta(1 - \boldsymbol{z}))$ denotes an affinity function with scaling factor α and sharpness ratio β. Here, $\boldsymbol{Y}_Q$ denotes the one-hot pseudo-label matrix for the cache. During inference, the cache-based prediction is integrated with the current prediction as $\mathbb{P}_{\text{cls}} = \boldsymbol{v}\boldsymbol{W}_{\text{VLM}}^{\text{T}} + \mathbb{P}_Q(\boldsymbol{v})$. In practice, test samples often arrive as a stream with evolving distributions, and continuous test-time adaptation aims to update the VLM online to maintain robust performance under distribution shifts.

3 The Proposed CURE

3.1 Framework Overview

In this paper, we study continuous test-time adaptation of VLMs and propose CURE, a framework that handles distribution shifts over different time scales through complementary caching and gradient-space disentanglement. Specifically, CURE integrates three connected mechanisms. First, neighbor-aware textual voting assesses pseudo-label reliability and determines which samples enter the short-term cache. Second, the short-term cache captures recent reliable observations, while the long-term cache retains reliable samples with limited contribution to recent updates, preserving complementary historical evidence. Third, short-term gradients are projected onto a principal low-rank subspace, while long-term corrections are constrained to its orthogonal complement to mitigate interference between adaptation signals. These mechanisms jointly coordinate reliability assessment, temporal memory, and parameter updates, balancing short-term plasticity with long-term stability. An overview of CURE is illustrated in Figure 2, with further details presented below.

3.2 Textual Affinity Structure for Majority Voting

Since entropy-based confidence is often unreliable for VLMs due to overconfident predictions [9, 23, 31], we construct a textual affinity graph and leverage adjacent text embeddings as a semantic voting committee, which evaluates prediction consistency to identify reliable predictions. In particular, for each class c, we construct multiple text prompts as $\boldsymbol{\mathcal{P}}^c = \{\boldsymbol{\mathcal{P}}_1^c, \ldots, \boldsymbol{\mathcal{P}}_K^c\}$, where $\boldsymbol{\mathcal{P}}_i^c$ denotes the i-th prompt and the corresponding feature can be extracted via the text encoder as $\boldsymbol{t}_i^c = \mathcal{F}_t(\boldsymbol{\mathcal{P}}_i^c)$. Instead of simply averaging the text features for the prediction, we exploit the subtle differences by calculating their semantic affinity in the textual space as:

$$r_i^c = \sum_{j=1, j\neq i}^{K} \mathrm{sim}(\boldsymbol{t}_i^c, \boldsymbol{t}_j^c), \tag{4}$$

where r_i^c denotes the intra-class similarity of the i-th prompt. We then sort the text features in ascending order according to the similarity and construct the neighbor-aware text embedding by progressively aggregating them from semantic outliers toward the class centroids. Specifically, we define the bin size as $B_m = \lfloor \frac{mK}{M} \rfloor$ for $m \in \{1, 2, \ldots, M\}$ and update each text embedding $\tilde{\boldsymbol{t}}_m^c$ by averaging the first B_m text features from the sorted sequence. Given the neighbor-aware text embedding sets $\tilde{\mathcal{T}} = \{\tilde{\boldsymbol{t}}_m^c | c \in [1, \ldots, C], m \in [1, \ldots, M]\}$, the pseudo-label set can be defined as:

$$\tilde{\mathcal{Y}} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_M\},\ \tilde{y}_m = \underset{c\in[1,\ldots,C]}{\arg\max}\, \mathrm{sim}(\boldsymbol{v}, \tilde{\boldsymbol{t}}_m^c). \tag{5}$$

We treat this set of pseudo-label predictions as a voting committee and define the most frequent label as the majority-voted pseudo-label as $y^* = \arg\max_{c\in[1,\ldots,C]} \sum_{m=1}^{M} \mathbb{I}(\tilde{y}_m = c)$. Based on this, we next quantify the stability of the majority-voted pseudo-label and its consistency with the original prediction, which can be defined as:

$$S = \frac{M}{n^*},\ R = \begin{cases} 1, & \text{if } y^* = y \\ \gamma, & \text{if } y^* \neq y \end{cases}, \tag{6}$$

where S and R denote the stability and consistency score of the prediction, respectively. Here n^* denotes the frequency of the majority prediction. A larger value of S reflects lower prediction stability, whereas $S = 1$ indicates full agreement. Moreover, $\gamma > 1$ is a penalty coefficient that amplifies the entropy term, thereby downweighting inconsistent predictions during cache system updating.

3.3 Complementary Cache Guidance with Sample Quality Monitoring

To accommodate dynamically evolving test streams, we maintain two complementary cache systems to balance instant and historical knowledge, where the short-term cache captures recent local dynamics and the long-term cache preserves stable historical knowledge. The two cache systems are updated using different criteria, so as to balance adaptation plasticity and long-term robustness.

Short-term Caching with Entropy Monitoring. Since relying solely on entropy may overestimate unreliable samples [59], we introduce a consistency-stability score to reweight cache samples instead

of relying solely on entropy for cache updating, which can be defined as:

$$H'(\boldsymbol{v}) = (1 + \log(RS)) \cdot H(\boldsymbol{v}), \tag{7}$$

where $H(\cdot)$ denotes the entropy function. The term $(1 + \log(RS))$ serves as an adaptive scaling factor that calibrates entropy according to the reliability of the pseudo-label. When the prediction is unstable or inconsistent, the term further amplifies the entropy score. As a result, we retain reliable samples with lower reweighted entropy in the short-term cache $\boldsymbol{Q}_s \in \mathbb{R}^{K' \times C \times d}$ according to the pseudo label and replace unreliable samples with larger reweighted entropy when the cache is full.

Long-term Caching with Consistency Monitoring. To avoid overfitting to recent local dynamics and reduce error accumulation in continuously evolving test streams, we maintain a long-term cache that preserves stable class-level feature representations over time. Specifically, we extract sample embeddings using the short-term adapted parameters and measure their feature consistency across successive updates. Note that the training loss relies exclusively on the current high confident sample. Accordingly, for each sample j within short cache, we estimate the first-order Taylor expansion of the loss on the current test sample i after the update at step $(t + 1)$ as follows:

$$\mathcal{L}(i, \boldsymbol{\theta}_{t+1}) \approx \mathcal{L}(i, \boldsymbol{\theta}_t) - \eta_t \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(i, \boldsymbol{\theta}_t)^\top \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(j, \boldsymbol{\theta}_t), \tag{8}$$

where $\boldsymbol{\theta}_t$ and η_t are the parameter and learning rate at step t. $\nabla_{\boldsymbol{\theta}_t}(\cdot)$ denotes the gradient of the loss function w.r.t. $\boldsymbol{\theta}_t$. The change in loss across optimization steps reflects the contribution of the sample within short-term cache. Thus, we define the contribution score function of j, namely $\Gamma_j(\cdot)$, and aggregate the contribution score over multiple streaming test samples $\mathcal{I}$ as:

$$s_j = \sum_{i \in \mathcal{I}} \Gamma_j(i) = \sum_{i \in \mathcal{I}} \eta_t \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(i, \boldsymbol{\theta}_t)^\top \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(j, \boldsymbol{\theta}_t). \tag{9}$$

When the contribution score is negative, the corresponding sample is considered detrimental to the current update. In practice, we employ a class-balanced strategy to select samples from the short-term cache to update the long-term cache. The class-wise contribution score matrix defined as:

$$\hat{s}_{jc} = \frac{\sum_i \mathbb{I}(y_i^t = c) \cdot \Gamma_j(i)}{\sum_i \mathbb{I}(y_i^t = c)}, \tag{10}$$

where $\hat{s}_{jc}$ denotes the contribution score of node j under class c. Assume that total of K' samples are selected for each class c, we choose $\frac{(1+\beta)K'}{C}$ samples with the lowest score to update the long-term cache $\boldsymbol{Q}_l \in \mathbb{R}^{K' \times C \times d}$. Since these samples have passed the reliability filter, low contribution does not imply low reliability. Instead, it indicates weak or negative alignment with recent updates, allowing the long-term cache to preserve complementary historical evidence. The subsequent null-space projection further prevents interference with the principal short-term directions. Similarly, we replace the sample with a lower score when the cache is full.

3.4 Gradient Space Disentanglement for Interference Mitigation

Given the constructed two complementary cache systems, we explicitly disentangle the gradient update space into the low-rank subspace and the null space to avoid the parameter interference during the adaptation process. In particular, the overall loss function consists of an entropy and an alignment loss to encourage alignment between textual and visual representations, which can be defined as:

$$\begin{aligned} \mathcal{L}^* &= \mathcal{L}_{\text{ent}} + \lambda \mathcal{L}^*_{\text{align}},\ \mathcal{L}_{\text{ent}} = H(\mathbb{P}_{\text{TPT}}(\boldsymbol{x})), \\ \mathcal{L}^*_{\text{align}} &= \frac{1}{C} \sum_{c=1}^{C} \left(-\log \frac{\exp((\tilde{\boldsymbol{t}}_M^c)^\top \boldsymbol{F}_*^c)}{\sum_{c'=1}^{C} \exp((\tilde{\boldsymbol{t}}_M^c)^\top \boldsymbol{F}_*^{c'})} - \log \frac{\exp((\tilde{\boldsymbol{t}}_M^c)^\top \boldsymbol{F}_*^c)}{\sum_{c'=1}^{C} \exp((\tilde{\boldsymbol{t}}_M^{c'})^\top \boldsymbol{F}_*^c)} \right), * \in \{s, l\} \end{aligned} \tag{11}$$

where $\boldsymbol{F}_*^c$ denotes the average prototype from the short and long-term cache. To enable robust continual adaptation under evolving test streams while maintaining efficiency, we restrict the updates to the affine parameters of the normalization layers instead of the entire model.

Low-rank Constrained Adaptation for Instant Knowledge Acquisition. Given a sequence of recent test batches, we collect the short-term adaptation gradient as $\boldsymbol{g}_t^s = \nabla_{\boldsymbol{\theta}_t} \mathcal{L}^s(i, \boldsymbol{\theta}_t)$. Let $\boldsymbol{G} = [\boldsymbol{g}_{t-t_s+1}^s, \ldots, \boldsymbol{g}_t^s] \in \mathbb{R}^{n \times t_s}$ denote the gradient over the latest t_s sequential batches, we extract the r-rank subspace of the gradient matrix for $r < t_s \ll n$ as:

$$\max_{\boldsymbol{U}_r \in \mathbb{R}^{n \times r}} \operatorname{tr}(\boldsymbol{U}_r^\top \boldsymbol{G} \boldsymbol{G}^\top \boldsymbol{U}_r),\ \text{s.t.}\ \boldsymbol{U}_r^\top \boldsymbol{U}_r = \boldsymbol{I}_{\boldsymbol{r}}, \tag{12}$$

where $\boldsymbol{U}_r \in \mathbb{R}^{n\times r}$ denote the orthonormal basis of the top-r principal subspace. In practice, we apply a truncated Singular Value Decomposition (SVD), namely $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, and retain the top-$r$ left singular vectors as $\boldsymbol{U}_r$. We then project the current gradient $\boldsymbol{g}_t$ onto this subspace and map it back to the original parameter space, and the model parameter can be updated as:

$$\boldsymbol{\theta}'_t = \boldsymbol{\theta}_t - \eta_t \tilde{\boldsymbol{g}}^s_t,\ \tilde{\boldsymbol{g}}_t = \boldsymbol{U}_r \boldsymbol{U}_r^\top \boldsymbol{g}^s_t. \tag{13}$$

Note that the update process is restricted to the principal gradient directions, which preserve the dominant signals while filtering out less correlated gradient components for short-term adaptation.

Null Space Adaptation for Historical Knowledge Consolidation. Since parameter updates from different short-term adaptations may conflict with each other, the model can gradually forget previously acquired knowledge. To mitigate this issue, we project the parameter updates onto the null space [44, 38], i.e., the orthogonal complement of previously learned directions, for long-term adaptation, which can be defined as:

$$\mathcal{P}^t(\bar{\boldsymbol{g}}^l_t) = (I - \boldsymbol{U}_r \boldsymbol{U}_r^\top)\bar{\boldsymbol{g}}^l_t, \tag{14}$$

where $\bar{\boldsymbol{g}}^l_t$ denotes the average differences between model parameters among t_l latest sequential batches. $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the i-th and j-th columns for left singular vectors $\boldsymbol{U}$ and right singular vectors $\boldsymbol{V}$, respectively. The parameter update can be:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_0 + \frac{\beta_{t-1}\Delta\boldsymbol{\theta}_t - \mathcal{P}^t(\bar{\Delta}\boldsymbol{\theta}^l_t)}{\beta_t},\ \Delta\boldsymbol{\theta}_t = \boldsymbol{\theta}'_t - \boldsymbol{\theta}_0,\ \bar{\boldsymbol{g}}^l_t = \frac{1}{t_l}\sum_{t=t-t_l+1}^{t} \boldsymbol{g}^l_t, \tag{15}$$

where $\boldsymbol{g}^l_t = \nabla_{\boldsymbol{\theta}_t}\mathcal{L}^l(i, \boldsymbol{\theta}_t)$ is the gradient for the long-term adaptation loss, $\boldsymbol{\theta}_0$ denotes the original parameter of the pretrained model. And we introduce the scaling factor β_t to adaptively maintain a consistent magnitude of the updated parameter:

$$\beta_1 = 1,\ \beta_t = \frac{\|\beta_{t-1}\Delta\boldsymbol{\theta}_{t-1} - \mathcal{P}^{t-1}(\bar{\boldsymbol{g}}^l_{t-1})\|_2}{\mathrm{Avg}(\|\Delta\boldsymbol{\theta}_t\|_2)}, \tag{16}$$

where $\mathrm{Avg}(\|\cdot\|)$ denotes the average norm. So in this way, we can ensure the Frobenius norm $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2$ remains stable for long-term adaptation process.

Theoretical Justification of Space Disentanglement. Let the updated parameter be $\theta \in \mathbb{R}^n$, and define the short-term and long-term gradients as $g^s_t := \nabla_{\theta_t}\mathcal{L}^s(i_t, \theta_t)$ and $g^l_t := \nabla_{\theta_t}\mathcal{L}^l(i_t, \theta_t)$. For the latest t_s streaming steps, let $G_t := [g^s_{t-t_s+1}, \ldots, g^s_t] \in \mathbb{R}^{n\times t_s}$, with top-$r$ left singular vectors $U_r \in \mathbb{R}^{n\times r}$. Define the orthogonal projectors $\Pi_t := U_r U_r^\top$ and $\Pi_t^\perp := I - \Pi_t$. We condition on the reliable short-term and long-term cache systems produced above, and regard the corresponding prototypes as fixed within one local update window. Under this conditioning, $\mathcal{L}^s(i, \theta)$ and $\mathcal{L}^l(i, \theta)$ are well-defined objectives with respect to θ. Next, we theoretically demonstrate that the low-rank subspace disentanglement effectively separates the short-term plasticity and long-term robustness during continual parameter update.

Assumption 3.1. *$\mathcal{L}^s(i, \theta)$ and $\mathcal{L}^l(i, \theta)$ are continuously differentiable and L-smooth in a neighborhood containing the adaptation trajectory, i.e.,*

$$\mathcal{L}^\star(i, \theta') \le \mathcal{L}^\star(i, \theta) + \langle \nabla_\theta \mathcal{L}^\star(i, \theta), \theta' - \theta\rangle + \frac{L}{2}\|\theta' - \theta\|_2^2, \qquad \star \in \{s, l\}. \tag{17}$$

Proposition 3.1. *Among all rank-r orthogonal projectors, Π_t preserves the maximum accumulated energy of recent short-term gradients:*

$$\Pi_t \in \arg\max_{\Pi=\Pi^\top=\Pi^2,\ \mathrm{rank}(\Pi)=r} \mathrm{tr}(\Pi G_t G_t^\top). \tag{18}$$

Proposition 3.2. *If $\theta'_t = \theta_t - \eta_t \Pi_t g^s_t$, then under Assumption 3.1, we have*

$$\mathcal{L}^s(i_t, \theta'_t) \le \mathcal{L}^s(i_t, \theta_t) - \eta_t \|\Pi_t g^s_t\|_2^2 + \frac{L\eta_t^2}{2}\|\Pi_t g^s_t\|_2^2. \tag{19}$$

When $\eta_t \le 1/\mathcal{L}$, the projected update yields a descent:

$$\mathcal{L}^s(i_t, \theta'_t) \le \mathcal{L}^s(i_t, \theta_t) - \frac{\eta_t}{2}\|\Pi_t g^s_t\|_2^2. \tag{20}$$

Method	Flo.	DTD	Pets	Cars	UCF	Cal.	Food	SUN	Air.	Euro.	Avg.	Rank
CLIP-ViT-B/16	67.28	44.44	88.06	65.28	65.03	92.94	83.82	62.59	23.82	41.38	63.46	15.7
CoOp	68.71	41.92	89.14	64.51	66.55	93.70	85.30	64.15	18.47	46.39	63.88	14.1
CoCoOp	70.85	45.45	90.46	64.90	68.44	93.79	83.97	66.89	22.29	39.23	64.63	12.5
MaPLe	72.23	46.49	90.49	65.57	68.69	93.53	86.20	67.01	24.74	48.06	66.30	9.0
TPT	68.98	47.75	87.79	68.50	68.04	94.16	84.67	65.50	24.78	42.44	65.26	10.8
DiffTPT	70.10	47.00	88.22	67.01	62.67	92.49	87.23	65.74	25.60	43.13	64.92	11.2
PromptAlign	72.39	47.24	90.76	68.50	69.47	94.01	86.65	68.50	24.80	47.86	67.02	6.4
ZERO	67.07	45.80	86.74	67.54	67.64	93.51	84.36	64.49	24.40	39.60	64.12	13.7
MTA	68.26	45.59	88.22	68.05	68.11	94.13	84.95	64.98	25.32	38.71	64.63	11.9
GS-Bias	71.94	46.10	90.38	67.33	67.59	94.60	86.09	67.40	26.49	52.42	67.03	8.1
TCA	73.33	46.16	89.53	65.33	**72.38**	93.69	85.31	65.92	24.87	**70.43**	68.70	8.1
TDA	71.42	47.40	88.63	67.28	70.66	94.24	86.14	67.62	23.91	58.00	67.53	8.0
DPE	75.07	54.20	91.14	67.31	70.44	94.81	86.17	70.07	28.95	55.79	69.40	3.9
BoostAdapter	71.66	45.69	89.51	**69.30**	71.93	94.77	87.17	68.09	27.45	61.22	68.68	5.1
HisTPT	71.20	48.90	89.10	69.20	70.10	94.50	**89.30**	67.20	26.90	49.70	67.61	6.1
BCA	73.12	53.49	90.43	66.86	67.59	94.69	85.97	68.41	28.59	56.63	68.58	6.5
CURE (Ours)	**75.48**	**54.67**	**91.20**	67.49	72.14	**94.93**	86.70	**70.49**	**29.40**	58.84	**70.13**	**2.2**

Table 1: Comparison of CURE in Cross-Datasets Generalization using ViT-B/16 as the backbone. Bold values indicate the best performance. **Abbreviations**: Flo. (Flowers102), Cal. (Caltech101), Food (Food101), SUN (SUN397), Air. (FGVC Aircraft), Euro. (EuroSAT), UCF (UCF101).

Proposition 3.3. *Let* $\bar{g}_t^l := \frac{1}{t_l} \sum_{\tau=t-t_l+1}^{t} g_\tau^l$ *be the averaged long-term gradient, and define the null-space correction* $\delta_t^l := \Pi_t^\perp \bar{g}_t^l$. *Then*

$$\langle u, \delta_t^l \rangle = 0, \qquad \forall u \in \mathrm{span}(U_r). \tag{21}$$

In particular, since $\Pi_t g_t^s \in \mathrm{span}(U_r)$,

$$\langle \Pi_t g_t^s, \delta_t^l \rangle = 0, \tag{22}$$

and therefore

$$\|\Pi_t g_t^s + \delta_t^l\|_2^2 = \|\Pi_t g_t^s\|_2^2 + \|\delta_t^l\|_2^2. \tag{23}$$

Thus, the long-term correction does not overlap with the short-term principal adaptation directions in the first-order update space. Proposition 3.1 establishes that the rank-r orthogonal projector Π_t maximizes the accumulated energy of recent short-term gradients. Proposition 3.1 guarantees that the corresponding projected gradient $\Pi_t g_t^s$ remains a valid descent direction. Proposition 3.3 further shows that the long-term correction $\Pi_t^\perp \bar{g}_t^l$ lies in the orthogonal complement of this principal short-term subspace. Therefore, the parameter update is decomposed into a principal short-term component that promotes plasticity and an orthogonal long-term component that mitigates interference, which is exactly the desired disentanglement of CURE. Proofs are provided in Appendix E.1 and E.2.

4 Experiment

4.1 Experimental Settings

Datasets. We evaluate CURE on two widely used TTA benchmarks: Cross-Dataset Generalization and Domain Generalization. For Cross-Dataset Generalization, we use 10 diverse image recognition datasets: Flowers102 [34], OxfordPets [36], StanfordCars [25], FGVC-Aircraft [33], Food101 [3], EuroSAT [16], UCF101 [42], DTD [5], SUN397 [50], and Caltech101 [10]. To assess robustness against domain shifts, we use four out-of-distribution (OOD) variants of ImageNet [6]: ImageNetV2 [40], ImageNet-Sketch [47], ImageNet-A [18], and ImageNet-R [17].

Baselines. We compare CURE with state-of-the-art vision-language adaptation approaches, categorized into three groups: (1) *Zero-Shot CLIP* [39]; (2) *Training-time adaptation*, including CoOp [66], CoCoOp [65], and MaPLe [24]; and (3) *Test-Time Adaptation (TTA)*. The TTA baselines include seven instance-wise methods (TPT [41], DiffTPT [11], PromptAlign [1], ZERO [9], MTA [55], GS-Bias [20], TCA [49]) and five episodic methods (TDA [22], BoostAdapter [62], HisTPT [58], DPE [56], BCA [67]). Following standard protocols, training-time methods are pretrained on ImageNet (16 shots per class), whereas TTA methods operate without access to source training data.

Implementation Details. We build CURE upon the pretrained CLIP model [39]. We evaluate two visual encoder backbones, ViT-B/16 [7] and ResNet-50 [15], alongside a Transformer [45] text

Method	Venue	IN	IN-A	IN-V2	IN-R	IN-S	Avg.	OOD Avg.	Rank
CLIP-ViT-B/16	ICML2021	66.73	47.87	60.86	73.98	46.09	59.11	57.20	14.0
CoOp	IJCV2022	71.51	49.71	64.20	75.21	47.99	61.72	59.28	9.4
CoCoOp	CVPR2022	71.02	50.63	64.07	76.18	48.75	62.13	59.91	9.1
MaPLe	CVPR2023	70.72	50.90	64.07	76.98	49.15	62.36	60.28	8.5
TPT	NIPS2022	68.98	54.77	63.45	77.06	47.94	62.44	60.81	10.8
DiffTPT	ICCV2023	70.30	55.68	65.10	75.00	46.80	62.58	60.65	8.8
ZERO	NIPS2024	69.06	**61.35**	64.13	77.28	48.29	64.02	62.76	7.6
MTA	CVPR2024	70.08	58.06	64.24	78.33	49.61	64.06	62.56	6.8
GS-Bias	ICML2025	70.57	56.61	64.62	80.49	50.33	64.52	63.01	5.3
TCA	ICCV2025	68.88	50.13	62.10	77.11	48.95	61.43	59.57	10.8
TDA	CVPR2024	69.51	60.11	64.67	80.24	50.54	65.01	63.89	5.6
DPE	NIPS2024	71.91	59.63	65.44	80.40	52.26	65.93	64.43	3.0
BCA	CVPR2025	70.22	61.14	64.90	**80.72**	50.87	65.57	64.41	3.6
CURE (Ours)	-	**72.15**	60.40	**65.72**	80.49	**52.49**	**66.25**	**64.78**	**1.7**

Table 2: Comparison of CURE on ImageNet, and ImageNet variants using ViT-B/16 as the backbone. Bold values indicate the best performance. **Abbreviations**: IN (ImageNet), IN-A (ImageNet-A), IN-V2 (ImageNet-V2), IN-R (ImageNet-R), IN-S (ImageNet-S).

encoder. To construct the neighbor-aware textual affinity graph, our prompt set $\mathcal{P}^c$ is built using a multi-prompt formulation that combines dataset-specific hand-crafted templates and class-specific CuPL prompts [37]. Other detailed hyperparameters are provided in Appendix F.2.

4.2 Comparisons with State-of-the-art

Results on the Cross-Domain Benchmark. As shown in Table 1, CURE achieves a new state-of-the-art average accuracy of 70.13% with the ViT-B/16 backbone, while ResNet-50 results are provided in Appendix G.1. It surpasses MaPLe by 3.83 percentage points and DPE by 0.73 percentage points, outperforming DPE on all ten datasets. Compared with instance-wise TTA methods such as TPT, CURE benefits from neighbor-aware reliability modeling, which reduces the impact of overconfident noisy pseudo-labels. Moreover, the proposed null-space projection mitigates error accumulation during continual updates, contributing to the best average rank of 2.2 across all datasets.

Results on the OOD Benchmark. Table 2 reports the robustness results on ImageNet and its four out-of-distribution variants. CURE obtains the highest average accuracy of 66.25% with ViT-B/16, outperforming recent competitive methods such as BCA and DPE. The gains are particularly meaningful under severe shifts such as ImageNet-A and ImageNet-R, where unconstrained adaptation methods are more vulnerable to noisy updates. By combining principal subspace adaptation with null-space interference mitigation, CURE maintains stable performance across diverse distribution shifts and achieves the best average rank.

4.3 Ablation study

To evaluate the contribution of each component in CURE, we conduct an ablation study in Figure 3. Removing the long-term caching, denoted as *w/o Long Caching*, causes the largest overall performance drop of 2.86%, highlighting the necessity of the dual-caching architecture for balancing short-term plasticity and long-term stability. The results also show that bypassing the geometric update constraints substantially degrades performance. Removing the null-space projection, denoted as *w/o Null-Space Proj.*, leads to a 2.54% drop, indicating its importance in mitigating catastrophic forgetting. Removing the low-rank projection, denoted as *w/o Low-Rank Proj.*, causes a 2.42% drop, suggesting that constraining updates to principal adaptation directions helps suppress transient batch noise. Finally, removing the neighbor-aware reliability module, denoted as *w/o Neighbor Rel.*, results in a 2.45% drop, confirming that semantic voting is essential for reducing the impact of overconfident pseudo-labels and preventing error accumulation. By integrating these complementary modules, the full CURE framework achieves an overall accuracy of 67.05%, with a 7.13% absolute improvement over the Zero-Shot CLIP baseline.

4.4 Analysis and Insights

Parameter Sensitivity Analysis. We evaluate the sensitivity of CURE to key hyper-parameters, as illustrated in Figure 4. For input reliability modeling, the consistency penalty γ peaks at $\gamma = 1.2$,

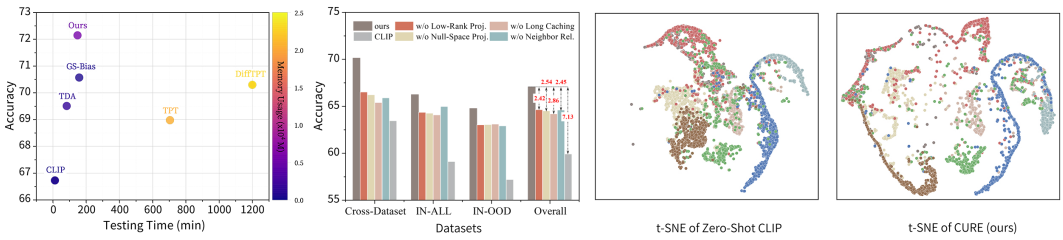

Figure 3: Comprehensive analysis of CURE. Left: testing time, accuracy, and memory usage comparison. Middle: accuracy drops caused by module removal. Right: t-SNE visualization on EuroSAT.

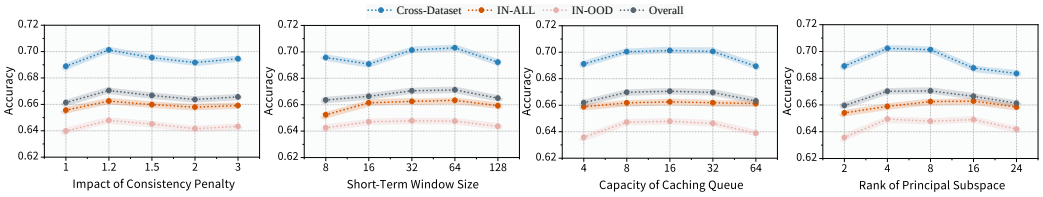

Figure 4: Parameter sensitivity analysis of CURE (ViT-B/16). The curves track top-1 accuracy variations across four evaluation groups.

confirming that moderate penalization effectively suppresses overconfident noisy pseudo-labels compared to the unpenalized baseline ($\gamma = 1.0$). Regarding the dual memory capacity, the overall accuracy reaches its optimum at $K = 16$. This demonstrates that an appropriate queue size adequately captures class-wise feature distributions, whereas excessively large queues degrade robustness by introducing outdated historical samples. Furthermore, for the geometric subspace updates, the short-term window size achieves the best performance at $t_s = 64$, balancing the responsiveness to rapid distribution shifts against transient noise suppression. Spatially, the principal subspace rank peaks at $r = 8$; this confirms that a properly sized adaptation subspace preserves sufficient plasticity, while overly expanding it restricts the complementary null-space, thereby weakening the geometric protection required to mitigate long-term interference. Additional experimental results and comprehensive parameter evaluations utilizing the ResNet-50 backbone are provided in the Appendix G.3.

Efficiency and Effectiveness. Figure 3 compares computational cost and adaptation performance. Optimization-based methods such as TPT and DiffTPT are accurate but expensive, requiring extensive augmentation and back-propagation, while cache-based methods such as TDA are efficient but less generalizable. CURE achieves the best accuracy of 72.15% (+5.42% over Zero-Shot CLIP) with moderate resource consumption, indicating a favorable efficiency-effectiveness trade-off for continuous streaming adaptation.

Visualization Analysis. Figure 3 visualizes the EuroSAT feature space using t-SNE. Zero-Shot CLIP produces scattered representations with evident inter-class overlap, whereas CURE forms more compact and separable clusters after online adaptation. This indicates that the proposed dual-memory prototypes and alignment objective help learn more discriminative decision boundaries under distribution shifts without target annotations.

5 Conclusion

In this paper, we propose CURE, a memory-enhanced framework designed to address error accumulation and parameter interference in continuous test-time adaptation. Our method introduces a space disentanglement view that decouples the adaptation process into short-term plasticity and long-term robustness. By constructing a local textual voting committee coupled with a dual memory bank, CURE reliably identifies and maintains target-domain knowledge from continuously changing test streams. Furthermore, by restricting parameter updates within disentangled low-rank and task-orthogonal null subspaces, the framework rapidly adapts to novel patterns while safeguarding previously learned representations against catastrophic forgetting. Experimental results across cross-domain and out-of-distribution benchmarks demonstrate that CURE achieves state-of-the-art performance, providing a flexible and scalable solution for deploying vision-language models in dynamic environments.

Acknowledgements

Tao Ren is supported by the National Natural Science Foundation of China (62276058) and the Fundamental Research Funds for the Central Universities (N25GFZ011). Yifan Wang is supported by the Fundamental Research Funds for the Central Universities in UIBE (Grant No. 23QN02), the Humanities and Social Sciences Research Fund of the Ministry of Education of China under Grant 25YJCZH275 and Open Research Fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems.

Competing Interests. The authors declare no competing interests.

References

[1] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 80396–80413, 2023.

[2] Jihwan Bang, Sumyeong Ahn, and Jae-Gil Lee. Active prompt learning in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27004–27014, 2024.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision.*, pages 446–461, Cham, Switzerland, 2014. Springer.

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, Piscataway, NJ, USA, 2014. IEEE.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, Piscataway, NJ, USA, 2009. IEEE.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, Virtual Event, 2021. OpenReview.net.

[8] Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

[9] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Proceedings of the Conference on Neural Information Processing Systems*, 37:129062–129093, 2024.

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 178–178, Piscataway, NJ, USA, 2004. IEEE.

[11] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023.

[12] Yiwei Fu, Hui Wan, Xiao Luo, and Minghua Deng. Star: Test-time adaptation can enhance universal prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31482–31492, 2026.

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.*, 132(2):581–595, 2024.

[14] Yiyang Gu, Jiayue Fan, Kaili Liu, Bohan Wu, Binqi Chen, Zequn Liu, Zhiping Xiao, Rong-Cheng Tu, Xiao Luo, and Ming Zhang. Evirag: Evidence-guided retrieval-augmented generation for medical vision-language models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3753–3758, 2026.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, Piscataway, NJ, USA, 2016. IEEE.

[16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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):2217–2226, 2019.

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8320–8329, Piscataway, NJ, USA, 2021. IEEE.

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, Piscataway, NJ, USA, 2021. IEEE.

[19] Jiazhen Huang, Zhiming Liu, Changhu Wang, Wei Ju, Ziyue Qiao, and Xiao Luo. Von mises-Fisher mixture model with dynamic shrinkage for realistic test-time transduction. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 47178–47211. PMLR, 06–11 Jul 2026.

[20] Zhaohong Huang, Yuxin Zhang, Jingjing Xie, Fei Chao, and Rongrong Ji. Gs-bias: Global-spatial bias learner for single-image test-time adaptation of vision-language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, Proceedings of Machine Learning Research, 2025. PMLR.

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, pages 4904–4916, 2021.

[22] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.

[23] Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10854–10863, 2024.

[24] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, Piscataway, NJ, USA, 2023. IEEE.

[25] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 554–561, Piscataway, NJ, USA, 2013. IEEE.

[26] Jogendra Nath Kundu, Suvaansh Bhambri, Akshay Kulkarni, Hiran Sarkar, Varun Jampani, and R Venkatesh Babu. Concurrent subsidiary supervision for unsupervised source-free domain adaptation. In *Proceedings of the European Conference on Computer Vision.*, pages 177–194. Springer, 2022.

[27] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

[28] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *Int. J. Comput. Vis.*, 133(1):31–64, 2025.

[29] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[30] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8602–8617, 2021.

[31] 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 ACM International Conference on Multimedia*, pages 4788–4797, 2025.

[32] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In *Proceedings of the International Conference on Learning Representations*, 2023.

[33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013.

[34] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729, Piscataway, NJ, USA, 2008. IEEE.

[35] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[36] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, Piscataway, NJ, USA, 2012. IEEE.

[37] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15645–15655. IEEE, 2023.

[38] Zihuan Qiu, Yi Xu, Chiyuan He, Fanman Meng, Linfeng Xu, Qingbo Wu, and Hongliang Li. Mingle: Mixture of null-space gated low-rank experts for test-time continual model merging. In *Proceedings of the Conference on Neural Information Processing Systems*, 2025.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[40] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the International Conference on Machine Learning*, pages 5389–5400, Proceedings of Machine Learning Research, 2019. PMLR.

[41] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Proceedings of the Conference on Neural Information Processing Systems*, 35:14274–14289, 2022.

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[43] Tao Sun, Cheng Lu, and Haibin Ling. Prior knowledge guided unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision.*, pages 639–655. Springer, 2022.

[44] Anke Tang, Enneng Yang, Li Shen, Yong Luo, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. Merging on the fly without retraining: A sequential approach to scalable continual model merging. In *Proceedings of the Conference on Neural Information Processing Systems*, 2025.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proceedings of the Conference on Neural Information Processing Systems*, 30, 2017.

[46] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the International Conference on Learning Representations*, 2021.

[47] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Proceedings of the Conference on Neural Information Processing Systems*, 32, 2019.

[48] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.

[49] Zixin Wang, Dong Gong, Sen Wang, Zi Huang, and Yadan Luo. Is less more? exploring token condensation as training-free test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 144–154, Piscataway, NJ, USA, 2025. IEEE.

[50] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[51] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[52] Shiqi Yang, Joost Van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 29393–29405, 2021.

[53] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2022.

[54] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909. IEEE, 2023.

[55] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23783–23793, Piscataway, NJ, USA, 2024. IEEE.

[56] Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 32111–32136, 2024.

[57] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18971–18981, 2023.

[58] Jingyi Zhang, Jiaxing Huang, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Historical test-time prompt tuning for vision foundation models. *Proceedings of the Conference on Neural Information Processing Systems*, 37:12872–12896, 2024.

[59] Kuan Zhang, Chengliang Chai, Jingzhe Xu, Chi Zhang, Han Han, Ye Yuan, Guoren Wang, and Lei Cao. Handling label noise via instance-level difficulty modeling and dynamic optimization. *Advances in Neural Information Processing Systems*, 38:46667–46696, 2026.

[60] Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: test time robustness via adaptation and augmentation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Proceedings of the Conference on Neural Information Processing Systems*, 2022.

[61] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Proceedings of the European Conference on Computer Vision.*, pages 493–510. Springer, 2022.

[62] Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. *Proceedings of the Conference on Neural Information Processing Systems*, 37:67795–67825, 2024.

[63] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28718–28728, 2024.

[64] Bowen Zhao, Chen Chen, and Shu-Tao Xia. Delta: Degradation-free fully test-time adaptation. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2023.

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[66] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022.

[67] Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29999–30009, Piscataway, NJ, USA, 2025. IEEE.

[68] Shunge Zou, Changhu Wang, Wei Ju, Ziyue Qiao, and Xiao Luo. Dance: Diversity-attended dynamic caching with asymmetric quantization for test-time adaptation of vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 37329–37343, 2026.

Appendix

A Limitations

CURE is primarily designed for continuous test-time adaptation of CLIP-like vision-language models under evolving distribution shifts. Consequently, its current formulation mainly focuses on image classification scenarios where class semantics can be represented by textual prompts. When extending the framework to more complex vision-language tasks, such as open-vocabulary detection, or multi-modal retrieval, task-specific adaptations may be required to handle structured outputs and heterogeneous visual-textual correspondences. One potential strategy to address this issue is to incorporate task-aware caching units and output-specific alignment objectives, allowing the proposed reliability modeling and subspace disentanglement mechanisms to operate beyond image-level classification.

B Large Language Model (LLM) Usage Statement

During the preparation of this manuscript, an LLM-based assistant was used to support language refinement, including identifying grammatical errors, improving the readability of descriptions, and exploring alternative expressions. All suggestions were carefully reviewed and revised by the authors. The research ideas, methodological development, experimental design, implementation, and result analysis were independently conducted by the authors. The LLM was not involved in any scientific decision-making or in the generation of the research findings. The authors take full responsibility for the accuracy and integrity of the entire manuscript.

C Broader impact

The proposed CURE framework advances continuous test-time adaptation of vision-language models by improving prediction reliability and mitigating catastrophic forgetting under evolving data distributions. By integrating complementary cache guidance and gradient space disentanglement, CURE enables models to adapt to changing environments while preserving previously acquired knowledge. By reducing the reliance on manually annotated data, CURE contributes to more reliable and adaptable vision-language systems in real-world environments.

D Related Works

D.1 Vision-Language Models

Traditional visual recognition approaches typically rely on training separate neural networks for specific tasks, which is notoriously labor-intensive [15]. Recently, Vision-Language Models (VLMs) have revolutionized this paradigm by learning rich cross-modal correlations from massive image-text pairs, enabling remarkable zero-shot predictions. For example, CLIP [39] pioneers this by utilizing contrastive learning to align image and text embeddings, while ALIGN [21] scales up the training with billion-level noisy pairs, and FILIP [53] captures fine-grained semantic interactions. Despite their robust zero-shot capabilities, an inherent distribution gap between pre-training data and downstream test environments often causes substantial performance degradation. To address this, existing literature typically adopts a few-shot adaptation paradigm, utilizing prompt learning methods like CoOp [66] and CoCoOp [65] to learn input context tokens, or adapter-based methods such as Tip-Adapter [61] and TaskRes [54] to modify extracted representations. However, their reliance on labeled target data incurs extra training costs and limits their real-world applicability.

D.2 Test-Time Adaptation for VLMs

TTA aims to align pre-trained models with unseen target distributions during inference, relying strictly on unlabeled test streams [46, 60]. Early TTA strategies primarily focused on single-modality vision models through techniques like entropy minimization [64] or batch normalization calibration [35]. Recently, extending this practical paradigm to Vision-Language Models has garnered significant attention. As a seminal work, TPT [41] pioneers this by optimizing learnable prompts

to enforce prediction consistency across diverse augmented views of a single test instance. This instance-specific optimization has inspired various extensions, such as leveraging diffusion-based augmentations (DiffTPT [11]) or explicitly aligning feature distributions (PromptAlign [1]). However, these approaches treat each test sample independently, which implicitly assumes a static target distribution and often requires computationally expensive backpropagation. To address dynamic real-world scenarios, an alternative trajectory treats test data as a continuous online stream. Methods like TDA [22] and BCA [67] establish memory caches or class priors to incrementally accumulate historical knowledge from incoming samples. Despite improving generalization, these cache-based methods typically rely only on a high-confidence subset of data or focus solely on a single modality, struggling to comprehensively model the continuous shift of real-world distributions.

E Proofs of Theoretical Results

E.1 Proof of Proposition 3.1 and Proposition 3.2

Proof. We prove the optimality and descent statements in turn.

For the optimality in Eq. (18), it follows from the Ky Fan maximum principle [8]: the rank-r orthogonal projector maximizing $\mathrm{tr}(\Pi G_t G_t^\top)$ is the projector onto the top-r eigenspace of $G_t G_t^\top$, i.e., $\mathrm{span}(U_r)$. Since Π_t is the orthogonal projector onto $\mathrm{span}(U_r)$, our target follows from the standard characterization of orthogonal projection.

For completeness, let the eigendecomposition of $G_t G_t^\top$ be

$$G_t G_t^\top = \sum_{k=1}^{n} \lambda_k u_k u_k^\top, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0. \tag{24}$$

Then for any rank-r orthogonal projector Π,

$$\mathrm{tr}(\Pi G_t G_t^\top) = \sum_{k=1}^{n} \lambda_k \langle u_k, \Pi u_k \rangle, \tag{25}$$

where $0 \leq \langle u_k, \Pi u_k \rangle \leq 1$ and $\sum_{k=1}^{n} \langle u_k, \Pi u_k \rangle = r$. Therefore,

$$\mathrm{tr}(\Pi G_t G_t^\top) \leq \sum_{k=1}^{r} \lambda_k = \mathrm{tr}(\Pi_t G_t G_t^\top), \tag{26}$$

which proves the optimality of Π_t in Eq. (18). Moreover, for any $z \in \mathrm{range}(\Pi_t)$, orthogonal projection gives the decomposition

$$\|g_t^s - z\|_2^2 = \|g_t^s - \Pi_t g_t^s\|_2^2 + \|\Pi_t g_t^s - z\|_2^2, \tag{27}$$

which immediately yields our target.

For the descent property, applying Assumption 3.1 with $\theta' = \theta_t - \eta_t \Pi_t g_t^s$ yields

$$\mathcal{L}^s(i_t, \theta_t') \leq \mathcal{L}^s(i_t, \theta_t) - \eta_t \langle g_t^s, \Pi_t g_t^s \rangle + \frac{L\eta_t^2}{2} \|\Pi_t g_t^s\|_2^2. \tag{28}$$

Since Π_t is symmetric and idempotent,

$$\langle g_t^s, \Pi_t g_t^s \rangle = (g_t^s)^\top U_r U_r^\top g_t^s = \|U_r^\top g_t^s\|_2^2 = \|\Pi_t g_t^s\|_2^2. \tag{29}$$

This proves Eq. (19). When $\eta_t \leq 1/L$, we have

$$\eta_t - \frac{L\eta_t^2}{2} \geq \frac{\eta_t}{2}, \tag{30}$$

which further proves Eq. (20). □

E.2 Proof of Proposition 3.3

Proof. For any $u \in \mathrm{span}(U_r)$, one has $\Pi_t u = u$. It follows that

$$\langle u, \delta_t^l \rangle = \langle u, (I - \Pi_t) \bar{g}_t^l \rangle = \langle (I - \Pi_t) u, \bar{g}_t^l \rangle = 0. \tag{31}$$

Taking $u = \Pi_t g_t^s \in \mathrm{span}(U_r)$ gives Eq. (22).

Following the Pythagorean theorem, by expanding the squared norm and using Eq. (22), we obtain

$$\|\Pi_t g_t^s + \delta_t^l\|_2^2 = \|\Pi_t g_t^s\|_2^2 + 2\langle \Pi_t g_t^s, \delta_t^l \rangle + \|\delta_t^l\|_2^2 = \|\Pi_t g_t^s\|_2^2 + \|\delta_t^l\|_2^2, \tag{32}$$

which is exactly Eq. (23). □

Remark on matrix-space projection. For simplicity, the above proof is written in vector form. The matrix-space projection used in Eq. (14) follows the same argument under the Frobenius inner product, since each rank-one basis can be vectorized as

$$\mathrm{vec}(u_i v_j^\top) = v_j \otimes u_i. \tag{33}$$

Indeed, for any two matrices A and B, the Frobenius inner product satisfies

$$\langle A, B \rangle_F = \mathrm{vec}(A)^\top \mathrm{vec}(B). \tag{34}$$

Therefore, orthogonality in the matrix space is equivalent to orthogonality after vectorization, and the above null-space argument applies directly to the matrix-space null projection in Eq. (14).

F Additional Implementation Details

F.1 Dataset Details

To comprehensively evaluate our proposed CURE approach across diverse distribution shifts, we utilize a wide range of image classification datasets. Table 3 summarizes the detailed statistics of all datasets included in our experimental setup. These datasets are broadly categorized into two groups. The first group consists of eleven standard datasets, ranging from general object and scene recognition, such as ImageNet, Caltech101, and SUN397, to highly specific fine-grained classification tasks, including FGVCAircraft and Food101. This diverse collection provides a comprehensive baseline for evaluating continual test-time adaptation capabilities across distinct visual domains. The second group comprises four specific ImageNet variants, namely ImageNet-V2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. These datasets are exclusively designed for evaluation purposes and contain only test sets. They are specifically incorporated to rigorously assess the model's robustness against out-of-distribution challenges, including domain shifts.

Dataset	Classes	Training	Validation	Testing	Task
Flowers102 [34]	102	4,093	1,633	2,463	Fine-grained flowers recognition
OxfordPets [36]	37	2,944	736	3,669	Fine-grained pets recognition
StanfordCars [25]	196	6,509	1,635	8,041	Fine-grained car recognition
FGVCAircraft [33]	100	3,334	3,333	3,333	Fine-grained aircraft recognition
Food101 [3]	101	50,500	20,200	30,300	Fine-grained food recognition
EuroSAT [16]	10	13,500	5,400	8,100	Satellite image recognition
UCF101 [42]	101	7,639	1,898	3,783	Action recognition
DTD [5]	47	2,820	1,128	1,692	Texture recognition
SUN397 [50]	397	15,880	3,970	19,850	Scene recognition
Caltech101 [10]	100	4,128	1,649	2,465	Object recognition
ImageNet [6]	1,000	1.28M	-	50,000	Object recognition
ImageNet-V2 [40]	1,000	-	-	10,000	Robustness to distribution shifts
ImageNet-Sketch [47]	1,000	-	-	50,889	Robustness of sketch domain
ImageNet-A [18]	200	-	-	7,500	Robustness of adversarial attack
ImageNet-R [17]	200	-	-	30,000	Robustness of multi-domains

Table 3: Detailed statistics of the experimental datasets. The four ImageNet variants are used exclusively for evaluation and contain only test samples.

F.2 Implementation details

We adopt the pre-trained Zero-Shot CLIP as our backbone, utilizing both ViT-B/16 and ResNet-50 as visual encoders. During the continuous test-time adaptation, the text encoder and the majority of the visual encoder remain frozen. Following standard continuous adaptation protocols, we only update the lightweight affine parameters of the normalization layers. We optimize the model using AdamW with a learning rate of $1 \times 10^{-}4$.

Unless otherwise specified, we maintain a unified set of hyper-parameters across all benchmark datasets to demonstrate the robustness of CURE. The dual-caching queue capacity is set to $K = 16$ samples per class. The entropy threshold for reliability filtering is fixed at $\tau = 0.7$, and the threshold trade-off for long-term caching is set to $\beta = 0.5$. For the low-rank subspace disentanglement, the rank

Method	Flo.	DTD	Pets	Cars	UCF	Cal.	Food	SUN	Air.	Euro.	Avg.	Rank
Visual Encoder Backbone: ResNet-50												
CLIP-ResNet50	62.77	40.37	82.97	55.89	59.48	87.26	74.82	60.85	16.11	25.79	56.63	8.9
CoOp	61.55	37.29	87.00	55.32	59.05	86.53	75.59	58.15	15.12	26.20	56.18	8.9
TPT	62.69	40.84	84.49	58.46	60.82	87.02	74.88	61.46	17.58	28.33	57.66	7.9
DiffTPT	63.53	40.72	83.40	60.71	62.67	86.89	79.21	62.72	17.60	41.04	59.85	6.1
TDA	68.74	43.74	86.18	57.78	64.18	89.70	77.75	62.53	17.61	42.11	61.03	4.7
DPE	67.60	50.18	85.97	59.26	61.98	**90.83**	77.83	64.23	19.80	41.67	61.94	3.7
BoostAdapter	68.25	43.85	85.75	59.67	**64.42**	88.48	78.78	62.83	18.93	**44.40**	61.54	3.4
HisTPT	67.60	41.30	84.90	**61.30**	64.10	87.20	**81.30**	63.50	18.10	42.50	61.18	4.0
BCA	66.30	48.58	85.58	58.13	63.51	89.70	77.19	63.38	19.89	42.12	61.44	4.6
CURE (Ours)	**69.75**	**52.48**	**87.03**	59.61	61.75	90.14	77.80	**64.62**	**20.85**	40.52	**62.46**	**3.0**

Table 4: Comparison of CURE in Cross-Datasets Generalization using ResNet-50 as the backbone. Bold values indicate the best performance. **Abbreviations**: Flo. (Flowers102), Cal. (Caltech101), Food (Food101), SUN (SUN397), Air. (FGVC Aircraft), Euro. (EuroSAT), UCF (UCF101).

of the principal subspace is consistently set to $r = 8$. Due to the inherent architectural differences between Transformers and CNNs, we apply slightly different temporal windows and penalty factors as analyzed in our parameter experiments:

- **For ViT-B/16:** The consistency penalty is set to $\gamma = 1.2$, and the short-term window size is $t_s = 64$.
- **For ResNet-50:** The consistency penalty is set to $\gamma = 2.0$, and the short-term window size is $t_s = 32$.

All experiments, including the efficiency comparisons, are conducted on a single NVIDIA RTX 5090D GPU (32GB VRAM) to ensure fair latency and memory measurements. The framework is implemented using PyTorch 2.8.

G Additional Experimental Results

G.1 Additional Results with ResNet-50 Architecture

To complement the main evaluations, we present the performance of CURE utilizing ResNet-50 as the visual encoder backbone. As detailed in Table 4, our approach maintains robust performance on the Cross-Dataset benchmark, achieving the highest average accuracy of 62.46%. This validates that the core mechanisms of our method are architecture-agnostic. They consistently mitigate catastrophic forgetting and seamlessly integrate target-domain knowledge into the long-term cache, proving highly effective on standard convolutional networks alongside Vision Transformers. Furthermore, we evaluate the ResNet-50 backbone on the Domain Generalization benchmark, encompassing ImageNet and its four out-of-distribution (OOD) variants, as summarized in Table 5. CURE attains a state-of-the-art average accuracy of 50.88% and secures the best average rank of 1.6. Despite the inherent structural differences between CNNs and ViTs, our subspace disentanglement updating mechanism effectively safeguards the network against OOD noise and severe distribution shifts.

G.2 Ablation Study with ResNet-50

To further validate the generalizability of our framework, we conduct a leave-one-out ablation study utilizing the ResNet-50 backbone (Figure 5). Consistent with the ViT observations, removing any single module results in severe performance degradation. Notably, discarding the null-space projection (*w/o Null-Space Proj.*) causes the most significant overall accuracy drop of 4.13%. This highlights that unconstrained parameter updates rapidly corrupt the pre-trained CNN representations, making geometric protection explicitly crucial for CNN architectures during continuous adaptation. Furthermore, eliminating the low-rank constraint (*w/o Low-Rank Proj.*, -4.07%) or the dual-caching architecture (*w/o Long Caching*, -4.03%) severely impairs the model's ability to filter transient noise and balance short-term plasticity with historical stability. Finally, omitting the semantic voting mechanism (*w/o Neighbor Rel.*, -3.75%) leads to error accumulation from overconfident pseudo-labels. Together, these complementary modules enable CURE to achieve a comprehensive 53.68%

Method	Venue	IN	IN-A	IN-V2	IN-R	IN-S	Avg.	OOD Avg.	Rank
Visual Encoder Backbone: ResNet-50									
CLIP-ResNet50	ICML2021	58.16	21.83	51.41	56.15	33.37	44.18	40.69	10.0
CoOp	IJCV2022	63.33	23.06	55.40	56.60	34.67	46.61	42.43	7.4
CoCoOp	CVPR2022	62.81	23.32	55.72	57.74	34.48	46.81	42.82	7.0
TPT	NIPS2022	60.70	26.56	54.77	59.02	35.13	47.24	43.87	7.6
DiffTPT	ICCV2023	60.80	**31.06**	55.80	58.80	37.10	48.71	45.69	5.2
GS-Bias	ICML2025	62.05	27.83	55.91	63.01	36.98	49.16	45.93	4.8
TDA	CVPR2024	61.35	30.29	55.54	62.58	38.12	49.58	46.63	5.2
DPE	NIPS2024	63.41	30.15	56.72	**63.72**	**40.03**	50.81	47.66	2.2
BCA	CVPR2025	61.81	30.35	56.58	62.89	38.08	49.94	46.98	4.0
CURE (Ours)	-	**63.60**	30.37	**56.86**	63.69	39.88	**50.88**	**47.70**	**1.6**

Table 5: Comparison on ImageNet and its variants using ResNet-50. **Abbreviations**: IN (ImageNet), IN-A (ImageNet-A), IN-V2 (ImageNet-V2), IN-R (ImageNet-R), IN-S (ImageNet-S). Rank is calculated across the 5 datasets (lower is better).

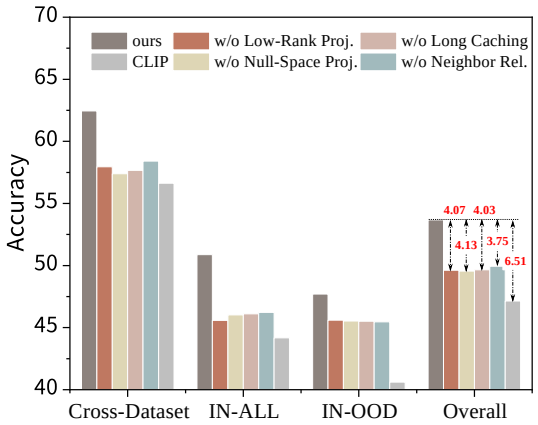

Figure 5: Ablation study of our proposed CURE framework utilizing the ResNet-50 backbone. Red dashed arrows indicate differences in overall accuracy relative to the complete model, including module ablations and the zero-shot CLIP baseline.

overall accuracy, securing a substantial +6.51% absolute improvement over the Zero-Shot CLIP baseline.

G.3 Parameter Sensitivity Analysis with ResNet-50

To evaluate the architectural adaptability of CURE, we perform a parameter sensitivity analysis utilizing the ResNet-50 backbone. While the optimal configurations largely align with the ViT backbone, notable differences emerge due to the inherent properties of CNNs. For input reliability modeling, the consistency penalty peaks at a much higher value ($\gamma = 2.0$) compared to ViT ($\gamma = 1.2$). This indicates that CNN representations produce more severe overconfident misclassifications under domain shifts, necessitating a harsher penalty from the semantic voting mechanism to effectively suppress unreliable predictions. Regarding the dual memory capacity, accuracy reaches its optimum at $K = 16$, demonstrating that an appropriate queue size adequately captures class-wise feature distributions without introducing outdated historical noise. Furthermore, for the geometric subspace updates, the optimal short-term window size is noticeably smaller ($t_s = 32$). This suggests that CNN feature distributions evolve more rapidly in continuous data streams; thus, extracting principal directions over a shorter temporal window prevents the low-rank subspace from lagging behind immediate distribution shifts. Spatially, the principal subspace rank peaks at $r = 8$, ensuring responsive short-term adaptation while maintaining the complementary null-space to preserve historical stability.

Statistical Significance Analysis. We further assess the statistical reliability of the results using critical difference diagrams in Figure 7, where lower average rank indicates better overall performance. Across both ViT-B/16 and ResNet-50 backbones, CURE consistently achieves the best average rank on Cross-Datasets and ImageNet-based evaluations. Compared with strong competitors such as DPE,

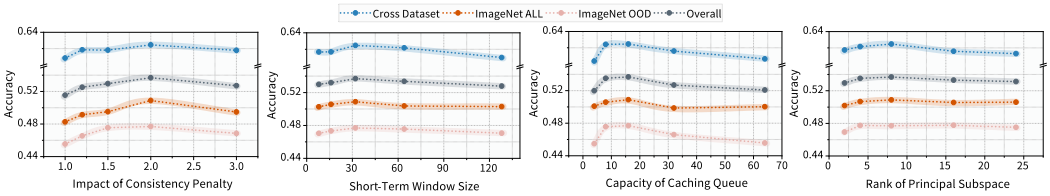

Figure 6: Parameter analysis of CURE utilizing the ResNet-50 backbone. Each subplot investigates the impact of a distinct parameter setting.

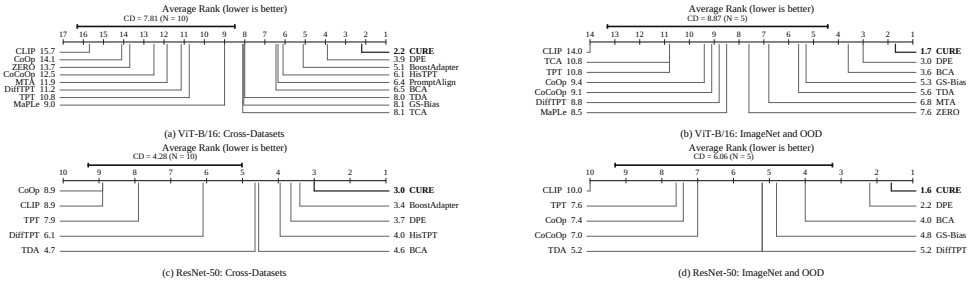

Figure 7: Critical difference diagrams of CURE and competing methods across different backbones and evaluation benchmarks. The diagrams are computed from average ranks using the Nemenyi post-hoc test, where lower ranks indicate better performance.

BCA, and TDA, CUREachieves the best average rank across different backbones and distribution-shift settings, demonstrating its consistent performance across evaluation benchmarks.

Domain Order	IN	IN-V2	IN-S	Avg.
IN $\rightarrow$ IN-V2 $\rightarrow$ IN-S	72.15	65.72	52.49	63.45
IN-S $\rightarrow$ IN-V2 $\rightarrow$ IN	71.30	65.00	53.33	63.21
IN-V2 $\rightarrow$ IN-S $\rightarrow$ IN	71.43	65.84	52.37	63.21

Table 6: Robustness to different domain arrival orders.

G.4 Robustness to Domain Arrival Order

We evaluate CURE under three domain arrival orders over ImageNet, ImageNet-V2, and ImageNet-Sketch, keeping all samples and hyperparameters fixed. As shown in Table 6, the average accuracy varies by only 0.24 percentage points across different orders, demonstrating the robustness of CURE to changes in domain arrival sequences.

Algorithm 1 CURE: Online Continual Test-Time Adaptation

Require: Pretrained VLM $(F_v, F_t, \boldsymbol{\theta}_0)$; test stream $\{x_t\}_{t=1}^{T}$; prompt sets $\{\mathcal{P}_c\}_{c=1}^{C}$; short- and long-term memory queues $(\mathcal{M}_s, \mathcal{M}_\ell)$; memory size K_m; group number M; windows (t_s, t_ℓ); rank r; penalty γ; admission ratio ξ; step sizes $\{\eta_t\}$; alignment weight λ.

Ensure: Online predictions $\{\hat{y}_t\}_{t=1}^{T}$ and adapted parameters $\boldsymbol{\theta}_{T+1}$.

Initialize $\boldsymbol{\theta}_1 \leftarrow \boldsymbol{\theta}_0$, $\beta_0 \leftarrow 1$, $\beta_1 \leftarrow 1$, and gradient buffers $\mathcal{G}_s, \mathcal{G}_\ell \leftarrow \emptyset$.

Build the textual voting committee $\widetilde{\mathcal{T}} = \{\widetilde{\boldsymbol{t}}_m^c\}_{c=1,m=1}^{C,M}$ from $\{\mathcal{P}_c\}_{c=1}^{C}$.

for $t = 1$ to T **do**

Predict before adaptation: $\boldsymbol{v}_t \leftarrow F_v(x_t; \boldsymbol{\theta}_t)$, $p_t \leftarrow p_{\boldsymbol{\theta}_t}(\cdot|x_t)$, $\hat{y}_t \leftarrow \arg\max_c p_t(c)$.

Obtain $\mathbb{P}_{\mathrm{TPT}}(x_t)$ using entropy-filtered augmented views $\{\mathcal{A}_n(x_t)\}_{n=1}^{N}$.

Estimate reliability by textual majority voting:

$$(y_t^\star, S_t, R_t, H_t') \leftarrow \mathrm{VoteReliability}(\boldsymbol{v}_t, \hat{y}_t, p_t, \widetilde{\mathcal{T}}, \gamma).$$

Update the short-term memory with reliability-calibrated entropy:

$$\mathcal{M}_s \leftarrow \mathrm{UpdateShort}(\mathcal{M}_s, \boldsymbol{v}_t, y_t^\star, H_t', K_m).$$

Compute the short-term objective $\mathcal{L}_t^s$ using $\mathbb{P}_{\mathrm{TPT}}(x_t)$, $\mathcal{M}_s$, and $\{\widetilde{\boldsymbol{t}}_M^c\}_{c=1}^{C}$.

Compute $\boldsymbol{g}_t^s \leftarrow \nabla_{\boldsymbol{\theta}_t} \mathcal{L}_t^s$ and update $\mathcal{G}_s \leftarrow \mathrm{Buffer}(\mathcal{G}_s, \boldsymbol{g}_t^s, t_s)$.

Stack recent gradients as $\boldsymbol{G}_t^s \leftarrow \mathrm{Concat}(\mathcal{G}_s) \in \mathbb{R}^{n \times |\mathcal{G}_s|}$ and extract $\boldsymbol{U}_r \leftarrow \mathrm{TopSVD}_r(\boldsymbol{G}_t^s)$.

Perform low-rank short-term adaptation:

$$\boldsymbol{\theta}_t' \leftarrow \boldsymbol{\theta}_t - \eta_t \boldsymbol{U}_r \boldsymbol{U}_r^\top \boldsymbol{g}_t^s.$$

Update the long-term memory using class-balanced contribution scores:

$$\mathcal{M}_\ell \leftarrow \mathrm{UpdateLong}(\mathcal{M}_\ell, \mathcal{M}_s, \boldsymbol{\theta}_t, \xi, K_m).$$

Compute the long-term objective $\mathcal{L}_t^\ell$ using $\mathbb{P}_{\mathrm{TPT}}(x_t)$, $\mathcal{M}_\ell$, and $\{\widetilde{\boldsymbol{t}}_M^c\}_{c=1}^{C}$.

Compute $\boldsymbol{g}_t^\ell \leftarrow \nabla_{\boldsymbol{\theta}_t} \mathcal{L}_t^\ell$ and update $\mathcal{G}_\ell \leftarrow \mathrm{Buffer}(\mathcal{G}_\ell, \boldsymbol{g}_t^\ell, t_\ell)$.

Obtain the null-space correction:

$$\bar{\boldsymbol{g}}_t^\ell \leftarrow \mathrm{Avg}(\mathcal{G}_\ell), \qquad \boldsymbol{\delta}_t^\ell \leftarrow (\boldsymbol{I} - \boldsymbol{U}_r \boldsymbol{U}_r^\top) \bar{\boldsymbol{g}}_t^\ell.$$

Set $\Delta\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t' - \boldsymbol{\theta}_0$ and update the calibration factor β_t.

Consolidate short-term plasticity and long-term robustness:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_0 + \frac{\beta_{t-1} \Delta\boldsymbol{\theta}_t - \boldsymbol{\delta}_t^\ell}{\beta_t}.$$

end for

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present abstract and introduction that accurately reflect the contribution of our work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide comprehensive proof in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: These can be seen in section implementation details. we also provide our code on github.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our code on github.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide these details in section implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide Experiment statistical significance in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We present it in section implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We fully abide the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential positive and negative societal impacts of CURE in Appendix C.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The assets we use have been properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide our code on anonymous github.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are part of the core method, and we describe their role in Sec. 3.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.