# THE SPACE BETWEEN VIEWS: A SIMPLEX APPROACH TO TEST-TIME ADAPTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We hypothesize that the region spanned by augmented views contains useful structure for test-time adaptation (TTA). Can TTA exploit it? Existing methods typically treat augmented views as a finite set, aggregating, filtering, or regularizing their predictions. We propose SimplexTTA, which models the convex hull of augmented test features and minimizes the expected adaptation loss over the resulting simplex. To make this objective practical, we derive a second-order approximation around the simplex barycenter. This, in turn, yields a curvature-weighted correction that captures the interaction between feature spread and the local curvature of the loss. The resulting objective can directly replace the standard adaptation loss in existing gradient-based TTA methods with minimal modification. Across multiple image classification benchmarks, including corruption benchmarks, SimplexTTA consistently improves state-of-the-art TTA methods, providing empirical support for our hypothesis.

## 1 INTRODUCTION

We propose *SimplexTTA*, a framework for test-time adaptation that minimizes the expected loss over convex combinations of augmented test features. Test-time adaptation (TTA) aims to improve a pretrained model after deployment using only unlabeled test data (Liang et al., 2025). TTA is particularly important for vision-language models, which are deployed in open and dynamic environments where the input data distribution may differ from training data distributions due to corruptions, domain shifts, or changes in visual style. In the standard TTA setting studied here, adaptation faces two difficulties: no labels are available to supervise the update, and each update relies on a single test image or a small batch, providing limited information about the test distribution.

A prominent approach to address these difficulties is to generate multiple augmented views of each test input and use their predictions in a self-supervised objective that encourages confident and consistent predictions across views. These views are correlated observations of the same image, rather than additional independent test examples. For example, MEMO updates the model by minimizing the entropy of the average predicted class distribution across augmented views (Zhang et al., 2022). TPT selects confident views and minimizes the entropy of their average prediction to tune the text prompt in a vision-language model (Shu et al., 2022). Augmented views can also be used without parameter updates: ZERO selects confident views and aggregates their predicted classes by voting (Farina et al., 2024). These methods use a finite set of view predictions for adaptation or aggregation, but do not explicitly optimize a loss over convex combinations of the augmented features.

With no labels and only a small number of test examples per update, objectives constructed from the model's predictions can provide an unreliable adaptation signal. For example, entropy-based selection can be unreliable under biased test distributions (Lee et al., 2024), while erroneous pseudo-labels can reinforce incorrect predictions (Mishra et al., 2026). Augmentations help mitigate this issue, but treating them only as a finite set of correlated views does not fully exploit the local region of feature space spanned by those views. We therefore hypothesize that accounting for the loss throughout the local feature region spanned by augmented views provides a more reliable adaptation signal than optimizing only the original sample or the sampled views.

Specifically, *SimplexTTA* is a general framework for test-time adaptation over the augmentation simplex, with feature representations of multiple augmented views of a test sample as vertices of the



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.

simplex. Instead of minimizing a test-time loss only on the original sample, the mean prediction, or a finite set of augmentations, *SimplexTTA* minimizes the expected loss over all convex combinations of the augmented features. This approach converts a finite collection of views into a continuous local adaptation space. The resulting objective encourages the model to perform well not only at the observed augmentations, but also throughout the local feature neighbourhood that they define.

We model convex combinations of the augmented features by drawing their mixing coefficients from a Dirichlet distribution over the coefficient simplex. For generic nonlinear test-time losses, the resulting expectation generally lacks a convenient closed-form expression. We therefore derive a second-order approximation around the simplex barycenter. We show that this second-order approximation decomposes the expected simplex loss into the loss evaluated at the mean representation and a covariance-weighted curvature term.

The first term adapts the model using the mean representation of the augmented views, while the second term accounts for the spread of the views under the local curvature of the loss. Therefore, *SimplexTTA* provides a principled way to use the augmentations not only as multiple discrete samples, but as a continuous feature-space region..

The proposed formulation is general. It applies to differentiable test-time objectives defined in the visual embedding space, including standard entropy minimization objectives, contrastive and pdeudo-label-based objectives. We show that for a broad class of differentiable test-time losses, *SimplexTTA* objective yields a compact approximation involving the Hessian-vector prodructs, avoiding explicit construction of the Hessian. Algorithm 1 summarizes the selected representations, loss evaluation, and parameter update; **??** details the four base methods and their simplex extensions, including HVP computation and gradient detachment. This makes the method compatible with deep learning frameworks such as pytorch that support automatic differentiation and allows it to be instantiated as a loss lift for existing gradient-based TTA methods.

We hypothesize that accounting for the loss over the local region spanned by augmented features can improve adaptation under distribution shift. The second-order objective incorporates both the loss at the mean representation and its curvature along directions of augmentation-induced variation. This provides a way to use the structure of the selected views when adapting from limited unlabeled data. The benefit may depend on the underlying test-time objective and the quality of the augmentations; i.e., mixtures of misleading views can still support incorrect predictions. In our CLIP ViT-B/16 evaluation, *SimplexTTA* raises the 13-dataset clean mean by $0.68$, $1.96$, and $1.46$ points for CLIPTTA, BATCLIP, and BITTA, respectively. It improves all 16 matched comparisons on four natural ImageNet shifts and all six BATCLIP/BITTA corruption-suite comparisons, with corruption gains of $1.01$ - $2.70$ points. TENT under corruption and all four evaluated methods on OfficeHome show that these gains depend on the base objective and distribution shift.

Our main contributions are threefold;

1. We formulate augmentation-aware TTA over the continuous feature-space simplex; the convex hull spanned by a finite set of augmented test representations.
2. We derive a second-order approximation of the expected simplex loss as a barycenter term plus a covariance-weighted curvature correction, and evaluate the correction using Hessian-vector products without explicitly constructing the Hessian.
3. We instantiate the lift in four gradient-based CLIP TTA objectives and empirically evaluate across clean classification, cross-domain transfer, natural distribution shifts, and corruption benchmarks.
4. In a controlled quadratic setting, we identify a covariance-matching mechanism: simplex adaptation retains an intermediate level of augmentation covariance and minimizes deployment risk when that covariance matches the target variation.

## 2 Related Work

**Test-time adaptation.** Test-time adaptation updates a pretrained model using unlabeled samples from the deployment distribution. TENT minimizes prediction entropy while adapting normalization parameters (Wang et al., 2021), whereas EATA filters unreliable or redundant samples and regularizes important parameters to limit forgetting (Niu et al., 2022). Augmentation provides an

additional source of test-time supervision: MEMO minimizes the entropy of the prediction averaged over augmented views of an individual input (Zhang et al., 2022). These approaches construct adaptation signals from test samples, selected views, or aggregated predictions. Reliability remains a central concern: SAR filters unreliable samples and promotes flatter solutions (Niu et al., 2023), while DeYO shows that entropy alone can be an unreliable selection signal under biased test distributions and incorporates a complementary prediction-difference criterion (Lee et al., 2024). Our objective instead evaluates the base adaptation loss over a continuous local region spanned by augmented representations.

**Test-time adaptation of vision-language models.** CLIP-style vision-language models transfer to downstream classification through image–text similarity (Radford et al., 2021). TPT adapts soft prompt tokens using low-entropy augmented views (Shu et al., 2022). Subsequent prompt-based approaches improve calibration through text-feature dispersion (Yoon et al., 2024), generate diverse diffusion-based views (Feng et al., 2023), or learn from contrasting augmented views with online and target prompts (Ma et al., 2023). TPS instead shifts cached class prototypes directly in the shared embedding space (Sui et al., 2025), while WATT adapts CLIP across prompt templates and averages the resulting model weights (Osowiechi et al., 2024). Training-free alternatives include confidence-based view selection and voting in ZERO (Farina et al., 2024) and a dynamic positive–negative feature cache in TDA (Karmanov et al., 2024). BCA jointly updates class embeddings and class priors (Zhou et al., 2025), whereas TT-RAA estimates the streaming test distribution in a Gaussian database and retrieves relevant multimodal information for prediction (Fan et al., 2025). Recent methods also adapt model parameters or batch-level objectives. CLIPTTA uses a batch-aware soft contrastive objective aligned with contrastive pretraining (Lafon et al., 2025); BATCLIP adapts the visual and textual branches with entropy, image–text alignment, and inter-class separation objectives (Maharana et al., 2025); and BITTA learns from low-entropy samples while using high-entropy samples to unlearn atypical features (Sun et al., 2026). SAT assigns pseudo-labels by aligning visual features with text-based semantic anchors (Mishra et al., 2026), whereas StatA anchors adaptation to the initial text encoder under variable-class and non-i.i.d. test batches (Zanella et al., 2025). A comparative study further examines prompt engineering and online adaptation strategies for vision–language models (Döbler et al., 2025), while the TTA-VLM benchmark identifies substantial sensitivity to evaluation protocols and trade-offs beyond classification accuracy (Sheng et al., 2025). *SimplexTTA* is complementary to these choices of trainable parameters and base objectives: it changes the feature domain over which an existing differentiable adaptation loss is evaluated.

**Interpolation in representation space.** Mixup trains on convex combinations of examples and their labels to encourage smoother behavior between observations (Zhang et al., 2018). Manifold Mixup extends this principle to hidden representations (Verma et al., 2019), while AugMix stochastically composes and mixes image transformations to improve training-time corruption robustness (Hendrycks et al., 2020). These methods mix examples, representations, or transformed images during training. In contrast, *SimplexTTA* constructs a local convex region from augmentations of the same unlabeled test input, defines a Dirichlet distribution over that region, and approximates the expected test-time loss with a barycenter term and a curvature-weighted correction.

## 3 METHODOLOGY

In this section, we introduce *SimplexTTA*, a general test-time adaptation objective that optimizes a model over the continuous region spanned by augmented test views.

### 3.1 PROBLEM SETUP

Consider an unlabeled test image $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^n$ is the input space. Given a set of $M$ augmentations $\mathcal{A} = \{a_1, \ldots, a_M\}$, with $a_i : \mathcal{X} \to \mathcal{X}$, we obtain the augmented views $\tilde{\boldsymbol{x}}_i = a_i(\boldsymbol{x})$. Let $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^d$ be a neural network with parameters $\boldsymbol{\theta} \in \Theta$, mapping each input to a $d$-dimensional representation. The representations of the augmented views are

$$\boldsymbol{z}_i = f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_i), \qquad i = 1, \ldots, M. \tag{1}$$

Adaptation minimizes a differentiable TTA loss $\ell : \mathbb{R}^d \times \Theta \to \mathbb{R}$. Denote by

$$\Delta^{M-1} = \left\{ \boldsymbol{\lambda} \in \mathbb{R}^M : \lambda_i \geq 0,\ \textstyle\sum_{i=1}^{M} \lambda_i = 1 \right\} \tag{2}$$

the probability simplex, and by $\mathrm{Dir}(\boldsymbol{\lambda};\boldsymbol{\alpha})$ the Dirichlet density on $\Delta^{M-1}$ with concentration parameters $\boldsymbol{\alpha} \in \mathbb{R}^M_{>0}$,

$$\mathrm{Dir}(\boldsymbol{\lambda};\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^M \Gamma(\alpha_i)} \prod_{i=1}^M \lambda_i^{\alpha_i - 1}, \qquad \alpha_0 = \textstyle\sum_{i=1}^M \alpha_i. \tag{3}$$

For weights $\boldsymbol{\lambda} \in \Delta^{M-1}$, we write $\boldsymbol{z}_{\boldsymbol{\lambda}} = \sum_{i=1}^M \lambda_i \boldsymbol{z}_i$ for the corresponding convex combination of the augmented representations, and $\mathcal{S} \coloneqq \{\boldsymbol{z}_{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \in \Delta^{M-1}\}$ for the set of all such combinations.

Existing TTA methods evaluate $\ell$ on the original test image, on an averaged representation or prediction, or on a finite set of augmented views (Wang et al., 2021; Zhang et al., 2022; Shu et al., 2022). We instead treat the augmented representations as vertices of a continuous region and integrate the loss over their convex combinations,

$$\mathcal{L}_\Delta(\boldsymbol{\theta}) \coloneqq \int_{\Delta^{M-1}} \ell(\boldsymbol{z}_{\boldsymbol{\lambda}}, \boldsymbol{\theta}) \, \mathrm{Dir}(\boldsymbol{\lambda};\boldsymbol{\alpha}) \, d\boldsymbol{\lambda}, \tag{4}$$

where $d\boldsymbol{\lambda}$ is the Lebesgue measure on $\Delta^{M-1}$.

Rather than constraining the model only at the augmented views, $\mathcal{L}_\Delta$ enforces the adaptation objective over the entire region they span, treating every convex combination of views as a plausible representation of the test image. The Dirichlet concentration $\boldsymbol{\alpha}$ determines how the loss is distributed over the simplex, making $\mathcal{L}_\Delta$ a flexible family of objectives. For the symmetric choice $\boldsymbol{\alpha} = \alpha\mathbf{1}$ and a continuous $\ell$, letting $\alpha \to 0$ concentrates the Dirichlet on the vertices of the simplex, and $\mathcal{L}_\Delta$ reduces to the average of the per-view losses $\frac{1}{M}\sum_{i=1}^M \ell(\boldsymbol{z}_i, \boldsymbol{\theta})$, a standard discrete-view objective. In this work, we instantiate the simplex lift for TENT (Wang et al., 2021), CLIPTTA (Lafon et al., 2025), BATCLIP (Maharana et al., 2025), and BITTA (Sun et al., 2026). Furthermore, an asymmetric $\boldsymbol{\alpha}$ shifts mass toward selected views, since $\mathbb{E}[\boldsymbol{\lambda}] = \boldsymbol{\alpha}/\alpha_0$. Existing view-based objectives are thus endpoints of $\mathcal{L}_\Delta$, and the Dirichlet provides a principled way to move beyond them.

### 3.2 Second-order approximation

For a general nonlinear loss, the integral in eq. (4) has no closed form. It can be estimated by Monte Carlo sampling of $\boldsymbol{\lambda}$, but this gives a stochastic objective. We instead derive a deterministic closed-form surrogate by expanding $\ell$ to second order about the mean of the Dirichlet mixture.

For $\boldsymbol{\lambda} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, the mean of $\boldsymbol{z}_{\boldsymbol{\lambda}}$ is the weighted barycenter

$$\bar{\boldsymbol{z}} \coloneqq \mathbb{E}[\boldsymbol{z}_{\boldsymbol{\lambda}}] = \sum_{i=1}^M \beta_i \boldsymbol{z}_i, \qquad \beta_i \coloneqq \frac{\alpha_i}{\alpha_0}, \tag{5}$$

where $\boldsymbol{\beta} = \boldsymbol{\alpha}/\alpha_0 = \mathbb{E}[\boldsymbol{\lambda}]$ are the Dirichlet mean weights. We denote by $\boldsymbol{\delta}_{\boldsymbol{\lambda}} \coloneqq \boldsymbol{z}_{\boldsymbol{\lambda}} - \bar{\boldsymbol{z}}$ the deviation of a mixture from the barycenter, and by $\boldsymbol{\delta}_i \coloneqq \boldsymbol{z}_i - \bar{\boldsymbol{z}}$ the deviation of the $i$-th view, which is the special case in which $\boldsymbol{\lambda}$ is the $i$-th vertex of the simplex. Let

$$\boldsymbol{g} \coloneqq \nabla_{\boldsymbol{z}} \ell(\boldsymbol{z}, \boldsymbol{\theta})\big|_{\boldsymbol{z}=\bar{\boldsymbol{z}}}, \qquad \boldsymbol{H} \coloneqq \nabla^2_{\boldsymbol{z}} \ell(\boldsymbol{z}, \boldsymbol{\theta})\big|_{\boldsymbol{z}=\bar{\boldsymbol{z}}} \tag{6}$$

be the gradient and Hessian of the loss with respect to the representation at the barycenter. Expanding $\ell$ to second order about $\bar{\boldsymbol{z}}$ gives

$$\ell(\boldsymbol{z}_{\boldsymbol{\lambda}}, \boldsymbol{\theta}) \approx \ell(\bar{\boldsymbol{z}}, \boldsymbol{\theta}) + \boldsymbol{g}^\top \boldsymbol{\delta}_{\boldsymbol{\lambda}} + \tfrac{1}{2} \boldsymbol{\delta}_{\boldsymbol{\lambda}}^\top \boldsymbol{H} \, \boldsymbol{\delta}_{\boldsymbol{\lambda}}. \tag{7}$$

We note that the second order is the lowest order that retains the spread of the views, since the first-order term integrates to zero about the Dirichlet mean $\bar{\boldsymbol{z}}$.

**Lemma 1.** *Let $\ell(\cdot, \boldsymbol{\theta})$ be twice differentiable at $\bar{\boldsymbol{z}}$. Integrating the right-hand side of eq.* (7) *against* $\mathrm{Dir}(\boldsymbol{\lambda};\boldsymbol{\alpha})$ *gives*

$$\mathcal{L}_{\texttt{sTTA}}(\boldsymbol{\theta}) \coloneqq \ell(\bar{\boldsymbol{z}}, \boldsymbol{\theta}) + \frac{1}{2(\alpha_0 + 1)} \sum_{i=1}^M \beta_i \, \boldsymbol{\delta}_i^\top \boldsymbol{H} \, \boldsymbol{\delta}_i. \tag{8}$$

The proof is given in **??**.

**Algorithm 1** One test-time update with *SimplexTTA*

**Require:** Test batch $\mathcal{B}$; model $f_{\boldsymbol{\theta}}$; TTA loss $\ell$; adaptation parameters $\boldsymbol{\theta}_{\text{adapt}}$; step size $\eta$.
1: **for** each $\boldsymbol{x} \in \mathcal{B}$ **do**
2: Encode its $M$ views as $\boldsymbol{z}_i = f_{\boldsymbol{\theta}}(a_i(\boldsymbol{x}))$.
3: Compute $\beta_i = \alpha_i/\alpha_0$, $\bar{\boldsymbol{z}} = \sum_i \beta_i \boldsymbol{z}_i$, and $\boldsymbol{\delta}_i = \boldsymbol{z}_i - \bar{\boldsymbol{z}}$.
4: Evaluate $\ell(\bar{\boldsymbol{z}}, \boldsymbol{\theta})$ and the HVPs $\boldsymbol{H}\boldsymbol{\delta}_i$ by automatic differentiation.
5: Form $\mathcal{L}_{\texttt{sTTA}}$ using eq. (8).
6: **end for**
7: Average the objectives and update $\boldsymbol{\theta}_{\text{adapt}}$ by gradient descent.
**Ensure:** Updated adaptation parameters.

The first term is the loss at the barycenter, and the second is a curvature correction that measures the spread of the views along the curvature directions of $\ell$, with strength set by $\alpha_0$. For the symmetric choice $\boldsymbol{\alpha} = \alpha\mathbf{1}$, we have $\beta_i = 1/M$ and $\alpha_0 = M\alpha$, so

$$\mathcal{L}_{\texttt{sTTA}}(\boldsymbol{\theta}) = \ell(\bar{\boldsymbol{z}}, \boldsymbol{\theta}) + \frac{1}{2M(M\alpha+1)} \sum_{i=1}^{M} \boldsymbol{\delta}_i^\top \boldsymbol{H}\, \boldsymbol{\delta}_i. \tag{9}$$

The next theorem bounds the error incurred by replacing the integral $\mathcal{L}_\Delta$ with the closed-form surrogate $\mathcal{L}_{\texttt{sTTA}}$. The formal statement and proof are given in **??**.

**Theorem 1** (Surrogate approximation error (informal))**.** *Suppose the Hessian of $\ell(\cdot, \boldsymbol{\theta})$ is $\rho$-Lipschitz over $\mathcal{S}$, and let $D = \max_i \|\boldsymbol{\delta}_i\|$. Then*

$$\left|\mathcal{L}_\Delta(\boldsymbol{\theta}) - \mathcal{L}_{\texttt{sTTA}}(\boldsymbol{\theta})\right| \le \frac{\rho\, D^3}{6(\alpha_0+1)}, \tag{10}$$

*and, for fixed $\boldsymbol{\beta}$, the error is $O(\alpha_0^{-3/2})$.*

The error is thus governed by how much the curvature of $\ell$ varies across the views. If the Hessian is constant over $\mathcal{S}$ ($\rho = 0$), then $\ell$ is quadratic there, its second-order expansion is exact, and $\mathcal{L}_{\texttt{sTTA}} = \mathcal{L}_\Delta$. More generally, the error relative to the natural scale $\|\boldsymbol{H}\|D^2/(\alpha_0+1)$ of the curvature correction is at most of order $\rho D/\|\boldsymbol{H}\|$, independent of $\alpha_0$. Asymptotically, the error decays as $\alpha_0^{-3/2}$, whereas a nonvanishing correction decays as $\alpha_0^{-1}$, so $\mathcal{L}_{\texttt{sTTA}}$ captures the entire leading-order effect of integrating over the simplex.

**Quadratic special case.** For a quadratic loss, the Taylor expansion is exact. Under the symmetric choice $\boldsymbol{\alpha} = \alpha\mathbf{1}$, let $\ell_{\text{vert}} = M^{-1}\sum_{i=1}^{M} \ell(\boldsymbol{z}_i, \boldsymbol{\theta})$ and $\rho_\alpha = (M\alpha+1)^{-1}$. Then

$$\mathcal{L}_\Delta = \mathcal{L}_{\texttt{sTTA}} = (1-\rho_\alpha)\ell(\bar{\boldsymbol{z}}, \boldsymbol{\theta}) + \rho_\alpha \ell_{\text{vert}}. \tag{11}$$

Thus, the simplex objective interpolates between barycenter and vertex adaptation, retaining an intermediate level of augmentation covariance. This identity motivates the controlled study in section 4.6 and **??**.

### 3.3 Practical computation and adaptation

The curvature term is evaluated as a Hessian–vector product without forming the Hessian explicitly; our implementations may detach this product during backpropagation, as detailed with the complete gradients and method-specific conventions in **????**.

### 3.4 Application to CLIP test-time adaptation

We instantiate *SimplexTTA* in the visual embedding space of CLIP. Let $\boldsymbol{t}_c$ be the normalized text embedding of class $c$ and let $\tau$ be the logit scale. For a visual representation $\boldsymbol{z}$, the class logit is

$$s_c(\boldsymbol{z}) = \tau\, \frac{\boldsymbol{z}^\top \boldsymbol{t}_c}{\|\boldsymbol{z}\|_2 \|\boldsymbol{t}_c\|_2}. \tag{12}$$

Each experiment uses eight vertices: the original image and seven augmented views. We L2-normalize every view embedding before forming the barycenter and curvature directions, and normalize the barycenter again before computing logits. We use the symmetric concentration $\alpha = 1$ (denoted $\beta = 1$ in the experimental tables), include all eight views without filtering, and reuse the encoded views for adaptation and prediction. The primary *SimplexTTA* prediction is the normalized barycenter readout; original-view prediction is reported only as a diagnostic. We replace the original adaptation loss of TENT, CLIPTTA, BATCLIP, or BITTA with the corresponding simplex surrogate while retaining that method's trainable parameters and optimization protocol. Full objective definitions and implementation details are provided in **??**.

# 4 EXPERIMENTS

We evaluate *SimplexTTA* by applying it to four gradient-based test-time adaptation methods for CLIP, on clean data, under distribution shift and corruption, and in controlled ablations and a mechanism study.

## 4.1 EXPERIMENTAL SETUP

**Evaluation protocol.** All experiments use CLIP with a ViT-B/16 image encoder (Radford et al., 2021). We follow the online test-time adaptation setting: the model adapts continually over the unlabeled test stream of each dataset or corruption type, and is reset to the pretrained weights before each new dataset or corruption.

Each base method uses the objective and optimizer settings from its original paper. **TENT** (Wang et al., 2021) minimizes prediction entropy by adapting the normalization affine parameters. **CLIPTTA** (Lafon et al., 2025) optimizes a bidirectional soft-contrastive objective with a marginal-entropy regularizer; its OfficeHome/PACS update-count sweep is given in **??**. **BATCLIP** (Maharana et al., 2025) and **BITTA** (Sun et al., 2026) use their original objectives; in the main tables, both keep the text encoder frozen (the *frozen-text* variants), and variants that also update the text encoder are compared in **??**. Optimizers, learning rates, batch sizes, update counts, and memory settings for every method and benchmark are listed in **??**, and full objectives in **??**.

**Simplex lift.** For each input, *SimplexTTA* constructs $M{=}8$ vertices: the original image and seven augmented views from the shared view recipe (**??**). Simplex weights are drawn from a symmetric Dirichlet distribution with $\beta{=}1$. No view filtering of any kind is applied: all eight views are used as vertices in every result except the independent candidate-filtering ablation in **??**. All views are encoded in a single forward pass and cached for reuse between adaptation and prediction. Each feature is L2-normalized before averaging, and the barycenter is renormalized before computing logits. We report predictions from both the adapted original view (*Orig.*) and the barycenter (*Bary.*); unless stated otherwise, *SimplexTTA* accuracy refers to the barycenter prediction. Each lifted method keeps the optimizer, number of updates, and batch protocol of its base method. The one exception to the shared view recipe is TENT+*SimplexTTA* on EuroSAT, which uses label-preserving right-angle rotations and flips suited to overhead imagery (**??**). Method-specific lifted losses and gradient conventions are given in **??**.

**Datasets and metrics.** The clean evaluation uses the standard CLIP transfer suite (Zhou et al., 2022): ImageNet, Caltech-101, Oxford Pets, Stanford Cars, Flowers-102, Food-101, FGVC Aircraft, SUN397, DTD, EuroSAT, and UCF-101, together with CIFAR-10 and CIFAR-100 (13 datasets in total). We additionally evaluate target-domain classification on OfficeHome (Venkateswara et al., 2017) and PACS (Li et al., 2017), and natural distribution shift on ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNet-Sketch (Wang et al., 2019), and ImageNet-V2 (Recht et al., 2019). The corruption evaluation uses all 15 CIFAR-10-C, CIFAR-100-C, and ImageNet-C corruptions at severity 5 (Hendrycks & Dietterich, 2019). We report top-1 accuracy (%). Each mean is taken over the datasets or corruptions listed with it, and every $\Delta$ compares a base method with its lifted counterpart on exactly the same conditions.

Table 1: **Clean-classification results.** Top-1 accuracy (%) for each baseline and its matched +*SimplexTTA* variant with barycenter prediction, together with the resulting change. Positive changes are green and negative changes are red. The better accuracy in each matched pair is bold; exact ties are both bold. TENT+*SimplexTTA* uses the EuroSAT-specific rotation-and-flip recipe described in **??**; all other simplex entries use the shared view recipe. CLIPTTA uses memory-off, one-step adaptation with batch size 64, except on ImageNet, where both sides use batch size 256. BATCLIP and BITTA use the frozen-text variants with batch size 64. Means are taken over all 13 datasets.

| Dataset | TENT | | | CLIPTTA | | | BATCLIP | | | BITTA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Base | +*SimplexTTA* | Δ | Base | +*SimplexTTA* | Δ | Base | +*SimplexTTA* | Δ | Base | +*SimplexTTA* | Δ |
| ImageNet | **67.13** | 66.91 | −0.22 | 67.84 | **68.10** | +0.26 | 60.91 | **61.96** | +1.04 | 60.94 | **62.03** | +1.10 |
| Aircraft | 23.49 | **24.24** | +0.75 | 25.56 | **27.15** | +1.59 | 25.17 | **25.95** | +0.78 | 24.63 | **25.02** | +0.39 |
| Caltech-101 | 93.96 | **94.56** | +0.61 | 92.98 | **94.04** | +1.05 | 92.90 | **93.87** | +0.97 | 92.94 | **93.83** | +0.89 |
| Stanford Cars | 63.93 | **64.23** | +0.30 | 66.14 | **66.67** | +0.53 | 60.00 | **61.19** | +1.18 | 60.03 | **61.48** | +1.46 |
| DTD | 44.15 | **44.86** | +0.71 | 44.44 | **44.80** | +0.35 | 44.39 | **45.15** | +0.77 | 44.56 | **44.92** | +0.35 |
| EuroSAT | 65.30 | **67.44** | +2.15 | 65.10 | **67.47** | +2.37 | 42.35 | **58.44** | +16.10 | 50.21 | **60.07** | +9.86 |
| Flowers-102 | 70.65 | **71.30** | +0.65 | 70.85 | **71.66** | +0.81 | **70.16** | 69.87 | −0.28 | **70.20** | 69.96 | −0.24 |
| Food-101 | 86.45 | **86.73** | +0.28 | 86.40 | **86.82** | +0.43 | 79.39 | **79.62** | +0.23 | 79.45 | **79.67** | +0.22 |
| Oxford Pets | **89.92** | 89.86 | −0.05 | 89.59 | **90.16** | +0.57 | **87.93** | 87.87 | −0.05 | **87.87** | **87.87** | 0.00 |
| SUN397 | 62.73 | **63.02** | +0.29 | 64.13 | **64.14** | +0.02 | 59.13 | **60.34** | +1.21 | 58.41 | **60.41** | +1.99 |
| UCF-101 | 68.04 | **69.31** | +1.27 | 68.25 | **69.07** | +0.82 | 66.88 | **67.54** | +0.66 | 66.96 | **67.59** | +0.63 |
| CIFAR-10 | 94.50 | **94.59** | +0.09 | 93.18 | **93.38** | +0.20 | **90.12** | 89.52 | −0.60 | **90.30** | 89.51 | −0.79 |
| CIFAR-100 | 72.30 | **73.45** | +1.15 | **71.97** | 71.75 | −0.22 | 63.57 | **67.05** | +3.48 | 64.36 | **67.41** | +3.05 |
| Mean | 69.43 | **70.04** | +0.61 | 69.73 | **70.40** | +0.68 | 64.84 | **66.80** | +1.96 | 65.45 | **66.91** | +1.46 |

Table 2: **Natural ImageNet distribution shifts.** Matched results on ImageNet-A, ImageNet-R, ImageNet-Sketch, and ImageNet-V2. TENT uses one update with batch size 64; CLIPTTA uses memory-off, one-step, batch-256 adaptation; BATCLIP and BITTA use the frozen-text variants with batch size 64. *SimplexTTA* uses barycenter prediction. All results are seed 42 top-1 accuracy (%); the higher accuracy within each pair is bold, and positive changes are green.

| Dataset | TENT | | | CLIPTTA | | | BATCLIP | | | BITTA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Base | +*SimplexTTA* | Δ | Base | +*SimplexTTA* | Δ | Base | +*SimplexTTA* | Δ | Base | +*SimplexTTA* | Δ |
| ImageNet-A | 48.32 | **49.75** | +1.43 | 48.25 | **50.00** | +1.75 | 46.85 | **47.36** | +0.51 | 46.61 | **47.33** | +0.72 |
| ImageNet-R | 75.52 | **76.16** | +0.65 | 75.51 | **76.06** | +0.55 | 70.03 | **70.81** | +0.78 | 69.89 | **70.84** | +0.95 |
| ImageNet-Sketch | 46.88 | **47.39** | +0.51 | 48.33 | **48.49** | +0.16 | 39.31 | **41.64** | +2.33 | 37.55 | **41.90** | +4.35 |
| ImageNet-V2 | 60.69 | **61.41** | +0.72 | 61.00 | **61.45** | +0.45 | 58.60 | **59.91** | +1.31 | 58.65 | **59.82** | +1.17 |

### 4.2 Clean Classification

Table 1 reports clean accuracy on all 13 datasets. *SimplexTTA* improves the macro-average of every base method, with the largest gains for frozen-text BATCLIP (+1.96 points) and BITTA (+1.46 points). For TENT, *SimplexTTA* improves 11 of 13 datasets and raises the macro-average by 0.61 points; EuroSAT uses the overhead-imagery view recipe described above. CLIPTTA+*SimplexTTA* gains 0.68 points, including an improvement on ImageNet from 67.84 to 68.10; results with memory enabled and ten updates are given in **??**.

The gains are not uniform across datasets: BITTA+*SimplexTTA* decreases CIFAR-10 accuracy from 90.30 to 89.51 but increases CIFAR-100 accuracy from 64.36 to 67.41. BITTA results at batch size 200 on CIFAR, which have no paired *SimplexTTA* run, are listed in **??**. Text-updating BATCLIP and BITTA results are reported in **??**.

**??** shows rolling-accuracy trajectories and complete-run means for the five clean datasets with the largest BATCLIP gains. Because these datasets are selected post hoc, the figure illustrates adaptation dynamics rather than providing additional aggregate evidence.

### 4.3 Cross-domain and Natural Distribution Shifts

*SimplexTTA* consistently improves accuracy under natural distribution shift and on PACS, but slightly reduces accuracy on OfficeHome. Full OfficeHome and PACS target-domain results are reported in **??**, with batch size 64 for all methods. TENT uses one update, and CLIPTTA uses ten updates with memory disabled, the setting selected from the sweep in **??**. Averaged over the four target domains, *SimplexTTA* reduces OfficeHome accuracy by 0.38, 0.72, 0.80, and 0.73 points for

Table 3: **Corruption robustness at severity 5.** Seed-42 top-1 accuracy (%), averaged over all 15 corruption types; the higher accuracy within each pair is bold. CLIPTTA uses memory-on, ten-step, batch-128 adaptation on CIFAR-10-C/100-C and memory-off, one-step, batch-256 adaptation on ImageNet-C. BATCLIP and BITTA use the frozen-text variants with batch size 64.

| | CIFAR-10-C | | | CIFAR-100-C | | | ImageNet-C | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Base | +*SimplexTTA* | Δ | Base | +*SimplexTTA* | Δ | Base | +*SimplexTTA* | Δ |
| TENT | **63.21** | 60.19 | −3.02 | **37.02** | 24.75 | −12.27 | **26.20** | 24.24 | −1.96 |
| CLIPTTA | 80.63 | **81.71** | **+1.09** | 52.54 | **53.19** | **+0.65** | **33.36** | 33.04 | −0.32 |
| BATCLIP | 71.09 | **72.10** | **+1.01** | 37.38 | **40.08** | **+2.70** | 27.40 | **28.93** | **+1.53** |
| BITTA | 72.25 | **73.52** | **+1.28** | 39.15 | **41.67** | **+2.51** | 30.33 | **31.87** | **+1.54** |

TENT, CLIPTTA, BATCLIP, and BITTA, respectively, and increases PACS accuracy by 0.32, 0.51, 0.30, and 0.32 points; an independent ablation of candidate filtering on both benchmarks, which is not used in any other result, is given in **??**. On the natural ImageNet shifts (table 2), with batch-64 TENT, one-step, batch-256 CLIPTTA, and frozen-text BATCLIP and BITTA, *SimplexTTA* improves all 16 method–dataset pairs, by 0.83, 0.73, 1.23, and 1.80 points on average for TENT, CLIPTTA, BATCLIP, and BITTA, respectively.

### 4.4 Corruption robustness

Table 3 summarizes corruption robustness; per-corruption results for CIFAR-10-C, CIFAR-100-C, and ImageNet-C are given in **??**, using the same protocols and barycenter prediction.

*SimplexTTA* improves every corruption benchmark for frozen-text BATCLIP and BITTA, but reduces robustness for TENT. BATCLIP+*SimplexTTA* gains 1.01, 2.70, and 1.53 points on CIFAR-10-C, CIFAR-100-C, and ImageNet-C, respectively, and BITTA+*SimplexTTA* gains 1.28, 2.51, and 1.54 points. CLIPTTA+*SimplexTTA* improves CIFAR-10-C by 1.09 points and CIFAR-100-C by 0.65 points, but reduces ImageNet-C accuracy by 0.32 points under the one-step, memory-off protocol; the two CLIPTTA corruption protocols are compared in **??**.

In contrast, TENT+*SimplexTTA* decreases accuracy by 3.02, 12.27, and 1.96 points on the three benchmarks. Over all 45 conditions, the adapted original view is already 4.47 points below the baseline, and the barycenter prediction lowers accuracy by a further 1.28 points, so most of the loss arises during adaptation rather than at prediction. This pattern is consistent with entropy collapse. In a controlled study on CIFAR-100-C (**??**), barycentric TENT adaptation concentrates its predictions on a few classes, and adding a batch-diversity term prevents this. The failure therefore stems from TENT's entropy-only objective, which *SimplexTTA* inherits and can accelerate, rather than from the simplex construction alone. Condition-level readout decompositions and feature-space drift diagnostics are given in **????**. Text-updating results are reported separately in **??**.

**??** also shows, for each benchmark, the corruption with the largest BATCLIP gain, illustrating where the final per-corruption differences accumulate over the test stream.

### 4.5 Ablation studies

We conduct three matched frozen-text BATCLIP ablations on CIFAR-10 and four CIFAR-10-C corruptions. All variants use batch size 64, eight cached views, barycenter prediction, and, unless varied, one update and $\beta = 1$.

**Update objective versus prediction.** Panel (a) separates prediction-time averaging from the update objective. The full objective improves over Base+Bary readout by 0.72/2.54 clean/corruption points and over discrete vertices by 0.24/2.22, but matches barycenter-only adaptation within 0.08/0.01. Moreover, sweeping the implied quadratic interpolation weight from 0.93 to 0.001 changes accuracy by at most 0.15 points, and the matched TENT curvature gradient is only 0.14% of the barycentric gradient (**??**). Thus, the gain on this subset is barycentric; detailed results are in **??**.

Table 4: **Controlled BATCLIP ablations.** Top-1 accuracy (%) for frozen-text BATCLIP with batch size 64 on CIFAR-10 (C10) and four severity-5 CIFAR-10-C corruptions; C10-C is their mean. Unless varied, *SimplexTTA* uses eight views, one update, $\beta = 1$, and barycenter prediction. Bold denotes the best value in panels (a)–(b) and the higher accuracy within each matched pair in panel (c). Detailed condition-level results and protocol variants are in **??????**.

| (a) Objective component | | | (b) Dirichlet concentration | | | (c) Number of updates | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variant | C10 | C10-C | $\beta$ | C10 | C10-C | Steps | Base | +*SimplexTTA* C10 | Base | +*SimplexTTA* C10-C |
| Base+Bary readout | 88.80 | 66.88 | 0.01 | **89.60** | 69.42 | 0 | 89.29 | **89.82** | **58.05** | 57.32 |
| Discrete vertices | 89.28 | 67.20 | 0.1 | 89.56 | 69.48 | 1 | **90.12** | 89.52 | 68.33 | **69.42** |
| Barycenter loss only | **89.60** | **69.43** | 1 | 89.52 | 69.42 | 2 | **88.60** | 88.53 | 66.75 | **67.17** |
| Full simplex objective | 89.52 | 69.42 | 10 | 89.48 | 69.37 | 5 | 87.46 | **88.02** | **64.41** | 62.30 |
| | | | 100 | 89.57 | **69.52** | | | | | |

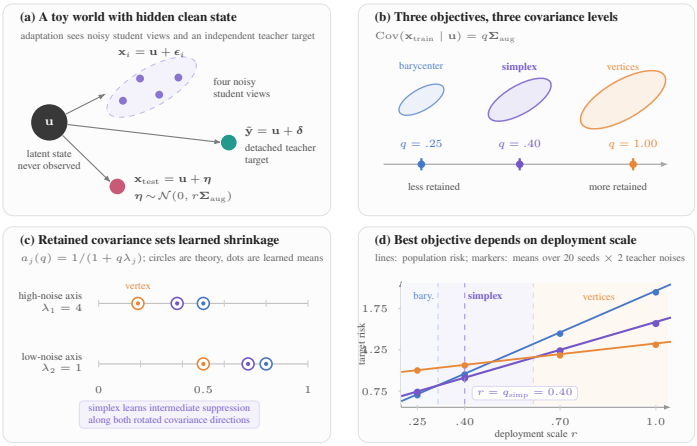


Figure 1: **Controlled covariance matching.** The three objectives retain distinct covariance scales and each wins at matched deployment variation.

**Sensitivity.** Accuracy varies by at most 0.15 points over the concentration sweep, while one update improves corruption accuracy by 1.09 points and five reduce it by 2.10. We therefore use $\beta = 1$ and one update; full sweeps are in the appendix.

### 4.6 Controlled quadratic mechanism study

With four views and $\beta = 1$, barycenter, simplex, and vertex adaptation retain covariance scales 0.25, 0.40, and 1. Across 20 paired seeds, each wins at matched deployment variation ($r = 0.25$, 0.40, and $0.70/1.00$), so simplex's advantage is conditional. In CLIP, full and barycenter-only adaptation are nearly identical (section 4.5), making barycenter versus vertices the relevant comparison; full details are in **??**.

## 5 Conclusion

We introduced *SimplexTTA*, which minimizes a test-time objective over the continuous feature-space simplex spanned by augmented views. A second-order approximation yields a practical barycenter-plus-curvature loss that lifts existing gradient-based TTA methods using Hessian–vector products. Across clean, natural-shift, and corruption benchmarks, the most consistent gains occur for BATCLIP and BITTA; TENT and OfficeHome show that improvements depend on the base objective and how augmentation variation relates to deployment shift. The quadratic study formalizes this covariance-matching mechanism, while the CLIP ablation identifies barycentric adaptation as the dominant empirical effect. These results motivate treating augmentations as a local adaptation region, rather than only as discrete observations.

### REPRODUCIBILITY STATEMENT

We support reproducibility by specifying the datasets, evaluation metrics, and reset rules in section 4.1, and the optimizer settings, learning rates, batch sizes, update counts, and memory settings for every method and benchmark in **??**. The complete simplex construction and objective are given in section 3, while method-specific objectives, trainable parameters, feature normalization, Hessian-vector product computation, and gradient conventions are documented in **??**. The appendix provides complete derivations of the second-order objective and its gradients, detailed corruption results, protocol and hyperparameter ablations, and the full controlled quadratic protocol. Main results report matched baseline and *SimplexTTA* comparisons under the same evaluation protocol.

### AI USE STATEMENT

In this work, we used generative AI tools for drafting and editing assistance. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Mario Döbler, Robert A. Marsden, Tobias Raichle, and Bin Yang. A lost opportunity for vision-language models: A comparative study of online test-time adaptation for vision-language models. In *Computer Vision – ECCV 2024 Workshops*, volume 15640 of *Lecture Notes in Computer Science*, pp. 117–133. Springer, 2025. doi: 10.1007/978-3-031-91672-4_8. URL https://link.springer.com/chapter/10.1007/978-3-031-91672-4_8.

Xinqi Fan, Xueli Chen, Luoxiao Yang, Chuin Hong Yap, Rizwan Qureshi, Qi Dou, Moi Hoon Yap, and Mubarak Shah. Test-time retrieval-augmented adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8810–8819, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Fan_Test-Time_Retrieval-Augmented_Adaptation_for_Vision-Language_Models_ICCV_2025_paper.html.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024. doi: 10.52202/079017-4099. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e92cb6f981a2cacb2a710ecaa0d7b141-Abstract-Conference.html.

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. URL https://openaccess.thecvf.com/content/ICCV2023/html/Feng_Diverse_Data_Augmentation_with_Diffusions_for_Effective_Test-time_Prompt_Tuning_ICCV_2023_paper.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

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. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_Generalization_ICCV_2021_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

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.

Marc Lafon, Gustavo Adolfo Vargas Hakim, Clément Rambour, Christian Desrosiers, and Nicolas Thome. CLIPTTA: Robust contrastive vision-language test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2412. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/67e2da8af2476c7441c881fc253042a6-Abstract-Conference.html.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9w3iw8wDuE.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5543–5551, 2017. doi: 10.1109/ICCV.2017.591. URL https://openaccess.thecvf.com/content_ICCV_2017/html/Li_Deeper_Broader_and_ICCV_2017_paper.html.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025. doi: 10.1007/s11263-024-02181-w. URL https://doi.org/10.1007/s11263-024-02181-w.

Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. SwapPrompt: Test-time prompt adaptation for vision-language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2847. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/cdd0640218a27e9e2c0e52e324e25db0-Abstract-Conference.html.

Sarthak Kumar Maharana, Baoming Zhang, Leonid Karlinsky, Rogerio Feris, and Yunhui Guo. BATCLIP: Bimodal online test-time adaptation for CLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1569–1579, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Maharana_BATCLIP_Bimodal_Online_Test-Time_Adaptation_for_CLIP_ICCV_2025_paper.html.

Shambhavi Mishra, Julio Silva-Rodriguez, Ismail Ben Ayed, Marco Pedersoli, and Jose Dolz. Semantic anchor transport: Robust test-time adaptation for vision-language models. In *ICLR Workshop on Test-Time Updates*, 2026. URL https://openreview.net/forum?id=dlC83Q3Gct.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

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 *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

David Osowiechi, Mehrdad Noori, Gustavo A. Vargas Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. WATT: Weight average test time adaptation of CLIP. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1522. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/55d16334943f8728073e17139e5baa3d-Abstract-Conference.html.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://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. URL https://proceedings.mlr.press/v97/recht19a.html.

Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? a critical look at test-time adaptation for vision-language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025. doi: 10.52202/085713-4167. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b57ddd8726c217a6fef9a48ce3e09ffd-Abstract-Datasets_and_Benchmarks_Track.html.

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 (NeurIPS)*, volume 35, pp. 14274–14289, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5bf2b802e24106064dc547ae9283bb0c-Abstract-Conference.html.

Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 825–835, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Sui_Just_Shift_It_Test-Time_Prototype_Shifting_for_Zero-Shot_Generalization_with_WACV_2025_paper.html.

Jingwei Sun, Jianing Zhu, Jiangchao Yao, Gang Niu, Masashi Sugiyama, and Bo Han. Bilateral information-aware test-time adaptation for vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vv8EcCoBfr.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Venkateswara_Deep_Hashing_Network_CVPR_2017_paper.html.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6438–6447. PMLR, 2019. URL https://proceedings.mlr.press/v97/verma19a.html.

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 (ICLR)*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jzzEHTBFOT.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zanella_Realistic_Test-Time_Adaptation_of_Vision-Language_Models_CVPR_2025_paper.html.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 38629–38642, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fc28053a08f59fccb48b11f2e31e81c7-Abstract-Conference.html.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1. URL https://doi.org/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.