ACME: ONLINE ADAPTATION VIA CROSS-MOMENT ESTIMATION FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) provide strong zero-shot predictions, but their performance can deteriorate under test-time distribution shift. Test-time adaptation (TTA) aims to mitigate this degradation using unlabeled test samples during inference. Existing cache-based TTA methods retain confidence-selected examples or hard pseudo-labels, which can discard useful target evidence and propagate unreliable predictions. Distributional methods instead estimate target-domain feature distributions, but often require costly covariance state or task-sensitive hyperparameters whose improper configuration can cause severe performance degradation. To address these issues, we introduce Adaptive Cross-Moment Estimation (ACME), a tuning-free and backpropagation-free online adapter that uses the complete frozen posterior of every test sample. ACME maintains exact streaming feature and posterior statistics together with their cross-covariance, from which it derives class-specific directions for a bounded residual correction to the frozen logits. ACME requires no source data, stored test samples, or full class-conditional covariance matrices. Extensive experiments across distribution-shift and cross-dataset benchmarks demonstrate that ACME achieves strong adaptation performance with a negligible additional GPU memory overhead of 3 MiB and the highest inference throughput among the evaluated TTA methods, retaining 88% of frozen CLIP's throughput on ImageNet-1K.

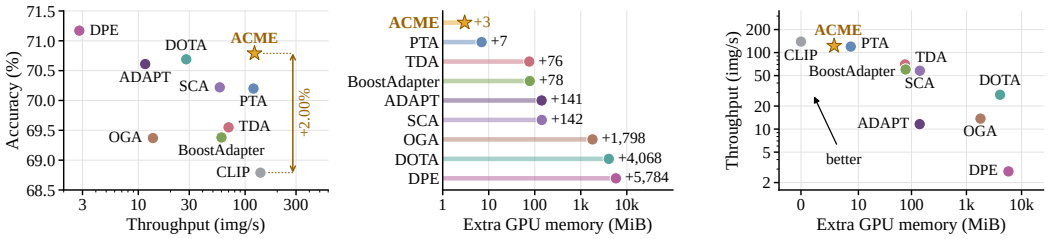
1 INTRODUCTION

Vision-language models (VLMs) such as CLIP (Radford et al., 2021) learn aligned visual and textual representations from large-scale image-text pairs, enabling strong zero-shot recognition across diverse downstream tasks. However, their performance can degrade substantially when the test distribution differs from the data observed during pre-training (Shu et al., 2022; Yoon et al., 2024). Prior work has addressed this limitation by adapting VLMs to downstream tasks using labeled examples (Gao et al., 2024; Zhang et al., 2022; Yu et al., 2023). Although effective, these approaches depend on high-quality, task-specific annotations, which are costly to obtain and often unavailable during deployment. Test-time adaptation (TTA) has therefore emerged as a promising solution that improves model robustness using only unlabeled test samples during inference, without requiring labeled target data or conventional supervised retraining (Wang et al., 2021).

Recent TTA methods can be broadly categorized into optimization-based and training-free approaches. Optimization-based methods refine textual prompts (Shu et al., 2022; Feng et al., 2023; Yoon et al., 2024) or jointly evolve textual and visual prototypes (Zhang et al., 2024a). Although effective, these methods rely on iterative optimization and backpropagation at test time, resulting in considerable inference and memory overhead. In contrast, backpropagation-free adapters keep the pretrained VLM frozen and adapt its predictions by accumulating information in auxiliary test-time state. Cache-based methods store and retrieve selected image representations (Karmanov et al., 2024; Zhang et al., 2024b;c), prototype-based methods recursively summarize test observations as compact class representations (Zhou et al., 2025; Huang et al., 2026), and statistics-based methods estimate target-domain feature statistics or class-conditional distributions to construct adaptive classifiers (Dai & Yang, 2025; Guan et al., 2025; Zhang et al., 2025; Han et al., 2025).

Despite recent advancements, existing backpropagation-free TTA methods retain several limitations: (i) cache-based adapters store confidence-selected test features using hard pseudo-labels. Erroneous

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Accuracy vs. speed (b) Extra memory (c) Speed vs. memory

Figure 1: **Accuracy-efficiency comparison of test-time adaptation methods** using CLIP ViT-B/16 on ImageNet-1K (Deng et al., 2009). (a) Accuracy versus inference throughput. (b) Additional GPU memory relative to frozen CLIP. (c) Throughput versus additional GPU memory, where the upper-left region indicates better efficiency. ACME improves frozen CLIP from 68.79% to 70.79% accuracy (+2.00%), while retaining 88% of its throughput and incurring a negligible additional GPU-memory overhead of 3 MiB. Compared with DPE (Zhang et al., 2024a), ACME is approximately 44× faster and reduces the additional GPU memory from 5,784 MiB to only 3 MiB.

pseudo-labels can place unreliable samples into the wrong class-specific cache and consequently harm adaptation performance, while informative high-entropy samples may be discarded(Karmanov et al., 2024; Zhang et al., 2024b;c). Moreover, although cache capacity is bounded for each class, the total cache size and retrieval cost grow with the number of classes, increasing memory overhead and reducing inference speed (Huang et al., 2026) in large-scale settings such as ImageNet-1K (Deng et al., 2009); (ii) prototype-based methods avoid explicit cache retrieval by compressing the accumulated observations into compact class representatives. Although more efficient, their updates remain dependent on the confidence-derived weighting rules, allowing incorrect evidence to shift the prototypes and influence subsequent predictions (Zhou et al., 2025; Huang et al., 2026); (iii) statistics-based methods preserve richer target-domain information, but estimating class-conditional distributions can require large covariance state, matrix inversion, historical knowledge banks, and considerable memory and computation at inference (Dai & Yang, 2025; Guan et al., 2025; Zhang et al., 2025; Han et al., 2025); and (iv) many methods depend on task-sensitive hyperparameters, including confidence thresholds, update rates, regularization strengths, and fusion coefficients (Zhou et al., 2025; Karmanov et al., 2024; Han et al., 2025). These hyperparameters are typically tuned using validation sets from specific datasets and may not transfer reliably to unseen target distributions. Such a mismatch can compromise the model's zero-shot robustness and, in extreme cases, cause catastrophic performance degradation (Boudiaf et al., 2022).

To address the above limitations, we introduce Adaptive Cross-Moment Estimation (ACME), a tuning-free and backpropagation-free test-time adapter for frozen VLMs. Our core insight is to accumulate rather than select from the evidence provided by test samples. For each observation, ACME relates the frozen model's complete posterior to its visual representation and updates the resulting posterior-feature cross-covariance, together with the required marginal statistics, through exact streaming recurrences. Every sample therefore contributes graded evidence to every class without confidence filtering, hard pseudo-label assignment, or example storage. ACME maintains only a $C \times D$ cross-covariance matrix and a small collection of marginal statistics, requiring fixed $O(CD)$ memory and computation per image, independent of the number of previously observed samples. It requires no full feature-covariance matrix, matrix inversion, or dataset-selected adaptation hyperparameters. At prediction time, the accumulated association produces class-specific correlation directions that provide a normalized and bounded correction to the frozen logits.

We conduct extensive experiments on cross-domain and natural distribution-shift benchmarks using multiple visual backbones and stream orders. ACME consistently improves upon the matched frozen classifier and prototype-accumulation baseline while achieving competitive or superior accuracy to recent cache-, prototype-, and distribution-based adapters. Its fixed-size state and closed-form updates retain near-frozen inference speed and require among the lowest adaptation memory of the evaluated methods. As shown in Figures 1a and 1b, ACME is the fastest among the evaluated adaptation methods, retaining 88% of frozen CLIP's throughput while incurring only 3 MiB of additional GPU memory, a negligible overhead. Notably, ACME matches or exceeds adapters that estimate full class-specific $D \times D$ covariance matrices, despite requiring only $O(CD)$ state rather than $O(CD^2)$.

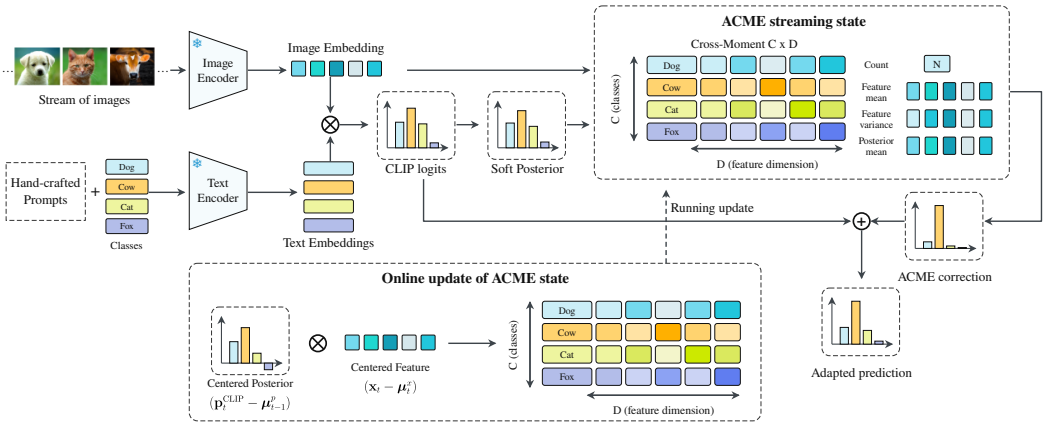


Figure 2: **Overview of our proposed Adaptive Cross-Moment Estimation (ACME).** Given a stream of unlabeled test samples, the frozen CLIP encoders produce image and text embeddings together with the corresponding soft posterior. ACME continuously updates a compact posterior-feature cross-covariance and the required marginal statistics without storing past examples. The resulting class-wise correction is combined with the frozen CLIP logits to obtain the adapted prediction.

This controlled comparison indicates that class-specific pairwise feature covariance is not required for effective adaptation on the evaluated benchmarks; the substantially smaller posterior-feature cross-covariance retains sufficient target-domain structure. Together, these results establish ACME as an accurate, efficient, and tuning-free approach to test-time adaptation of frozen VLMs.

2 Related Work

2.1 Vision-Language Models

Contrastively pretrained vision-language models, including CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), learn a shared embedding space in which images can be classified by comparison with textual class descriptions. Their zero-shot transfer has motivated several approaches to downstream adaptation. CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) learn textual prompts, while MaPLe learns coupled visual and textual prompts (Khattak et al., 2023). Other approaches introduce lightweight feature or classifier corrections, as in CLIP-Adapter (Gao et al., 2024) and TaskRes (Yu et al., 2023), or construct a labeled few-shot cache, as in Tip-Adapter (Zhang et al., 2022). These methods improve task-specific transfer when annotated support data is available. We instead consider online adaptation from an unlabeled test stream, without updating either pretrained encoder or accessing labeled target examples.

2.2 Test-Time Adaptation

Test-time adaptation uses unlabeled observations to respond to distribution shift after deployment. Early approaches such as TENT (Wang et al., 2021) update model parameters by minimizing prediction entropy. For vision-language models, TPT, DiffTPT, and C-TPT optimize prompts using augmented test views (Shu et al., 2022; Feng et al., 2023; Yoon et al., 2024). DPE (Zhang et al., 2024a) evolves visual and textual prototypes but also optimizes per-sample residuals. These approaches can improve recognition, although their test-time optimization introduces additional computation.

Backpropagation-free methods instead retain information across the test stream. TDA (Karmanov et al., 2024) and BoostAdapter (Zhang et al., 2024b) maintain selected test features in pseudo-labeled caches, while dual-memory approaches (Zhang et al., 2024c) combine stored test features with other forms of support. BCA (Zhou et al., 2025) updates class embeddings and class priors. More compact representations include PTA's confidence-weighted class prototypes (Huang et al., 2026) and SCA's incrementally accumulated statistics for a regression-based classifier (Guan et al., 2025). Distribution-based approaches model test-time features using online Gaussian estimates or proba-

bilistic updates, including OGA (Fuchs et al., 2025), FreeTTA (Dai & Yang, 2025), ADAPT (Zhang et al., 2025), and DOTA (Han et al., 2025).

Statistical moments and distribution embeddings have been widely used to characterize distributions and support adaptation under distribution shift. Kernel mean embeddings provide a general mechanism for comparing distributions (Gretton et al., 2012), while CORAL aligns second-order statistics between source and target domains (Sun et al., 2016; Sun & Saenko, 2016). These statistical tools and alignment methods were developed primarily for distribution comparison and batch domain adaptation rather than online TTA. Related normalization-based methods adjust feature statistics using unlabeled target samples (Li et al., 2018; Schneider et al., 2020). Unlike these source-target alignment and normalization-based approaches, ACME estimates an online cross-moment between frozen class posteriors and visual features, using the resulting association to correct the frozen CLIP logits without accessing source data or modifying the model parameters during inference.

3 Methodology

3.1 Preliminaries

Contrastive language-image pretraining. Contrastive vision-language models such as CLIP (Radford et al., 2021) consist of a frozen image encoder f_{v} and a frozen text encoder f_{t} that map images and textual descriptions into a shared D-dimensional representation space. Given C downstream classes, the prompt templates associated with each class are encoded once. Their normalized text representations are averaged and renormalized to obtain the fixed text classifier

$$\mathbf{W} = [\mathbf{w}_1 \quad \cdots \quad \mathbf{w}_C] \in \mathbb{R}^{D\times C}, \qquad \|\mathbf{w}_c\| = 1. \tag{1}$$

For the t-th image I_t, its normalized image representation, frozen logits, and frozen posterior are

$$\mathbf{x}_t = \frac{f_{\mathrm{v}}(I_t)}{\|f_{\mathrm{v}}(I_t)\|}, \qquad \boldsymbol{\ell}_t^{\mathrm{CLIP}} = \tau \mathbf{W}^\top \mathbf{x}_t, \qquad \mathbf{p}_t^{\mathrm{CLIP}} = \mathrm{softmax}\left(\boldsymbol{\ell}_t^{\mathrm{CLIP}}\right), \tag{2}$$

where $\mathbf{x}_t \in \mathbb{R}^D$, $\mathbf{p}_t^{\mathrm{CLIP}} \in \mathbb{R}^C$, and τ is the fixed inverse temperature of the softmax function.

3.2 Online State

Let $\mathcal{S}_t$ denote the adaptation state after processing the t-th test observation. ACME follows an update-before-predict protocol (Karmanov et al., 2024; Huang et al., 2026): the current feature and frozen posterior are first incorporated into the state, after which the corresponding prediction is produced:

$$\mathcal{S}_t = \mathcal{U}\left(\mathcal{S}_{t-1}, \mathbf{x}_t, \mathbf{p}_t^{\mathrm{CLIP}}\right), \qquad \widehat{y}_t = \arg\max_c \ell_{t,c}^{\mathrm{final}}. \tag{3}$$

After observing t samples, the state is

$$\mathcal{S}_t = \left(t, \boldsymbol{\mu}_t^x, \boldsymbol{\mu}_t^p, \mathbf{v}_t^x, \mathbf{C}_t^{px}\right), \tag{4}$$

where $\boldsymbol{\mu}_t^x \in \mathbb{R}^D$ and $\boldsymbol{\mu}_t^p \in \mathbb{R}^C$ are the empirical feature and posterior means, $\mathbf{v}_t^x \in \mathbb{R}^D$ is the coordinate-wise empirical feature variance, and $\mathbf{C}_t^{px} \in \mathbb{R}^{C\times D}$ is the empirical posterior-feature cross-covariance. Both second-order statistics use the population normalization $1/t$, rather than the unbiased normalization $1/(t-1)$, and all four statistics are initialized to zero at $t = 0$.

The central quantity estimated by ACME is the population cross-covariance between the frozen posterior and visual representation under the target distribution:

$$\mathbf{C}^{px} = \mathrm{Cov}\left(\mathbf{p}^{\mathrm{CLIP}}, \mathbf{x}\right) = \mathbb{E}\left[\left(\mathbf{p}^{\mathrm{CLIP}} - \boldsymbol{\mu}^p\right)\left(\mathbf{x} - \boldsymbol{\mu}^x\right)^\top\right] \in \mathbb{R}^{C\times D}. \tag{5}$$

For any feature direction $\mathbf{a} \in \mathbb{R}^D$,

$$[\mathbf{C}^{px}]_{c,:}\mathbf{a} = \mathrm{Cov}\left(p_c^{\mathrm{CLIP}}, \mathbf{a}^\top \mathbf{x}\right). \tag{6}$$

Thus, the c-th row of $\mathbf{C}^{px}$ characterizes how movement in the visual feature space is associated with changes in the frozen evidence for class c. Unlike hard pseudo-label updates, this association does not assign an observation to a single class.

The feature and posterior means obey the standard online-average recurrences

$$\boldsymbol{\mu}_t^x = \boldsymbol{\mu}_{t-1}^x + \frac{1}{t}\left(\mathbf{x}_t - \boldsymbol{\mu}_{t-1}^x\right), \qquad \boldsymbol{\mu}_t^p = \boldsymbol{\mu}_{t-1}^p + \frac{1}{t}\left(\mathbf{p}_t^{\mathrm{CLIP}} - \boldsymbol{\mu}_{t-1}^p\right). \tag{7}$$

The second-order statistics are updated without retaining previous observations using Welford-style online recurrences (Welford, 1962). The coordinate-wise feature variance follows

$$\mathbf{v}_t^x = \mathbf{v}_{t-1}^x + \frac{1}{t}\left[\left(\mathbf{x}_t - \boldsymbol{\mu}_{t-1}^x\right) \odot \left(\mathbf{x}_t - \boldsymbol{\mu}_t^x\right) - \mathbf{v}_{t-1}^x\right]. \tag{8}$$

Similarly, the posterior-feature cross-covariance follows the paired recurrence

$$\mathbf{C}_t^{px} = \mathbf{C}_{t-1}^{px} + \frac{1}{t}\left[\left(\mathbf{p}_t^{\mathrm{CLIP}} - \boldsymbol{\mu}_{t-1}^p\right)\left(\mathbf{x}_t - \boldsymbol{\mu}_t^x\right)^\top - \mathbf{C}_{t-1}^{px}\right]. \tag{9}$$

These recurrences retain the exact empirical means, coordinate-wise variance, and posterior-feature cross-covariance of the observed stream without storing any test example.

3.3 Standardized Readout

The visual representation may exhibit substantially different scales across feature dimensions. Consequently, the magnitude of a cross-covariance entry may reflect the variance along the corresponding dimension rather than the strength of its association with a class. To remove this scale dependence, we normalize the current feature deviation and posterior-feature cross-covariance using empirical standard deviations estimated from the observed stream. The resulting directional comparison is therefore invariant to feature-wise rescaling.

The standardized feature residual of the current observation is

$$\mathbf{z}_t = \left(\mathbf{x}_t - \boldsymbol{\mu}_t^x\right) \oslash \sqrt{\mathbf{v}_t^x}. \tag{10}$$

For class c, the corresponding standardized posterior–feature direction is

$$\mathbf{r}_{t,c} = \left([\mathbf{C}_t^{px}]_{c,:}\right)^\top \oslash \sqrt{\mathbf{v}_t^x} \in \mathbb{R}^D. \tag{11}$$

We then measure the agreement between these standardized quantities using cosine similarity:

$$s_{t,c}^{\mathrm{ACME}} = \frac{\mathbf{z}_t^\top \mathbf{r}_{t,c}}{\|\mathbf{z}_t\|\|\mathbf{r}_{t,c}\|}. \tag{12}$$

Standardizing the cross-covariance by the posterior standard deviation is unnecessary in this readout. For each class, it would rescale the entire direction $\mathbf{r}_{t,c}$ by a single positive factor, which cancels under cosine normalization. We therefore do not retain posterior variances in the adaptation state. For numerical stability, we lower-bound each feature variance by a small constant ($\varepsilon = 10^{-8}$); the cosine score is defined as zero whenever either vector has zero norm.

Evaluating equation 12 for all classes gives

$$\mathbf{s}_t^{\mathrm{ACME}} = \begin{bmatrix} s_{t,1}^{\mathrm{ACME}} & \cdots & s_{t,C}^{\mathrm{ACME}} \end{bmatrix}^\top \in [-1, 1]^C. \tag{13}$$

The adaptive readout therefore depends on directional agreement between the current standardized deviation and each target-derived class association, rather than the magnitude of either vector.

3.4 Maturity-Weighted Logit Fusion

Early cross-covariance estimates are supported by only a small number of observations and should not override the frozen classifier. We define the sample-to-dimension maturity

$$m_t = \frac{t}{t + D}, \tag{14}$$

which increases monotonically toward one as target observations accumulate.

The fully target-informed predictor is obtained by adding the maturity-weighted ACME score to the frozen logits:

$$\widetilde{\boldsymbol{\ell}}_t = \boldsymbol{\ell}_t^{\mathrm{CLIP}} + \tau m_t \mathbf{s}_t^{\mathrm{ACME}}, \qquad \widetilde{\mathbf{p}}_t = \mathrm{softmax}\left(\widetilde{\boldsymbol{\ell}}_t\right). \tag{15}$$

Table 1: **Comparison of ACME on Cross-Domain Datasets using ViT-B/16 as the backbone. Bold** and underlined values indicate the best and second-best performance among all methods, respectively. For a fair comparison, all reproduced methods use the handcrafted prompt templates described in CLIP (Radford et al., 2021). We report their mean accuracies over five randomly shuffled test orders. Here, † denotes results reproduced using the officially released code. Other results are taken from the corresponding papers using consistent prompt templates.

Method	Aircraft	Caltech	Cars	DTD	EuroSAT	Flower	Food101	Pets	SUN397	UCF101	Average
CLIP-ViT-B/16† [ICML'21]	24.90	93.87	66.17	44.21	48.46	70.77	85.89	89.10	66.27	67.54	65.72
Backpropagation-Based Test-Time Adaptation											
TPT [NeurIPS'22]	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT [ICCV'23]	25.60	92.49	67.01	47.00	43.13	70.10	**87.23**	88.22	65.74	68.22	65.47
C-TPT [ICLR'24]	24.00	93.60	65.80	46.00	43.20	69.80	83.70	88.20	64.80	65.70	64.48
DPE† [NeurIPS'24]	25.83	94.49	67.82	46.16	57.75	72.49	86.30	90.17	68.42	73.04	68.25
Backpropagation-Free Test-Time Adaptation											
OGA† [CVPRW'25]	25.03	93.62	67.96	46.03	64.36	72.93	86.00	91.73	68.21	71.28	68.71
TDA† [CVPR'24]	25.38	94.01	67.43	45.47	61.15	71.67	86.11	89.72	67.68	70.89	67.95
BoostAdapter† [NeurIPS'24]	**27.21**	94.51	68.93	45.93	61.62	71.55	**87.23**	89.69	68.14	71.63	68.64
BCA† [CVPR'25]	25.21	93.87	66.53	44.22	49.59	70.47	85.91	88.72	66.49	67.90	65.89
PTA† [ICML'26]	25.45	**94.58**	68.61	47.86	61.55	74.71	86.32	91.20	69.10	73.33	69.27
SCA† [NeurIPS'25]	25.20	93.59	67.89	**48.70**	49.42	74.03	86.07	90.19	68.94	71.60	67.56
ADAPT† [NeurIPS'25]	25.33	93.10	68.83	46.74	**64.50**	72.98	85.62	91.85	68.97	71.08	68.90
DOTA† [NeurIPS'25]	26.18	94.20	69.64	47.73	62.51	**75.28**	86.91	**91.88**	69.74	72.69	69.68
ACME (Ours)	27.01	94.52	**69.78**	48.53	62.46	75.14	86.37	91.47	**69.99**	**74.10**	**69.94**

The frozen posterior preserves the pretrained zero-shot evidence, whereas $\widetilde{\mathbf{p}}_t$ incorporates the target-domain association estimated by ACME. Since labeled target data are unavailable for favoring either predictor, ACME combines them symmetrically using their normalized geometric mean:

$$[\mathbf{p}_t^{\text{final}}]_c = \frac{\sqrt{[\mathbf{p}_t^{\text{CLIP}}]_c [\widetilde{\mathbf{p}}_t]_c}}{\sum_{k=1}^{C} \sqrt{[\mathbf{p}_t^{\text{CLIP}}]_k [\widetilde{\mathbf{p}}_t]_k}}. \tag{16}$$

Substituting the two predictors into equation 16 gives the equivalent logit form

$$\boldsymbol{\ell}_t^{\text{final}} = \boldsymbol{\ell}_t^{\text{CLIP}} + \frac{\tau m_t}{2} \mathbf{s}_t^{\text{ACME}}. \tag{17}$$

The factor $1/2$ follows from treating the frozen and target-informed predictors symmetrically. Consequently, neither the maturity nor the fusion requires a dataset-selected adaptation parameter: m_t is determined solely by the observed sample count and feature dimension, while $1/2$ is induced by the symmetric geometric mean. Since $\mathbf{s}_t^{\text{ACME}} \in [-1, 1]^C$, each logit correction is bounded in magnitude by $\tau m_t/2$. For $t < 2$, the empirical moments do not define a nontrivial direction, and ACME reduces to the frozen prediction.

4 Experiments

4.1 Experimental Setup

Datasets. Following prior work (Shu et al., 2022; Karmanov et al., 2024; Huang et al., 2026), we evaluate ACME on two widely adopted TTA benchmarks: cross-domain generalization and out-of-distribution (OOD) generalization. The cross-domain benchmark comprises 10 diverse classification datasets: Flowers102 (Nilsback & Zisserman, 2008), OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), FGVC-Aircraft (Maji et al., 2013), Food101 (Bossard et al., 2014), EuroSAT (Helber et al., 2019), UCF101 (Soomro et al., 2012), DTD (Cimpoi et al., 2014), SUN397 (Xiao et al., 2010), and Caltech101 (Li et al., 2004). For OOD generalization, we evaluate on ImageNet-1K (Deng et al., 2009) and its four natural distribution shifts: ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a).

Table 2: **Comparison of ACME on ImageNet and OOD datasets using CLIP ViT-B/16. Bold** and underlined values indicate the best and second-best performance among all methods, respectively. For a fair comparison, all reproduced methods use the handcrafted prompt templates described in CLIP (Radford et al., 2021). Here, † denotes results reproduced using the officially released code. Other results are taken from the corresponding papers using the consistent prompt templates. OOD Average and Average denote the mean accuracy over 4 OOD datasets and all 5 datasets, respectively.

Method	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	OOD Average	Average
CLIP-ViT-B/16† [ICML'21]	68.79	50.68	62.36	77.68	48.26	59.74	61.55
Backpropagation-Based Test-Time Adaptation							
TPT [NeurIPS'22]	68.98	54.77	63.45	77.06	47.94	60.81	62.44
DiffTPT [ICCV'23]	70.30	55.68	65.10	75.00	46.80	60.65	62.58
C-TPT [ICLR'24]	68.50	51.60	62.70	76.00	47.90	59.55	61.34
DPE† [NeurIPS'24]	**71.17**	59.49	64.58	80.12	51.37	63.89	65.35
Backpropagation-Free Test-Time Adaptation							
OGA† [CVPRW'25]	69.37	59.99	63.99	80.86	50.13	63.74	64.87
TDA† [CVPR'24]	69.55	59.95	64.52	80.69	50.70	63.96	65.08
BoostAdapter† [NeurIPS'24]	69.38	**63.83**	**65.49**	80.69	51.55	**65.39**	**66.19**
BCA† [CVPR'25]	68.86	59.79	64.20	80.42	49.69	63.52	64.59
PTA† [ICML'26]	70.20	60.35	64.72	80.34	50.71	64.03	65.26
SCA† [NeurIPS'25]	70.22	60.91	64.53	80.88	51.27	64.40	65.56
ADAPT† [NeurIPS'25]	70.61	60.55	64.67	80.94	51.63	64.45	65.68
DOTA† [NeurIPS'25]	70.69	61.24	64.90	**81.21**	51.62	64.74	65.93
ACME (Ours)	70.79	60.67	64.84	80.62	**51.69**	64.45	65.72

Baselines. We compare ACME with zero-shot CLIP (Radford et al., 2021) and representative backpropagation-based methods, including TPT (Shu et al., 2022), DiffTPT (Feng et al., 2023), C-TPT (Yoon et al., 2024) and DPE (Zhang et al., 2024a). We further compare with recent backpropagation-free methods: OGA (Fuchs et al., 2025), TDA (Karmanov et al., 2024), BoostAdapter (Zhang et al., 2024b), BCA (Zhou et al., 2025), PTA (Huang et al., 2026), SCA (Guan et al., 2025), ADAPT (Zhang et al., 2025) and DOTA (Han et al., 2025).

Evaluation protocol. In line with prior work (Shu et al., 2022; Karmanov et al., 2024; Huang et al., 2026), we use publicly available pretrained CLIP models (Radford et al., 2021) for image recognition. Our main experiments employ ViT-B/16 (Dosovitskiy et al., 2021) as the visual encoder and a Transformer (Vaswani et al., 2017) as the text encoder, while results with ResNet-50 (He et al., 2016) are reported in Appendix A.1. For each reproduced method, we follow its officially released implementation. We consider the single-image test-time adaptation setting, with the batch size set to 1 to reflect online streaming evaluation. It is important to highlight that we adopt the ensemble prompting strategy described in CLIP (Radford et al., 2021), which uses handcrafted templates, ensuring a fair comparison among TTA methods. All experiments are conducted on a single NVIDIA RTX PRO 6000 Blackwell Server Edition MIG 1g.24gb instance.

4.2 Comparison with State-of-the-Art

Cross-domain generalization. Table 1 reports performance on the ten cross-domain datasets. ACME achieves the highest average accuracy of 69.94%, outperforming backpropagation-based TPT (Shu et al., 2022), DiffTPT (Feng et al., 2023), and our reproduced DPE (Zhang et al., 2024a) baseline by 4.84%, 4.47%, and 1.69%, respectively. It also surpasses all evaluated backpropagation-free methods, including DOTA (Han et al., 2025) and PTA (Huang et al., 2026), by 0.26% and 0.67%, respectively. Notably, ACME is tuning-free across datasets, whereas several competing methods employ dataset-specific hyperparameter configurations. These results demonstrate that ACME effectively adapts across diverse target domains by accumulating graded evidence from the complete test stream, rather than relying on confidence-filtered per-class caches. Additional experiments with CLIP ResNet-50 further demonstrate ACME's effectiveness on a convolutional backbone, yielding the highest cross-domain average and ImageNet-1K accuracy among the evaluated methods (Appendix A.1).

ImageNet and OOD generalization. Table 2 reports performance on ImageNet-1K and its four natural distribution shifts. Among all evaluated methods, ACME achieves the second-highest ImageNet-1K accuracy of 70.79%, improving zero-shot CLIP (Radford et al., 2021) by 2.00%. While DPE (Zhang et al., 2024a) obtains the highest accuracy, it performs gradient-based adaptation over a batch containing the original image and 63 augmented views for every test image and therefore incurs substantially greater inference cost. Across the four OOD datasets, ACME achieves an average accuracy of 64.45%. DOTA (Han et al., 2025) and BoostAdapter (Zhang et al., 2024b) exceed this result by 0.29% and 0.94%, respectively, but their released implementations employ dataset-specific hyperparameter configurations and maintain expensive class-conditioned online states through full covariance matrices and sample caches, respectively. In contrast, ACME is tuning-free across datasets and requires neither backpropagation nor per-example caches, providing a favorable balance between ImageNet accuracy, OOD robustness, and deployment efficiency.

Runtime and memory efficiency. Table 3 compares the efficiency of various online test-time adaptation methods on the ImageNet-1K validation set. ACME processes all 50,000 images in 6.8 minutes, retaining 88.07% of frozen CLIP's throughput while improving performance by 2.00%. ACME introduces only approximately 3 MiB of memory overhead compared with frozen CLIP (652 MiB), which is substantially lower than other methods.

Table 3: **End-to-end efficiency on ImageNet-1K.**

Method	Testing Time	Rel. Speed	Memory	Gain
CLIP	6.0 min	100.00%	652 MiB	0.00
DPE	4 h 59 min	1.92%	6,436 MiB	**+2.38**
OGA	1 h 1 min	9.85%	2,450 MiB	+0.58
BoostAdapter	13.9 min	43.20%	730 MiB	+0.59
TDA	11.9 min	50.24%	728 MiB	+0.76
PTA	7.0 min	86.03%	659 MiB	+1.41
SCA	14.4 min	41.67%	794 MiB	+1.43
ADAPT	1 h 12 min	8.32%	793 MiB	+1.82
DOTA	29.6 min	20.19%	4,720 MiB	+1.90
ACME (Ours)	**6.8 min**	**88.07%**	**655 MiB**	+2.00

ACME is 1.75×, 2.04×, 4.35×, 10.59×, and 43.97× faster than TDA (Karmanov et al., 2024), BoostAdapter (Zhang et al., 2024b), DOTA (Han et al., 2025), ADAPT (Zhang et al., 2025), and DPE (Zhang et al., 2024a), respectively. These results demonstrate that ACME offers a favorable accuracy-efficiency trade-off, delivering substantial adaptation gains while preserving near-frozen-CLIP inference speed and memory consumption.

4.3 Ablation Studies

Effect of feature centering and scaling. We study the importance of feature centering and coordinate-wise scaling in the ACME cross-moment readout. Given a visual feature $\mathbf{x}_t$, we consider four transformations:

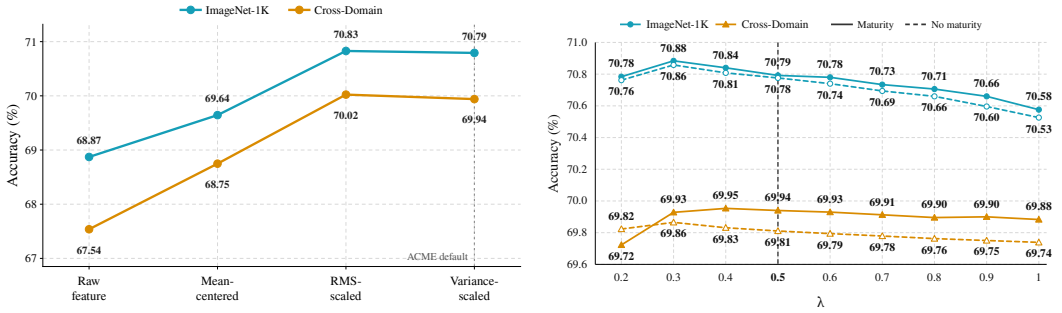
$$\mathbf{z}_t = \begin{cases} \mathbf{x}_t, & \text{Raw Feature,} \\ \mathbf{x}_t - \boldsymbol{\mu}_t^x, & \text{Mean-Centered,} \\ (\mathbf{x}_t - \boldsymbol{\mu}_t^x) \oslash \sqrt{\mathbf{v}_t^x + \boldsymbol{\mu}_t^x \odot \boldsymbol{\mu}_t^x}, & \text{RMS-Scaled,} \\ (\mathbf{x}_t - \boldsymbol{\mu}_t^x) \oslash \sqrt{\mathbf{v}_t^x}, & \text{Variance-Scaled (ACME),} \end{cases} \tag{18}$$

The class-specific cross-moment directions remain unscaled for the Raw Feature and Mean-Centered variants, whereas the RMS- and variance-scaled variants apply the same coordinate-wise scaling to both the directions and the query feature.

Figure 3a shows that both mean centering and coordinate-wise scaling are important: omitting either operation leads to a clear degradation in performance on ImageNet-1K and the cross-domain benchmark. The RMS-scaled variant marginally outperforms the variance-scaled ACME readout by 0.04% and 0.08%, respectively, indicating that both scaling strategies are similarly effective. RMS scaling uses the uncentered second moment and may therefore avoid amplifying coordinates with small centered variance but relatively large mean magnitude. We retain variance scaling in ACME because it produces standardized coordinates and preserves the direct correlation-based interpretation of the cross-moment readout.

Fusion weight and maturity. We also study the sensitivity of the target-informed predictor,

$$\boldsymbol{\ell}_t^{\text{final}} = \boldsymbol{\ell}_t^{\text{CLIP}} + \lambda \cdot \tau m_t \mathbf{s}_t^{\text{ACME}}, \tag{19}$$



(a) (b)

Figure 3: **Ablation studies.** (a) Effect of feature centering and coordinate-wise scaling; (b) sensitivity to the fusion weight λ with and without maturity scaling on ImageNet-1K and Cross-Domain.

to the fusion coefficient λ and maturity factor $m_t = t/(t + D)$. Figure 3b reports results for $\lambda \in [0.2, 1.0]$, with and without the maturity factor, on ImageNet-1K and the cross-domain benchmark. Performance remains stable across a broad range of coefficients, demonstrating robustness to the choice of λ. The maturity factor further improves performance slightly by suppressing over-corrections when the streaming statistics are still unreliable.

Transductive setting. We further examine whether ACME can exploit complete target-stream statistics under a two-pass, label-free protocol. During the first pass, each adaptation method constructs its state from the complete unlabeled test stream. This state is subsequently frozen, and predictions are obtained in a second pass over the same samples without further updates. For ACME, we fix the maturity factor after observing the complete stream as $m_N = N/(N + D)$, where N is the number of test samples. As shown in Table 4, ACME achieves the highest cross-domain accuracy while remaining competitive on ImageNet. DOTA (Han et al., 2025) obtains the highest ImageNet accuracy, reflecting its richer class-specific covariance estimates, but incurs substantially greater memory and inference costs.

Table 4: **Two-pass transductive accuracy (%).**

Method	ImageNet	Cross-Domain
TDA	69.45	67.71
PTA	70.13	69.79
ADAPT	71.16	70.11
DOTA	**71.43**	70.41
ACME (Ours)	71.17	**70.48**

5 Conclusions

We introduced Adaptive Cross-Moment Estimation (ACME), a tuning-free and backpropagation-free method for online adaptation of frozen vision-language models. Its central advantage is that it requires no validation-based selection of adaptation hyperparameters. ACME uses the complete frozen posterior of each test sample to update a compact posterior-feature cross-covariance through exact streaming recurrences, without storing previous examples or assigning hard pseudo-labels. Maturity-aware fusion then combines the resulting correction with the frozen predictions. Across cross-domain and distribution-shift benchmarks, ACME achieves strong adaptation performance while retaining near frozen-CLIP inference speed and introducing only approximately 3 MiB of additional memory on ImageNet-1K. Evaluations across visual backbones, prompt formulations, corruption shifts, and random stream permutations further support its robustness (Appendix A).

Our evaluation focuses on closed-set image classification. More broadly, posterior-feature cross-covariance provides a general interface between model predictions and learned representations. Promising future directions include extending streaming cross-moment estimation to dense prediction tasks, such as semantic segmentation, and to temporal multimodal settings involving video-language or audio-language models.

AI Use Statement

We used ChatGPT to assist with identifying relevant literature, polishing the language and clarity of the manuscript, checking experimental configurations, and debugging evaluation code. All

suggested references were verified against their original publications, and all AI-assisted code and configuration suggestions were reviewed and validated by the authors. The research methodology, experimental analysis, and scientific conclusions were developed by the authors, who take full responsibility for the content of the final manuscript.

Reproducibility Statement

The main paper provides a complete description of ACME, including its streaming updates, standardized readout, and maturity-aware fusion. The experimental section documents the datasets, model backbones, baselines, evaluation protocol, and efficiency measurements, while further experimental details and evaluations are provided in Appendix A. The attached supplementary material includes the complete PyTorch implementation, dependency specifications, dataset preparation instructions, and executable scripts for the cross-domain and ImageNet distribution-shift benchmarks. These resources provide the configurations and implementation details required to reproduce the reported results.

References

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Qiyuan Dai and Sibei Yang. Free on the fly: Enhancing flexibility in test-time adaptation with online EM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9538–9548, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dai_Free_on_the_Fly_Enhancing_Flexibility_in_Test-Time_Adaptation_with_CVPR_2025_paper.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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 *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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*, pp. 2704–2714, 2023.

Clément Fuchs, Maxime Zanella, and Christophe De Vleeschouwer. Online gaussian test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 128–137, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/MULA2025/html/Fuchs_Online_Gaussian_Test-Time_Adaptation_of_Vision-Language_Models._CVPRW_2025_paper.html.

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. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x. URL https://doi.org/10.1007/s11263-023-01891-x.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Zenghao Guan, Yucan Zhou, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 6616–6639. Curran Associates, Inc., 2025. doi: 10.52202/085713-0230. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/09eadc995e2a0293f1bc622a3e68fb18-Paper-Conference.pdf.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. DOTA: Distributional test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/d24189d251aa453b6d2eb3421ec48bcc-Abstract-Conference.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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. doi: 10.1109/JSTARS.2019.2918242.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

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*, pp. 8340–8349, 2021a.

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*, pp. 15262–15271, 2021b.

Zhaohong Huang, Yuxin Zhang, Wenjing Liu, Fei Chao, and Rongrong Ji. Prototype-based test-time adaptation of vision-language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ejmfnVdW2X.

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 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

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*, pp. 14162–14171, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Karmanov_Efficient_Test-Time_Adaptation_of_Vision-Language_Models_CVPR_2024_paper.html.

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*, pp. 19113–19122, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Fei-Fei Li. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Fei-Fei Li, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178–178, 2004.

Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15691–15701, October 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/5bf2b802e24106064dc547ae9283bb0c-Abstract-Conference.html.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops*, 2016.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pp. 2058–2065, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962. doi: 10.1080/00401706.1962.10490022. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1962.10490022.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 8636–8660, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/23ff02034404b65776080cbf7148addd-Paper-Conference.pdf.

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*, pp. 10899–10909, 2023.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://papers.nips.cc/paper_files/paper/2024/hash/38b787fc530d0b31825827e2cc306656-Abstract-Conference.html.

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 Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 493–510, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19833-5.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 67795–67825. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2165. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7d60bfd8458b67acbbaf18b892338d00-Paper-Conference.pdf.

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*, pp. 28718–28728, 2024c.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 76280–76310. Curran Associates, Inc., 2025. doi: 10.52202/085713-2562. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6e9a0a72da9b76c3ebc8cc33ff10ac29-Paper-Conference.pdf.

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*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022b. doi: 10.1007/s11263-022-01653-1.

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*, pp. 29999–30009, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_Bayesian_Test-Time_Adaptation_for_Vision-Language_Models_CVPR_2025_paper.html.

A APPENDIX

Table 5: **Comparison of ACME on Cross-Domain Datasets using ResNet-50 as the backbone. Bold** and underlined values indicate the best and second-best performance among the evaluated methods, respectively. For reproduced methods, we use the handcrafted prompt templates described in CLIP (Radford et al., 2021). Here, † denotes results reproduced using the officially released implementations, while other results are taken from their respective papers.

Method	Aircraft	Caltech	Cars	DTD	EuroSAT	Flower	Food101	Pets	SUN397	UCF101	Average
CLIP-ResNet-50† [ICML'21]	16.92	87.87	56.27	42.97	36.19	65.98	77.43	85.75	61.24	61.96	59.26
Backpropagation-Based Test-Time Adaptation											
TPT [NeurIPS'22]	17.58	87.02	58.46	40.84	28.33	62.69	74.88	84.49	61.46	60.82	57.66
DiffTPT [ICCV'23]	17.60	86.89	60.71	40.72	41.04	63.53	**79.21**	83.40	62.72	62.67	59.85
DPE† [NeurIPS'24]	17.10	**89.82**	57.83	44.62	39.56	67.44	77.63	86.10	63.17	64.82	60.81
Backpropagation-Free Test-Time Adaptation											
OGA† [CVPRW'25]	16.74	87.67	58.72	44.33	43.25	66.75	76.40	86.75	63.55	63.89	60.80
TDA† [CVPR'24]	16.98	89.57	57.68	43.32	39.63	68.41	77.73	85.77	62.35	63.65	60.51
BoostAdapter† [NeurIPS'24]	**18.72**	88.36	59.99	44.33	40.09	67.48	78.66	86.07	63.14	64.13	61.10
PTA† [ICML'26]	16.77	88.32	59.08	44.33	49.37	68.25	76.94	86.70	63.18	66.30	61.92
SCA† [NeurIPS'25]	17.61	88.64	58.64	45.57	48.47	**69.18**	77.38	85.88	63.17	66.88	62.14
ADAPT† [NeurIPS'25]	17.25	88.24	58.85	45.15	43.95	67.56	76.57	87.57	64.06	65.08	61.43
ACME (Ours)	17.52	88.97	**60.78**	**46.10**	**50.69**	68.98	77.19	**87.63**	**64.66**	**68.04**	**63.06**

A.1 EVALUATION ON RESNET-50 BACKBONE

To further evaluate the architectural robustness of ACME, we conduct additional experiments using the ResNet-50 (He et al., 2016) as the visual backbone. For a fair comparison, all the reproduced methods are evaluated using the same handcrafted prompt templates described in CLIP (Radford et al., 2021) and online evaluation protocol.

As reported in Tables 5 and 6, ACME achieves the highest average accuracy on the cross-domain benchmark among the evaluated methods. On the ImageNet and OOD benchmarks, ACME obtains the best ImageNet-1K performance while remaining competitive across the four natural distribution shifts. These results highlight the architectural robustness of ACME, demonstrating that its streaming cross-moment estimation and maturity-aware fusion remain effective across both the transformer-based ViT-B/16 (Dosovitskiy et al., 2021) and the convolutional ResNet-50 (He et al., 2016) backbone configurations, preserving both strong performance and computational efficiency.

DOTA (Han et al., 2025) is excluded from this comparison because its officially released implementation supports only ViT-B/16, while the paper does not provide a reproducible ResNet-50 configuration for the complete cross-domain and OOD benchmarks.

A.2 ROBUSTNESS TO TEST-STREAM ORDER

In real-world streaming scenarios, the order in which unlabeled test samples arrive is unpredictable. Because the adaptation state evolves sequentially, early observations may disproportionately influence the subsequent adaptation trajectory. To assess the stability of ACME, we evaluate it over ten random permutations of the test stream. As reported in Table 7, ACME exhibits consistent performance, with standard deviations of only 0.07% on ImageNet-1K and 0.08% on the cross-domain benchmark. Its streaming cross-moment estimates progressively aggregate information, while maturity-aware fusion limits the influence of less reliable estimates early in the stream. These properties are consistent with the observed low sensitivity to stream order. The results also indicate that ACME's gains are not attributable to a favorable sample sequence.

Table 6: **Comparison on ImageNet and OOD Datasets using ResNet-50 as the backbone. Bold** and underlined values indicate the best and second-best performance among the evaluated methods, respectively. For reproduced methods, we use the handcrafted prompt templates described in CLIP (Radford et al., 2021). Here, † denotes results reproduced using the officially released implementations, while other results are taken from their respective papers.

Method	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	OOD Average	Average
CLIP-ResNet-50† [ICML'21]	60.34	23.87	53.34	60.72	35.48	43.35	46.75
Backpropagation-Based Test-Time Adaptation							
TPT [NeurIPS'22]	60.74	26.67	54.70	59.11	35.09	43.89	47.26
DiffTPT [ICCV'23]	60.80	31.06	55.80	58.80	37.10	45.69	48.71
DPE† [NeurIPS'24]	62.07	30.01	55.53	62.85	**38.96**	46.84	49.88
Backpropagation-Free Test-Time Adaptation							
OGA† [CVPRW'25]	61.08	29.88	52.62	61.79	36.69	45.25	48.41
TDA† [CVPR'24]	61.30	30.69	54.96	**62.94**	37.62	46.55	49.50
BoostAdapter† [NeurIPS'24]	61.09	**35.61**	**56.41**	62.72	38.84	**48.40**	**50.93**
PTA† [ICML'26]	60.49	29.07	54.66	60.56	36.81	45.28	48.32
SCA† [NeurIPS'25]	61.25	31.27	54.97	62.67	37.86	46.69	49.60
ADAPT† [NeurIPS'25]	61.78	30.25	54.84	61.95	37.25	46.07	49.21
ACME (Ours)	**62.16**	30.44	55.69	62.02	38.82	46.74	49.83

Table 7: **Sensitivity of ACME to test-stream order.** We report the results under 10 random permutations of the test stream. Average denotes the mean accuracy across the ten cross-domain datasets. Std. denotes the population standard deviation across the ten permutations.

Dataset	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7	Exp. 8	Exp. 9	Exp. 10	Mean ± Std.
Aircraft	27.03	26.67	27.18	27.15	27.03	26.73	26.58	26.34	26.76	27.33	26.88 ± 0.30
Caltech101	94.60	94.93	94.08	94.52	94.48	94.52	94.08	94.60	94.16	93.87	94.39 ± 0.31
Stanford Cars	69.94	69.52	69.53	69.84	70.07	69.85	70.39	69.88	69.73	69.90	69.87 ± 0.24
DTD	48.88	48.64	47.70	48.23	49.23	49.29	49.29	48.35	48.70	48.88	48.72 ± 0.49
EuroSAT	62.51	62.48	62.65	62.41	62.27	63.01	62.75	63.43	62.35	62.88	62.67 ± 0.34
Flowers102	75.11	74.95	75.36	75.03	75.27	75.60	74.91	75.19	74.50	74.91	75.08 ± 0.28
Food101	86.34	86.32	86.39	86.40	86.39	86.34	86.33	86.46	86.35	86.50	86.38 ± 0.06
Oxford Pets	91.74	91.17	91.77	91.28	91.41	91.58	91.77	91.44	91.71	91.20	91.51 ± 0.23
SUN397	69.97	70.15	70.06	69.73	70.07	69.85	69.95	70.22	70.13	69.85	70.00 ± 0.15
UCF101	73.75	74.41	74.02	74.73	73.59	74.25	74.39	74.15	73.88	74.78	74.20 ± 0.38
Average	69.99	69.92	69.87	69.93	69.98	70.10	70.04	70.01	69.83	70.01	**69.97 ± 0.08**
ImageNet	70.79	70.82	70.71	70.77	70.71	70.80	70.58	70.76	70.72	70.68	**70.73 ± 0.07**

A.3 Robustness to Prompt Templates

To further investigate the robustness of ACME to textual guidance, we evaluate it under three prompt formulations: Vanilla template, handcrafted CLIP templates (Radford et al., 2021), and fine-grained GPT-generated class descriptions following DPE (Zhang et al., 2024a). As reported in Table 8, ACME consistently achieves the highest performance across all three settings, with average accuracies of 68.83%, 69.94%, and 71.03% under Vanilla, CLIP-template, and GPT prompts, respectively.

Notably, the improvement over the corresponding zero-shot CLIP classifier is largest under the Vanilla setting, reaching $+5.05\%$ when textual guidance is least informative. While methods such as SCA (Guan et al., 2025) and ADAPT (Zhang et al., 2025) exhibit sharper degradation under sparse guidance, ACME retains strong discriminative performance with only a single generic template. It also preserves gains of $+4.22\%$ and $+2.60\%$ with the richer CLIP-template and GPT formulations, showing that its adaptation remains effective even with more informative textual priors. Overall, these results establish the robustness of ACME across substantial variations in prompt granularity.

Table 8: **Sensitivity to different prompt formulations.** We report accuracy (%) on ten cross-domain datasets. **Bold** and underlined values indicate the best and second-best results within each prompt setting, respectively. The GPT formulation follows the DPE protocol and combines handcrafted CLIP templates with fine-grained CuPL class descriptions (Pratt et al., 2023).

Method	Aircraft	Caltech	Cars	DTD	EuroSAT	Flower	Food101	Pets	SUN397	UCF101	Average
Vanilla: "a photo of a {class}."											
CLIP† [ICML'21]	23.85	93.31	65.55	44.33	42.07	67.44	85.25	88.20	62.56	65.27	63.78
DPE† [NeurIPS'24]	26.07	94.20	67.69	46.93	51.81	69.83	86.01	88.77	66.25	69.60	66.72
OGA† [CVPRW'25]	23.79	93.18	67.94	46.99	53.54	69.87	85.76	89.53	65.35	68.46	66.44
TDA† [CVPR'24]	24.48	93.71	66.68	46.04	55.05	69.83	85.77	88.63	65.83	68.52	66.45
BoostAdapter† [NeurIPS'24]	25.92	**94.24**	68.14	46.16	55.69	70.04	**86.92**	88.88	66.20	69.73	67.19
PTA† [ICML'26]	24.48	94.08	68.52	47.34	56.40	72.63	85.97	89.42	67.19	69.92	67.60
SCA† [NeurIPS'25]	25.14	93.39	67.42	48.35	54.67	73.65	85.63	89.07	67.40	69.63	67.43
ADAPT† [NeurIPS'25]	24.21	92.78	68.52	47.34	54.21	70.36	85.43	90.02	66.78	68.62	66.83
DOTA† [NeurIPS'25]	25.65	93.91	69.43	47.58	**60.17**	72.76	86.68	90.32	67.97	70.47	68.50
ACME (Ours)	**26.25**	94.16	**69.83**	**49.47**	57.60	**73.89**	86.25	**90.46**	**68.65**	**71.72**	**68.83**
CLIP Template: ensemble of handcrafted templates											
CLIP† [ICML'21]	24.90	93.87	66.17	44.21	48.46	70.77	85.89	89.10	66.27	67.54	65.72
DPE† [NeurIPS'24]	25.83	94.49	67.82	46.16	57.75	72.49	86.30	90.17	68.42	73.04	68.25
OGA† [CVPRW'25]	25.03	93.62	67.96	46.03	64.36	72.93	86.00	91.73	68.21	71.28	68.71
TDA† [CVPR'24]	25.38	94.01	67.43	45.47	61.15	71.67	86.11	89.72	67.68	70.89	67.95
BoostAdapter† [NeurIPS'24]	**27.23**	94.53	68.92	45.95	61.61	71.48	**87.24**	89.68	68.12	71.73	68.65
PTA† [ICML'26]	25.45	**94.58**	68.61	47.86	61.55	74.71	86.32	91.20	69.10	73.33	69.27
SCA† [NeurIPS'25]	25.20	93.59	67.89	**48.70**	49.42	74.03	86.07	90.19	68.94	71.60	67.56
ADAPT† [NeurIPS'25]	25.33	93.10	68.83	46.74	**64.50**	72.98	85.62	91.85	68.97	71.08	68.90
DOTA† [NeurIPS'25]	26.18	94.20	69.64	47.73	62.51	**75.28**	86.91	**91.88**	69.74	72.69	69.68
ACME (Ours)	27.01	94.52	**69.78**	48.53	62.46	75.14	86.37	91.47	**69.99**	**74.10**	**69.94**
GPT: CLIP templates with fine-grained CuPL descriptions											
CLIP† [ICML'21]	27.27	94.16	66.32	53.25	55.60	72.92	85.90	90.30	68.08	70.47	68.43
DPE† [NeurIPS'24]	28.86	**95.21**	67.52	54.49	58.02	74.95	86.26	91.28	69.87	72.51	69.90
OGA† [CVPRW'25]	27.57	93.55	67.73	54.37	57.69	74.42	85.95	91.80	69.83	73.57	69.65
TDA† [CVPR'24]	29.16	94.36	66.86	54.79	57.09	73.81	86.19	91.06	69.15	72.30	69.48
BoostAdapter† [NeurIPS'24]	**30.48**	95.09	68.45	55.38	56.93	73.93	**87.28**	91.17	69.58	72.96	70.12
PTA† [ICML'26]	28.71	94.89	68.26	55.56	54.77	77.22	86.36	91.44	70.49	74.02	70.17
SCA† [NeurIPS'25]	28.62	94.44	67.77	**56.15**	57.11	76.74	86.03	91.66	70.20	72.80	70.15
ADAPT† [NeurIPS'25]	28.89	93.75	68.83	54.91	57.47	74.79	85.46	91.82	70.56	73.04	69.95
DOTA† [NeurIPS'25]	29.16	94.56	69.69	55.69	**59.40**	77.06	87.01	**92.48**	71.01	74.17	71.02
ACME (Ours)	29.58	94.81	**69.71**	55.97	56.30	**78.85**	86.44	92.37	**71.15**	**75.15**	**71.03**

A.4 Online Adaptation Dynamics

We further analyze how adaptation performance evolves over the unlabeled test stream. Figure 4 reports cumulative online accuracy for frozen CLIP (Radford et al., 2021), TDA (Karmanov et al., 2024), PTA (Huang et al., 2026), DOTA (Han et al., 2025), and ACME. We observe that ACME rapidly establishes a clear advantage on FGVC-Aircraft (Maji et al., 2013), DTD (Cimpoi et al., 2014), and UCF101 (Soomro et al., 2012), and consistently preserves this advantage as adaptation proceeds. This indicates that its improvements emerge early and are sustained throughout the stream, rather than being driven by a small subset of later observations. On the substantially longer ImageNet-1K stream (Deng et al., 2009), the improvement develops more gradually as the accumulated statistics mature, with ACME progressively separating from the competing methods. Moreover, its trajectories remain stable without exhibiting late-stage performance deterioration. These results demonstrate that ACME supports effective and reliable information accumulation throughout online adaptation while limiting the propagation of early prediction errors.

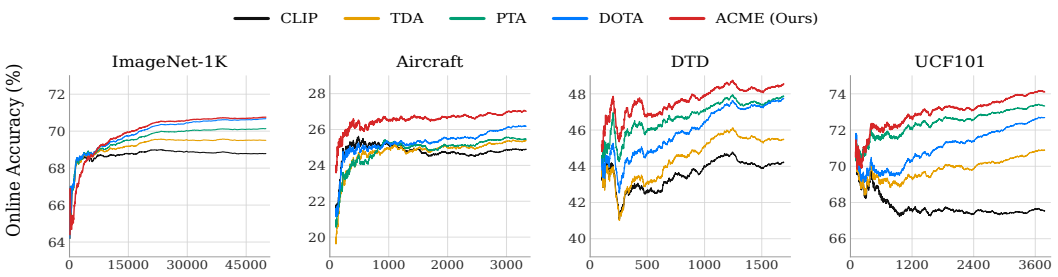


Figure 4: **Online adaptation dynamics.** Cumulative accuracy over the unlabeled test stream.

A.5 Evaluation on ImageNet Corruptions

We further evaluate ACME on ImageNet-C (Hendrycks & Dietterich, 2019) to assess its robustness under common corruption shifts. ImageNet-C comprises 15 corruption types spanning noise, blur, weather, and digital distortions. We evaluate all corruptions at severity level 5, corresponding to the most severe setting. We consider both a single Vanilla prompt and an ensemble of handcrafted CLIP templates (Radford et al., 2021). All methods use the same ViT-B/16 backbone and test-stream order, with the adaptation state reset before each corruption.

As shown in Table 9, ACME demonstrates strong adaptation performance under both prompt formulations. It improves over frozen CLIP by 2.58% and 1.50% under the Vanilla and CLIP-template settings, respectively, while outperforming PTA by 1.80% and 1.23%. TDA obtains marginally higher average accuracy, exceeding ACME by only 0.10% and 0.43% under the two settings. Despite this small difference, ACME is substantially more efficient in both inference time and memory consumption, providing a more favorable balance between corruption robustness and computational cost. These results demonstrate that ACME remains effective under severe and diverse corruption shifts while preserving its lightweight adaptation characteristics.

Table 9: **Comparison on ImageNet-C at corruption severity level 5 using ViT-B/16.**

Method	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG	Average
Vanilla Prompt																
CLIP† [ICML'21]	13.25	14.16	13.50	24.25	15.73	24.45	22.62	33.07	31.07	37.61	55.61	17.10	13.45	33.01	33.67	25.50
TDA† [CVPR'24]	15.95	17.14	16.54	**26.27**	**18.00**	**27.03**	**24.80**	**35.90**	**34.13**	**40.51**	57.80	**19.62**	**16.36**	**36.46**	36.21	**28.18**
PTA† [ICML'26]	13.91	15.11	14.46	24.84	16.23	24.90	22.11	33.98	31.93	38.87	57.60	17.17	13.96	34.21	34.96	26.28
ACME (Ours)	**16.31**	**17.24**	**17.23**	26.21	17.50	26.64	24.13	35.58	33.38	40.48	**59.17**	18.66	16.06	36.20	**36.37**	28.08
CLIP Template Ensemble																
CLIP† [ICML'21]	15.15	16.31	15.24	25.84	16.86	26.35	24.43	34.55	32.99	39.14	57.78	18.45	14.71	35.65	35.79	27.28
TDA† [CVPR'24]	**16.99**	**18.09**	17.77	**27.34**	**18.63**	**27.95**	**25.69**	**36.96**	**35.27**	**41.32**	59.35	**20.51**	**17.29**	**37.72**	**37.28**	**29.21**
PTA† [ICML'26]	15.41	16.47	15.82	26.08	17.02	26.27	23.49	35.02	33.13	39.83	58.93	18.06	15.21	36.09	36.41	27.55
ACME (Ours)	16.93	17.94	**17.96**	27.03	17.84	27.12	24.79	36.38	34.15	41.17	**60.13**	19.27	16.74	37.08	37.24	28.78