ARE WE MEASURING THE RIGHT ANGLES? SPHERICAL TEST-TIME ADAPTATION OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) for vision–language models (VLMs) aims to improve robustness under distribution shifts by exploiting unlabeled test data during inference. Existing CLIP-TTA methods either optimize prompts through backpropagation or estimate Euclidean statistics from normalized CLIP features, despite CLIP performing classification through cosine similarity on the unit hypersphere. We propose **STRA**, a **S**pherical **T**est-**T**ime **R**obust **A**daptation framework for closed-form, backpropagation-free online CLIP adaptation. Our central view is that closed-form CLIP-TTA should recover a target-domain angular coordinate system under a pullback-vMF model, rather than fit an ambient Euclidean classifier. STRA instantiates this view through label-free geometry calibration from streaming second-order statistics and text-anchored spherical prototype updates driven by sparse soft pseudo-labels. This design avoids raw-feature caching and the selection bias induced by confidence-filtered geometry, decouples marginal geometry estimation from semantic pseudo-label updates, and couples them only through a shared bilateral angular map, while maintaining a single fixed configuration across datasets and backbones. Extensive experiments across natural distribution shifts, fine-grained recognition, and corruption robustness demonstrate consistent improvements over recent CLIP-TTA baselines with favorable computational efficiency. The code is available at https://anonymous.4open.science/r/STRA.

1 INTRODUCTION

CLIP (Radford et al., 2021) provides strong zero-shot recognition but degrades under domain and corruption shifts. While supervised prompt or adapter tuning (Zhou et al., 2022b;a; Zhang et al., 2022) uses labeled target data, online VLM adaptation (Shu et al., 2022; Karmanov et al., 2024; Zhang et al., 2024a; Guan et al., 2025; Zhang et al., 2026) must learn from an unlabeled stream. We study closed-set, source-free, label-free adaptation in which each sample is predicted immediately using only the current input and a compact past state, with frozen encoders and prompts. We call updates *closed-form* when they use streaming statistics and algebraic estimation rather than optimization or trainable adaptation modules.

The central challenge is that target shifts alter the angular structure on which CLIP's cosine classifier relies. A few nuisance directions can dominate normalized image features and distort angular neighborhoods, while useful class-specific evidence must be recovered only from noisy pseudo-labels. Existing CLIP-TTA methods address this issue through test-time prompt optimization (Shu et al., 2022; Feng et al., 2023; Yoon et al., 2024; Zhang et al., 2024a;b; Xiao et al., 2025), which requires backpropagation, feature caches (Karmanov et al., 2024; Fan et al., 2025; Zhang et al., 2024c), which are vulnerable to confident pseudo-label errors, or closed-form Euclidean summaries (Guan et al., 2025; Zhang et al., 2026; Zhou et al., 2025), which do not explicitly preserve the spherical decision geometry used by CLIP at inference time.

We instead treat target marginal geometry and class semantics as distinct statistical objects: the former uses every unlabeled feature, whereas the latter uses uncertain pseudo-label evidence. They interact only at prediction through a shared sphere-to-sphere map. A pullback-vMF model interprets the resulting bilateral cosine rule for a fixed map, rather than deriving the online estimators.

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.

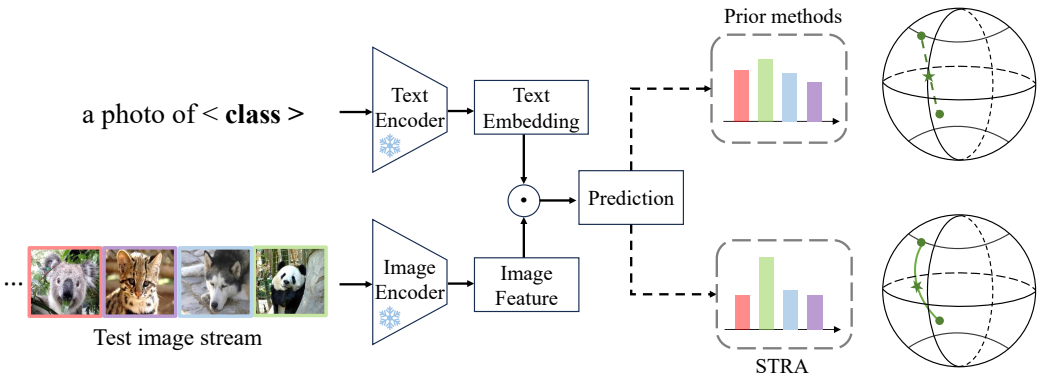


Figure 1: Comparing on the sphere versus comparing in the ambient space. A target shift spreads CLIP features anisotropically, so prior closed-form methods, which summarize features with Euclidean statistics, effectively compare two points by a straight line through the interior of the shifted cloud (up). STRA instead applies the coordinate map estimated from all unlabeled features, which restores the cloud to the sphere, and compares the same two points by the angle between them (down).

Figure 1 contrasts the two ways of comparing a shifted feature with a class direction: prior closed-form methods compare them by a straight line in the ambient space, whereas STRA restores the sphere first and compares them by the angle along it. We address this setting with **STRA** (**Spherical Test-Time Robust Adaptation**), using compact streaming statistics without caching raw test features.

Our main contributions are summarized as follows:

- We introduce a decoupled estimation formulation for closed-form CLIP-TTA. We show that confidence-selected geometry differs from the target marginal moment by an explicit covariance bias, motivating a pseudo-label-free geometry estimator while reserving text-anchored pseudo-label evidence for class semantics.
- We show that calibrating the label-free marginal second moment has a unique SPD solution, the inverse square root of the stream moment, whose power family is a log-Euclidean path from the identity to full calibration. For the resulting map, a pullback-vMF model yields the bilateral class-normalized angular readout.
- We instantiate the formulation as a practical online algorithm with compact streaming statistics, no backpropagation, and no raw-feature caching. Matched-stream, baseline-retrofit, imbalance, confusable-class, and larger-backbone studies evaluate both its accuracy–efficiency trade-off and the interaction between decoupled estimation and coupled spherical decision.

2 Related Work

Vision–language adaptation. CLIP (Radford et al., 2021) enables open-vocabulary recognition by aligning image and text embeddings and classifying normalized features with cosine similarity. Prompt-learning and adapter methods (Zhou et al., 2022b;a; Zhang et al., 2022) improve CLIP when labels or offline tuning data are available, but they are less suitable for online streams with only unlabeled test samples.

Online test-time adaptation for VLMs. Online CLIP-TTA adapts from unlabeled test data mainly by optimizing prompts (Shu et al., 2022; Feng et al., 2023; Yoon et al., 2024; Zhang et al., 2024a;b; Xiao et al., 2025; Sui et al., 2025; Meng et al., 2025), storing confident feature caches (Karmanov et al., 2024; Fan et al., 2025; Zhang et al., 2024c;a), or maintaining closed-form statistics (Guan et al., 2025; Zhang et al., 2026; Zhou et al., 2025). Prompt tuning requires backpropagation, while cache methods depend on pseudo-label confidence and memory policies. Among the closest closed-form approaches, SCA estimates a classifier from streaming statistics, ADAPT performs probabilistic Gaussian alignment, and BCA updates class centers in the original CLIP space.

Whitening and spherical modeling of CLIP features. CLIP features are normalized directions on $\mathbb{S}^{d-1}$, yet Euclidean summaries treat them as ambient vectors. Whitened CLIP (Betser et al., 2025) only applies offline whitening for image–caption scoring, and vMFCoOp (Shao et al., 2026) studies hyperspherical prompt learning with labels, neither of which performs source-free online estimation

for a frozen model. STRA instead decouples pseudo-label-free marginal geometry from text-anchored class directions and combines them through one bilateral normalized metric (Section 4.3). The von Mises–Fisher distribution (Banerjee et al., 2005) supplies the probabilistic interpretation.

3 METHOD

3.1 PRELIMINARIES

Problem setting and zero-shot CLIP. We assume a closed-set target stream whose C classes match the available text prototypes. Let $f_\theta(\cdot)$ be the frozen CLIP visual encoder and let $\mathbf{W} = [\boldsymbol{\mu}_1^{\text{txt}}, \dots, \boldsymbol{\mu}_C^{\text{txt}}] \in \mathbb{R}^{d\times C}$ collect the ℓ_2-normalized prototypes produced by the frozen text encoder and fixed prompt templates. At step t, the learner receives only the unlabeled image $\mathbf{x}_t$ and its previous compact state s_{t-1}; it has no access to future samples, target labels, or source-domain data. The image is mapped to $\mathbf{z}_t = f_\theta(\mathbf{x}_t) \in \mathbb{S}^{d-1}$ and scored by the frozen zero-shot logits $\boldsymbol{\ell}_t^{\text{zs}} = \kappa^* \mathbf{W}^\top \mathbf{z}_t$, with $\kappa^* = 100$. STRA follows an update-then-predict protocol: it algebraically updates s_{t-1} with the current unlabeled feature, outputs $\hat{y}_t$, and discards $\mathbf{z}_t$. It never updates either encoder or the prompts, and uses no backpropagation, optimizer state, trainable adaptation module, or raw-feature cache. Its pseudo-labels come exclusively from the frozen zero-shot branch, avoiding a feedback loop from adapted predictions to later semantic updates.

Theorem 1 (Pullback-vMF decision rule)**.** *Let $\mathbf{A} \succ 0$ and define the angular coordinate map $T_{\mathbf{A}}(\mathbf{z}) = \mathbf{A}\mathbf{z}/\|\mathbf{A}\mathbf{z}\|_2$ for $\mathbf{z} \in \mathbb{S}^{d-1}$. Suppose that, in the transformed coordinates, $T_{\mathbf{A}}(\mathbf{z}) \mid y = c$ follows a vMF distribution with direction $\boldsymbol{\nu}_c \in \mathbb{S}^{d-1}$ and shared concentration κ, and that class priors are equal. Then the Bayes classifier on the original sphere is*

$$h^\star(\mathbf{z}) = \arg\max_c \ \boldsymbol{\nu}_c^\top T_{\mathbf{A}}(\mathbf{z}). \tag{1}$$

Moreover, if $\boldsymbol{\nu}_c = T_{\mathbf{A}}(\boldsymbol{\mu}_c)$, this rule becomes the normalized metric score

$$\arg\max_c \ \frac{\mathbf{z}^\top \mathbf{A}^\top \mathbf{A}\boldsymbol{\mu}_c}{\sqrt{\mathbf{z}^\top \mathbf{A}^\top \mathbf{A}\mathbf{z}}\,\sqrt{\boldsymbol{\mu}_c^\top \mathbf{A}^\top \mathbf{A}\boldsymbol{\mu}_c}}, \tag{2}$$

which is the bilateral normalized angular score adopted by STRA when $\mathbf{A} = \mathbf{A}_t^{(\gamma)}$ and $\boldsymbol{\mu}_c = \boldsymbol{\mu}_{t,c}$.

3.2 STRA: CLOSED-FORM SPHERICAL ADAPTATION ON THE UNIT SPHERE

Overview. Figure 2 summarizes the procedure. A domain shift redistributes the features over the sphere and changes angles along the directions where the stream varies most. At decision time, STRA applies a map $\mathbf{A}$ determined by the anisotropy of the target second moment, and applies it exactly once, to the current feature and to every class direction. Theorem 1 states that the resulting cosine is in fact the Bayes rule under a shared-concentration vMF.

How the shift enters: anisotropy of the marginal second moment. Since $\mathbf{z}_t$ is a direction, the only structure a target domain can impose on it is the way directions are spread over the sphere, and the uncentered second moment is the natural summary of that spread: its leading eigenvectors are the directions along which the stream varies most, and whenever it is anisotropic, angles are stretched exactly along those directions, that is, exactly where CLIP measures them. Adapting the decision therefore means re-measuring angles in coordinates where the stream's spread is isotropic. The remaining question is which samples should contribute to that measurement. Keeping only confident samples is tempting, since their pseudo-labels are more reliable, but confidence is itself a function of direction: such a rule estimates a confidence-conditioned geometry rather than the marginal one. Let $\mathbf{M} = \mathbb{E}[\mathbf{z}\mathbf{z}^\top]$ and let $w(\mathbf{z}) \in [0, 1]$ denote any confidence-based inclusion or weighting rule with $\mathbb{E}[w] > 0$. The geometry seen by the weighted stream is $\mathbf{M}_w = \mathbb{E}[w(\mathbf{z})\mathbf{z}\mathbf{z}^\top]/\mathbb{E}[w(\mathbf{z})]$.

Proposition 1 (Confidence-selection bias)**.** *The weighted and marginal geometries satisfy*

$$\mathbf{M}_w - \mathbf{M} = \frac{\mathrm{Cov}\big(w(\mathbf{z}), \mathbf{z}\mathbf{z}^\top\big)}{\mathbb{E}[w(\mathbf{z})]}, \tag{3}$$

where $\mathrm{Cov}(w, \mathbf{z}\mathbf{z}^\top) := \mathbb{E}[(w - \mathbb{E}[w])\mathbf{z}\mathbf{z}^\top]$. Thus confidence filtering changes the population geometry unless confidence is uncorrelated with feature direction.

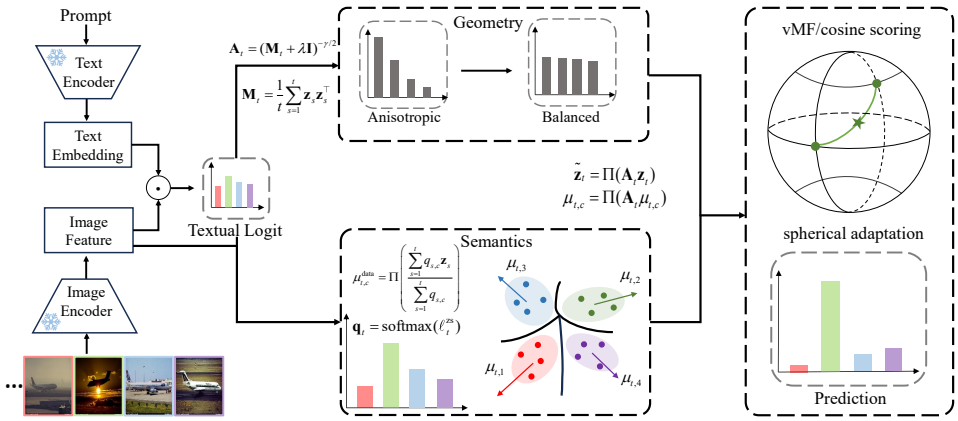


Figure 2: Overview of STRA. Frozen CLIP produces the image feature, the text prototypes, and the zero-shot logits. Two quantities are then estimated from the stream using different evidence: the angular coordinate map uses every unlabeled feature and no pseudo-label, whereas the class directions use sparse pseudo-labels of the frozen branch and stay text-anchored. At prediction one map is applied to both, both are projected back onto the sphere.

Setting $w \equiv 1$ makes the covariance term vanish identically, so the estimator targets the marginal geometry that the decision actually depends on. At test step t, STRA updates the uncentered second-moment statistic

$$\mathbf{G}_t = \mathbf{G}_{t-1} + \mathbf{z}_t \mathbf{z}_t^\top, \qquad \widehat{\mathbf{M}}_t = \mathbf{G}_t / n_t, \tag{4}$$

where n_t is the number of accumulated features.

Undoing the anisotropy in closed form. It remains to convert the measured anisotropy into a coordinate map. Let $\widehat{\mathbf{M}}_{t,\lambda} = \widehat{\mathbf{M}}_t + \lambda \mathbf{I} \succ 0$ with a fixed ridge λ. Asking that the stream become isotropic after the change of variables, $\mathbf{A}\widehat{\mathbf{M}}_{t,\lambda}\mathbf{A} = \mathbf{I}$, and measuring the residual in the Frobenius norm gives the calibration problem

$$\mathbf{A}_t^\star := \arg\min_{\mathbf{A} \succ 0} \left\| \mathbf{A}\widehat{\mathbf{M}}_{t,\lambda}\mathbf{A} - \mathbf{I} \right\|_F^2. \tag{5}$$

Proposition 2 (Marginal-calibration solution)**.** *The unique minimizer of Eq. (5) is* $\mathbf{A}_t^\star = \widehat{\mathbf{M}}_{t,\lambda}^{-1/2}$. *Moreover, for* $0 \le \gamma \le 1$*, the power family* $\mathbf{A}_t^{(\gamma)} = \widehat{\mathbf{M}}_{t,\lambda}^{-\gamma/2}$ *is the log-Euclidean geodesic from the identity to* $\mathbf{A}_t^\star$*;* $\gamma > 1$ *continues the same path by extrapolation.*

The calibration problem therefore selects the inverse square root of the stream moment, and the exponent interpolates continuously between the identity and full calibration. STRA uses the resulting power map

$$\mathbf{A}_t^{(\gamma)} = \big(\widehat{\mathbf{M}}_t + \lambda \mathbf{I}\big)^{-\gamma/2} = \mathbf{V}_t \operatorname{diag}\big(m_{t,i}^{-\gamma/2}\big)\mathbf{V}_t^\top, \tag{6}$$

where $m_{t,i}$ are the eigenvalues of $\widehat{\mathbf{M}}_{t,\lambda}$ and γ controls how far along the geodesic the correction travels. The default $\gamma = 1$ exactly minimizes the pre-projection marginal-moment objective in Eq. (5), while $\gamma = 0$ recovers the original CLIP geometry. This exactness concerns the calibration surrogate, not downstream classification risk: the subsequent sample-wise re-projection is nonlinear, and marginal isotropy need not maximize class separation. Values $\gamma > 1$ can therefore improve severe shifts by extrapolating beyond moment calibration. The selected map enters STRA only as a change of angular coordinates: both the feature and every prototype are mapped back to the sphere,

$$\tilde{\mathbf{z}}_t = \Pi\big(\mathbf{A}_t^{(\gamma)}\mathbf{z}_t\big), \qquad \tilde{\boldsymbol{\mu}}_{t,c} = \Pi\big(\mathbf{A}_t^{(\gamma)}\boldsymbol{\mu}_{t,c}\big), \qquad \Pi(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2. \tag{7}$$

The map is thus a pure change of coordinates: it is obtained without labels and is applied to both arguments of the decision, so it changes how angles are measured but never the sphere on which the decision is taken. Applying it to the feature alone would reduce the score to an unnormalized metric; re-projecting both sides is what keeps the decision inside CLIP's own angular domain. What is still missing is the second argument, the class directions.

Class directions from the frozen branch. Class directions are the only part of the model that requires labels, and the only supervision available comes from the frozen zero-shot branch, $\mathbf{q}_t = \mathrm{softmax}(\boldsymbol{\ell}_t^{\mathrm{zs}})$, where $\boldsymbol{\ell}_t^{\mathrm{zs}} = \kappa^* \mathbf{W}^\top \mathbf{z}_t$; STRA never feeds its own adapted predictions back into this update, so an erroneous decision cannot reinforce itself at a later step. A hard pseudo-label discards the ambiguity that the softmax encodes, while using the full dense tail lets many implausible classes contribute noise. STRA instead retains the smallest set of classes that carries most of the mass. Let $p_t^{\max} = \max_c q_{t,c}$ and define $b_t = p_t^{\max} + (1 - p_t^{\max})\tau_0$. After sorting classes by $q_{t,c}$ in descending order, K_t is the smallest top set whose cumulative probability reaches b_t; only this set is kept and renormalized:

$$\tilde{p}_{t,c} = \frac{q_{t,c}\mathbf{1}[c \in K_t]}{\sum_j q_{t,j}\mathbf{1}[j \in K_t]}. \tag{8}$$

This preserves ambiguity among the most plausible classes while discarding low-probability noise. Given $\tilde{p}_{t,c}$, the class-wise first moments are accumulated in the same native coordinates in which the pseudo-labels were produced. With $a_{t,c} = \mathbf{1}[\tilde{p}_{t,c} > 0]$, only active classes are decayed and updated:

$$\mathbf{S}_{t,c}^{(v)} = \rho^{a_{t,c}} \mathbf{S}_{t-1,c}^{(v)} + \tilde{p}_{t,c}\mathbf{z}_t, \qquad S_{t,c}^{(w)} = \rho^{a_{t,c}} S_{t-1,c}^{(w)} + \tilde{p}_{t,c}. \tag{9}$$

The data-induced direction is

$$\boldsymbol{\mu}_c^{\mathrm{data}} = \Pi\left(\frac{\mathbf{S}_{t,c}^{(v)}}{\max(S_{t,c}^{(w)}, \varepsilon)}\right). \tag{10}$$

When the accumulated evidence $S_{t,c}^{(w)}$ exceeds the minimum threshold, this direction is blended with the frozen text prototype as a stabilizing prior,

$$\boldsymbol{\mu}_{t,c} = \Pi\big(\alpha \boldsymbol{\mu}_c^{\mathrm{data}} + (1 - \alpha)\boldsymbol{\mu}_c^{\mathrm{txt}}\big); \tag{11}$$

otherwise it uses the fallback $\boldsymbol{\mu}_{t,c} = \boldsymbol{\mu}_c^{\mathrm{txt}}$. Because this update lives entirely in the native coordinates and never enters Eq. (4), pseudo-label noise can displace a class direction but cannot contaminate the geometry estimate; the text term mainly improves stability when pseudo-label evidence is sparse or noisy.

Spherical readout. The readout is where the two estimates meet. Given $\mathbf{A}_t := \mathbf{A}_t^{(\gamma)}$, STRA first obtains $\tilde{\mathbf{z}}_t$ and $\tilde{\boldsymbol{\mu}}_{t,c}$ by Eq. (7), then computes adapted vMF/cosine logits

$$\ell_{t,c}^{\mathrm{vmf}} = \kappa^* \tilde{\boldsymbol{\mu}}_{t,c}^\top \tilde{\mathbf{z}}_t. \tag{12}$$

Equivalently, with $\mathbf{M}_{\gamma,t} := \mathbf{A}_t^\top \mathbf{A}_t = (\widehat{\mathbf{M}}_t + \lambda \mathbf{I})^{-\gamma}$, the underlying angular score is

$$s_\gamma(\mathbf{z}, \boldsymbol{\mu}_c) = \frac{\mathbf{z}^\top \mathbf{M}_{\gamma,t} \boldsymbol{\mu}_c}{\sqrt{\mathbf{z}^\top \mathbf{M}_{\gamma,t}\mathbf{z}}\sqrt{\boldsymbol{\mu}_c^\top \mathbf{M}_{\gamma,t}\boldsymbol{\mu}_c}}. \tag{13}$$

The denominator is class-dependent, so this score is not equivalent to a Mahalanobis/LDA rule or to temperature scaling of CLIP logits; it reduces to the original cosine score only when $\gamma = 0$ or $\mathbf{M}_{\gamma,t}$ is isotropic. Finally, STRA fuses the adapted and frozen branches with an entropy gate. It maintains branch-entropy EMAs $\bar{H}_t^b = \rho_H \bar{H}_{t-1}^b + (1 - \rho_H)H_t^b$ for $b \in \{\mathrm{zs}, \mathrm{vmf}, \mathrm{raw}\}$, sets the online calibration multiplier $\tau^{(t)} = (\bar{H}_t^{\mathrm{zs}} + \varepsilon)/(\bar{H}_t^{\mathrm{vmf,raw}} + \varepsilon)$, recomputes H_t^{vmf} from $\mathrm{softmax}(\tau^{(t)}\boldsymbol{\ell}_t^{\mathrm{vmf}})$, and uses

$$c_t = \sigma\big(\beta(H_t^{\mathrm{vmf}} - H_t^{\mathrm{zs}})\big), \qquad \boldsymbol{\ell}_t = c_t \boldsymbol{\ell}_t^{\mathrm{vmf}} + (1 - c_t)\boldsymbol{\ell}_t^{\mathrm{zs}}. \tag{14}$$

Thus the zero-shot branch remains available as a conservative fallback when the adapted branch is less reliable. All constants are fixed globally and listed in Appendix E.

Complexity. STRA is closed-form and stream-length independent in memory, storing only the second-moment matrix and the class-direction statistics in $\mathcal{O}(d^2 + Cd)$ space. Streaming-statistic updates require $\mathcal{O}(d^2)$ computation and class scoring costs $\mathcal{O}(Cd)$. Computing the full whitening operator requires an $\mathcal{O}(d^3)$ matrix decomposition, but refreshing it only once every R samples reduces this term to an amortized $\mathcal{O}(d^3/R)$ cost. In practice, R can be increased substantially: on ViT-L/14, changing it from 10 to 1000 reduces latency from 24.50 to 22.07 ms/sample while changing accuracy by only 0.02 points (Appendix J).

Algorithm 1 STRA: online test-time adaptation.

Input: unlabeled test stream $\mathcal{D}_u$; frozen CLIP encoder f_θ; text prototypes $\mathbf{W}$.
Initialize the geometry statistic $\mathbf{G}_0$, the class-direction moments, and $\boldsymbol{\mu}_{0,c} \leftarrow \boldsymbol{\mu}_c^{\text{txt}}$.
for $\mathbf{x}_t \in \mathcal{D}_u$ **do**
  Encode $\mathbf{x}_t$ to $\mathbf{z}_t$ and compute zero-shot logits $\boldsymbol{\ell}_t^{\text{zs}}$.
  Accumulate $\mathbf{z}_t\mathbf{z}_t^\top$ into $\mathbf{G}_t$ and refresh $\mathbf{A}_t^{(\gamma)}$ when scheduled.
  Apply TRUNCSOFTMAX to $\boldsymbol{\ell}_t^{\text{zs}}$ to obtain sparse soft pseudo-labels $\tilde{\mathbf{p}}_t$.
  Update the class-direction moments and prototypes $\{\boldsymbol{\mu}_{t,c}\}_{c=1}^C$ with the text prior.
  Whiten and re-project $\mathbf{z}_t$ and $\{\boldsymbol{\mu}_{t,c}\}_{c=1}^C$ by Eq. (7).
  Compute adapted logits $\boldsymbol{\ell}_t^{\text{vmf}}$ and fuse them with $\boldsymbol{\ell}_t^{\text{zs}}$ by Eq. (14).
  Output $\hat{y}_t = \arg\max_c \ell_{t,c}$.
end for
return predictions $\{\hat{y}_t\}_{t=1}^T$.

Table 1: Top-1 accuracy (%) on the natural distribution shift benchmark, averaged over five random stream orders. Best results are in bold and second-best results are underlined among the methods listed. CLIP-R: CLIP-ResNet-50, CLIP-V: CLIP-ViT-B/16.

Method	IN	IN-A	IN-V	IN-R	IN-S	**Shift Avg.**	**All Avg.**
CLIP-R	58.16	21.83	51.41	56.15	33.37	40.69	44.18
BCA	61.81	30.35	56.58	62.89	38.04	46.97	49.93
TDA	61.35	30.29	55.54	62.58	38.12	46.63	49.58
Dota	61.82	30.81	55.27	62.81	37.52	46.60	49.65
SCA	63.11	30.64	56.51	62.85	39.33	47.33	50.49
ADAPT	62.16	33.08	55.97	62.69	40.21	47.99	50.82
Ours	**64.39**	**34.68**	**57.07**	**63.38**	**40.68**	**48.95**	**52.04**
CLIP-V	66.73	47.87	60.86	73.98	46.09	57.20	59.11
BCA	70.22	61.14	64.90	80.72	50.87	64.41	65.57
TDA	69.51	60.11	64.67	80.24	50.54	63.89	65.01
Dota	70.68	61.19	64.41	81.17	51.33	64.53	65.76
SCA	71.75	60.33	65.38	80.85	52.50	64.77	66.16
ADAPT	70.91	63.32	64.64	80.66	53.13	65.44	66.53
Ours	**72.91**	**64.58**	**65.77**	**81.50**	**53.73**	**66.40**	**67.70**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate STRA on three tasks that cover natural distribution shift, fine-grained categorization, and corruption robustness. For natural distribution shift, we use ImageNet (Deng et al., 2009), ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-Sketch (Wang et al., 2019). For corruption robustness, we evaluate on ImageNet-C (Hendrycks & Dietterich, 2019) at severity level 5, which contains 15 corruption types spanning noise, blur, weather, and digital artifacts. We also evaluate on 10 fine-grained recognition datasets: FGVC-Aircraft (Maji et al., 2013), Caltech-101 (Fei-Fei et al., 2004), Stanford Cars (Krause et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Oxford Flowers (Nilsback & Zisserman, 2008), Food-101 (Bossard et al., 2014), Oxford Pets (Parkhi et al., 2012), SUN397 (Xiao et al., 2010), and UCF-101 (Soomro et al., 2012).

Baselines. We compare STRA with representative CLIP adaptation methods(TDA (Karmanov et al., 2024), BCA (Zhou et al., 2025), Dota (Han et al., 2024), SCA (Guan et al., 2025), ADAPT (Zhang et al., 2026)). Following prior online evaluations (Karmanov et al., 2024; Guan et al., 2025; Zhang et al., 2026), we focus on the batch-size-one online setting and exclude transductive methods. The complete per-dataset results for more baseline, for each backbone separately, are given in Appendix B.

Table 2: Top-1 accuracy (%) on the fine-grained categorization benchmark, averaged over five random stream orders. Best results are in bold and second-best results are underlined among the methods listed. CLIP-R: CLIP-ResNet-50, CLIP-V: CLIP-ViT-B/16.

Method	Aircraft	Caltech	Cars	DTD	EuroSAT	Flowers	Food101	Pets	SUN397	UCF101	Avg.
CLIP-R	15.66	85.88	55.70	40.37	23.69	61.75	73.97	83.57	58.80	58.84	55.82
BCA	19.89	89.70	58.13	48.58	42.12	66.30	77.19	85.58	63.38	63.51	61.44
TDA	17.61	89.70	57.78	43.74	42.11	68.74	77.75	86.18	62.53	64.18	61.03
Dota	18.06	88.84	58.72	45.80	47.15	68.53	**78.61**	87.33	63.89	65.08	62.20
SCA	**20.82**	**91.03**	59.73	53.31	43.33	71.09	76.51	86.15	64.50	66.40	63.29
ADAPT	20.10	88.72	59.94	52.60	**49.70**	68.33	76.06	87.33	65.24	65.77	63.38
Ours	20.56	89.88	**60.56**	**54.37**	44.48	**71.81**	78.28	**88.02**	**65.87**	**66.67**	**64.05**
CLIP-V	23.67	93.35	65.48	44.27	42.01	67.44	83.65	88.25	62.59	65.13	63.58
BCA	**28.59**	94.69	66.86	53.49	56.63	73.12	85.97	90.43	68.41	67.59	68.58
TDA	23.91	94.24	67.28	47.40	58.00	71.42	86.14	88.63	67.62	70.66	67.53
Dota	25.59	94.32	69.48	47.87	57.65	74.67	**87.02**	91.69	69.70	72.06	69.01
SCA	28.50	**94.85**	68.49	**57.09**	57.16	76.09	86.09	91.44	70.27	73.43	70.34
ADAPT	27.99	93.87	69.08	55.38	58.88	75.27	85.32	92.12	70.31	73.88	70.21
Ours	27.33	94.71	**70.14**	56.49	**60.39**	**78.83**	87.01	**92.79**	**71.33**	**75.26**	**71.43**

Table 3: Top-1 accuracy (%) on the corruption robustness benchmark, averaged over five random stream orders. Best results are in bold and second-best results are underlined among the methods listed. CLIP-R: CLIP-ResNet-50, CLIP-V: CLIP-ViT-B/16.

Method	Blur				Weather				Digital				Noise			Avg.
	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brig.	Cont.	Elas.	Pix.	JPEG	Gauss.	Shot	Impu.	
CLIP-R	9.54	3.40	7.46	12.62	12.29	15.72	22.08	41.69	6.24	4.67	11.01	14.24	2.43	3.07	2.52	11.27
TDA	9.84	4.40	7.38	13.74	13.74	17.16	23.76	44.16	7.00	5.79	11.24	15.26	2.54	3.26	2.72	12.13
SCA	**11.88**	4.96	9.28	14.91	15.45	19.10	25.78	46.42	8.42	6.91	13.63	17.26	3.13	3.85	3.41	13.63
ADAPT	10.54	4.44	8.57	14.34	13.85	17.84	24.56	45.67	7.76	5.85	11.96	15.86	2.91	3.77	2.92	12.72
Ours	11.58	**5.75**	**9.87**	**16.47**	**16.33**	**20.67**	**28.04**	**47.37**	**8.55**	**8.65**	**14.23**	**17.66**	**3.89**	**4.65**	**4.07**	**14.52**
CLIP-V	24.25	15.71	24.46	22.60	33.08	31.06	37.61	55.62	17.11	13.43	33.04	33.70	13.25	14.16	13.48	25.50
TDA	26.53	17.91	27.35	25.90	36.50	34.84	40.53	58.57	20.16	16.62	35.65	36.69	15.42	16.46	16.03	28.34
SCA	26.31	18.13	27.15	25.14	35.70	34.09	40.62	60.25	19.96	16.46	37.27	37.57	15.77	16.94	15.99	28.49
ADAPT	26.30	18.01	27.31	25.54	36.19	34.67	40.96	60.29	19.95	16.09	37.44	37.22	15.76	16.84	15.90	28.56
Ours	**28.10**	**19.75**	**28.81**	**27.22**	**38.39**	**36.92**	**42.97**	**61.18**	**21.13**	**19.40**	**38.97**	**37.99**	**17.88**	**18.90**	**18.92**	**30.43**

Implementation details. Consistent with prior closed-form CLIP-TTA work (Guan et al., 2025; Zhang et al., 2026), our experiments adopt CLIP ViT-B/16 and RN50 as visual encoders, with batch size 1 to satisfy online processing. For textual prompts, we use the prompt-ensemble strategy adopted in prior CLIP-TTA studies, averaging and re-normalizing normalized CLIP text embeddings from dataset-specific templates and descriptive prompts. Augmentation-based baselines keep their official settings. We use one fixed STRA configuration for all datasets, corruption types, stream orders, and backbones, with all constants reported in Appendix E; no validation set or dataset-specific hyperparameter selection is used. We report means over five random stream orders. All experiments are conducted on a single NVIDIA A100 GPU, using Top-1 accuracy as the evaluation metric.

4.2 Main Results

Task 1: Natural Distribution Shift. As shown in Table 1, STRA achieves the highest overall average accuracy of $\mathbf{67.70}\%$ with ViT-B/16 and $\mathbf{52.04}\%$ with RN50 on ImageNet and four natural-shift variants. It outperforms ADAPT by 1.17 and 1.22 points on the overall average and by 0.96 points on the shift average for the two backbones, respectively, while ranking first on all five datasets with ViT-B/16 and on ImageNet, ImageNet-A, and ImageNet-Sketch with RN50. This demonstrates consistent robustness under natural distribution shifts. The complete per-dataset comparison against every baseline is given in Tables 8 and 11.

Table 4: Ablation of STRA components with CLIP ViT-B/16.

Method	Task 1	Task 2	Task 3
Baseline	59.11	63.58	25.50
w/o prototype update	66.85	69.57	26.96
w/o geometry map	66.84	69.86	27.82
w/o spherical re-proj	66.88	69.77	27.05
w/o text prior	67.60	71.30	29.45
w/o entropy fusion	**67.71**	71.37	30.29
STRA	67.70	**71.43**	**30.43**

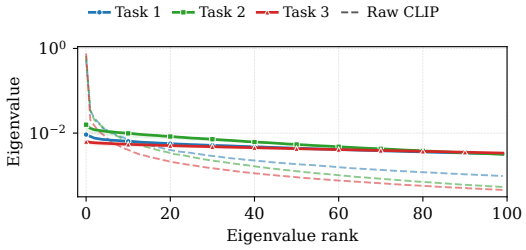


Figure 3: Geometry diagnostics of image features with CLIP ViT-B/16.

Table 5: Post-hoc spherical re-projection on closed-form baselines with ViT-B/16. †: only the final readout is replaced by bilateral spherical re-projection.

Method	Task 1	Task 2	Task 3
SCA	66.16	70.34	28.49
SCA†	66.29	66.50	27.71
ADAPT	66.53	70.21	28.56
ADAPT†	66.31	68.11	28.14
BCA	65.57	68.58	27.87
BCA†	65.74	68.31	27.86
STRA	**67.70**	**71.43**	**30.43**

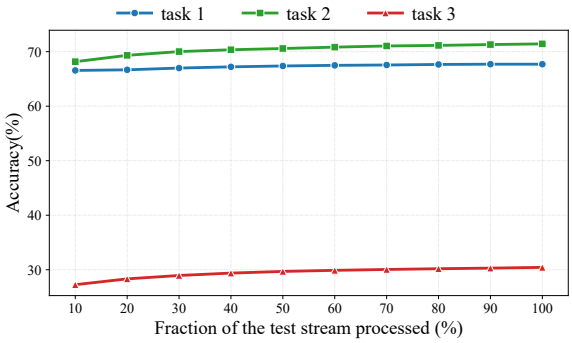


Figure 4: Online adaptation dynamics of STRA with CLIP ViT-B/16.

Task 2: Fine-Grained Categorization. Table 2 reports results on ten fine-grained recognition datasets under both backbones. STRA obtains the best average accuracy of $\mathbf{71.43}\%$ with ViT-B/16 and $\mathbf{64.05}\%$ with RN50, exceeding ADAPT by 1.22 and 0.67 points and SCA by 1.09 and 0.76 points, respectively. These results support generalization to fine-grained target domains without optimizing language-side parameters.

Task 3: Corruption Robustness. Table 3 evaluates the 15 ImageNet-C corruptions at severity 5 under both backbones. STRA achieves the best mean accuracy of $\mathbf{30.43}\%$ with ViT-B/16 and $\mathbf{14.52}\%$ with RN50, outperforming ADAPT by 1.87 and 1.80 points, respectively, and ranking first on all 15 corruption types with ViT-B/16 and on 14 of 15 with RN50 (Tables 10 and 13). These results support transfer across CLIP backbones without backbone-specific hyperparameter selection.

Online dynamics. Figure 4 plots task-level accuracy from the first 10% checkpoint to the end of each stream. The curves remain stable after the early checkpoints without late-stage degradation, indicating no noticeable error accumulation from the online geometry and prototype updates.

4.3 Ablation Studies and Discussion

Component ablation. The Baseline in Table 4 is the frozen zero-shot branch under the same prompts and protocol. Removing either the marginal-geometry map or the prototype update degrades STRA. Removing spherical re-projection retains the estimated transform but replaces the normalized score in Eq. (13) with an unnormalized metric score, yielding near-baseline natural-shift and below-baseline corruption accuracy. This supports the bilateral readout constraint in Theorem 1. The text prior mainly helps corruptions, whereas entropy fusion has only a small effect.

Transferability of the angular-coordinate readout. A natural question is whether STRA's gain comes solely from the spherical readout. To test this, we re-run SCA, ADAPT, and BCA with their update rules left intact and only their final readout replaced by bilateral spherical re-projection († in Table 5). None of the resulting methods comes close to STRA. The gain therefore comes not from the readout but from its coupling with the estimation.

Table 6: Accuracy–efficiency trade-off with ViT-B/16.

Task	Method	Acc.(mean±std)	Latency(ms) ↓	Memory(GB) ↓
Task 1	SCA	66.16±0.07	83.66	**0.692**
	ADAPT	66.53±0.01	151.71	0.931
	STRA	**67.70±0.04**	**78.77**	0.696
Task 2	SCA	70.34±0.09	17.49	**0.351**
	ADAPT	70.21±0.06	35.89	0.435
	STRA	**71.43±0.07**	**15.16**	0.355
Task 3	SCA	28.49±0.02	17.83	**0.356**
	ADAPT	28.56±0.01	79.53	0.771
	STRA	**30.43±0.02**	**13.94**	0.363

Table 7: Task-level accuracy (%) under different test-time stream orders with ViT-B/16.

Stream order	Task 1	Task 2	Task 3
Class-blocked	66.34	69.10	30.36
Easy-to-Hard	67.91	71.72	31.51
Hard-to-Easy	66.70	69.46	29.78
Random (Ours)	67.70	71.43	30.43

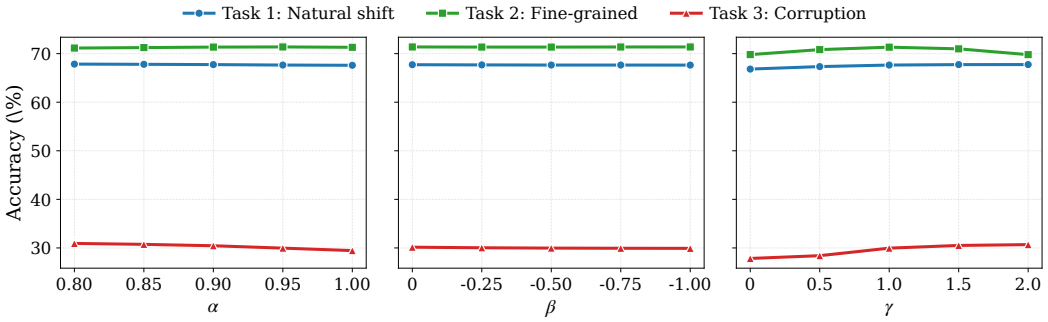


Figure 5: Task-level sensitivity of STRA to α, β, and γ on CLIP ViT-B/16.

Geometry diagnostics. Figure 3 shows that the geometry map together with spherical re-projection flattens the second-moment spectrum, confirming reduced marginal anisotropy. Spectrum flattening alone does not identify nuisance directions; its task benefit is established by the accuracy ablations.

Matched-stream reliability and efficiency. Table 6 reports the accuracy–efficiency comparison across all three benchmark groups under the same batch-size-one protocol and five matched stream orders. STRA achieves the highest accuracy, the lowest latency, and the highest throughput in every group, with peak memory close to SCA and well below ADAPT. The gains of 1.17, 0.67, and 1.87 points on Tasks 1–3 are substantially larger than the observed standard deviations.

Hyperparameter Analysis. Figure 5 shows broad stable ranges for α and β. The geometry parameter γ follows the path in Proposition 2: $\gamma = 1$ exactly calibrates the pre-projection moment. Larger γ performs better on Task 3, because severe corruptions inject strong nuisance variance that stronger whitening can suppress. We keep $\gamma = 1$ for all tasks rather than tuning it per benchmark, since a single fixed configuration is what makes STRA usable without target-side validation.

Robustness to test-time sample order. We compare random streams with class-blocked order and label-free easy-to-hard, hard-to-easy orders based on frozen CLIP entropy (Table 7), indicating that the stream order does have some influence on the results, while STRA remains stable throughout.

5 Limitations and Conclusion

Limitations. The full-matrix geometry map uses $\mathcal{O}(d^2)$ memory and periodic $\mathcal{O}(d^3)$ factorization, although infrequent refreshes substantially amortize this cost on ViT-L/14. but it remains super-linear in dimension and would need a structured or low-rank variant for much larger embedding widths.Proposition 3 gives sufficient rather than universal conditions for inverse-distortion recovery.

Conclusion. We introduced **STRA**, a closed-form framework that separates label-free marginal geometry from text-anchored class semantics and couples them through a bilateral spherical decision rule. Its theory and results across natural distribution shifts, fine-grained categorization, and corruption robustness benchmark support geometry-consistent, efficient online CLIP adaptation.

AI USE STATEMENT

In this work, we used generative AI tools for copy-editing and formatting assistance, namely polishing the wording of a few sentences and checking layout commands. We have not used generative AI tools for proposing research ideas, deriving the method or its proofs, writing experiment code, or producing any of the reported numbers; all ideas, derivations, code, and experiments were conceived and implemented by the authors. We have reviewed all AI-assisted work and verified that the suggested edits preserve the intended technical meaning. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

ETHICS STATEMENT

This work studies test-time adaptation of frozen, publicly released vision–language models and does not involve human subjects, personally identifiable information, or the release of new datasets; all benchmarks used are publicly available. Because STRA builds on a frozen CLIP backbone, it may inherit biases from the pretrained representation and from the prompts and class taxonomies used at inference time, and online adaptation under class imbalance or non-stationary streams may amplify such biases. We discuss potential societal impacts in more detail in Appendix K. We recommend task-specific validation, calibration checks, and fairness audits, together with human oversight or abstention mechanisms, before deployment in high-stakes settings.

REPRODUCIBILITY STATEMENT

All experiments use publicly available datasets and frozen, publicly released CLIP checkpoints, as detailed in Section 4.1; no new data collection is involved. STRA is closed-form and training-free, so results do not depend on optimizer seeds: the only source of variability is the order of the unlabeled test stream, which we quantify in Appendix H. Complete proofs of all propositions are given in Appendix A, the full hyperparameter configuration and its sensitivity analysis are given in Appendix E (Table 14), and additional backbone and online-dynamics results are reported in Appendices B and C. Anonymous source code is available at https://anonymous.4open.science/r/STRA.

REFERENCES

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened clip as a likelihood surrogate of images and captions. *arXiv preprint arXiv:2505.06934*, 2025.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Xinqi Fan, Xueli Chen, Luoxiao Yang, 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*, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

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.

Zenghao Guan, Zhou Yucan, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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. *arXiv preprint arXiv:2409.19375*, 2024.

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.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the 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, et al. 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.

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.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

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

Fan'an Meng, Chaoran Cui, Hongjun Dai, and Shuai Gong. Black-box test-time prompt tuning for vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6099–6107, 2025.

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

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Minye Shao, Sihan Guo, Xinrun Li, Xingyu Miao, Haoran Duan, and Yang Long. vmfcoop: Towards equilibrium on a unified hyperspherical manifold for prompting biomedical vlms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8851–8859, 2026.

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. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

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.

Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 825–835. IEEE, 2025.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

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

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees GM Snoek. Dynaprompt: Dynamic test-time prompt tuning. *arXiv preprint arXiv:2501.16404*, 2025.

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. *arXiv preprint arXiv:2403.14119*, 2024.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. *Advances in Neural Information Processing Systems*, 37:32111–32136, 2024a.

Jingyi Zhang, Jiaxing Huang, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Historical test-time prompt tuning for vision foundation models. *Advances in Neural Information Processing Systems*, 37: 12872–12896, 2024b.

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 *European conference on computer vision*, pp. 493–510. Springer, 2022.

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. *Advances in Neural Information Processing Systems*, 38:76280–76310, 2026.

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(9):2337–2348, 2022b.

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 Computer Vision and Pattern Recognition Conference*, pp. 29999–30009, 2025.

A Derivations for Spherical Whitening

This appendix develops the theory behind STRA in three steps: confidence selection can bias marginal geometry, the inverse square root of the unlabeled stream moment is the unique SPD solution of a label-free moment-calibration problem, and the resulting fixed map admits a pullback-vMF decision rule after bilateral re-projection. We prove Propositions 1 and 2, give an idealized recovery result for latent angular distortions, and then clarify the relationship to shared-covariance Euclidean classifiers.

Proof of Proposition 1. By adding and subtracting $\mathbb{E}[w]\mathbb{E}[\mathbf{z}\mathbf{z}^\top]$,

$$\mathbf{M}_w - \mathbf{M} = \frac{\mathbb{E}[w\mathbf{z}\mathbf{z}^\top] - \mathbb{E}[w]\mathbb{E}[\mathbf{z}\mathbf{z}^\top]}{\mathbb{E}[w]} \tag{15}$$

$$= \frac{\mathrm{Cov}(w, \mathbf{z}\mathbf{z}^\top)}{\mathbb{E}[w]}. \tag{16}$$

Hence confidence weighting is unbiased for the marginal geometry only when the matrix-valued covariance vanishes. Setting $w \equiv 1$ makes this term identically zero. □

Proof of Proposition 2. Write $\mathbf{P} = \widehat{\mathbf{M}}_{t,\lambda} \succ 0$. The objective in Eq. (5) is nonnegative, and $\mathbf{A} = \mathbf{P}^{-1/2}$ attains zero because $\mathbf{P}^{-1/2}\mathbf{P}\mathbf{P}^{-1/2} = \mathbf{I}$. If another SPD matrix $\widetilde{\mathbf{A}}$ attains zero, then $\widetilde{\mathbf{A}}\mathbf{P}\widetilde{\mathbf{A}} = \mathbf{I}$ and therefore $\mathbf{P} = \widetilde{\mathbf{A}}^{-2}$. Uniqueness of the SPD square root gives $\widetilde{\mathbf{A}} = \mathbf{P}^{-1/2}$. The log-Euclidean geodesic from $\mathbf{I}$ to $\mathbf{P}^{-1/2}$ is

$$\exp\Big((1-\gamma)\log\mathbf{I} + \gamma\log\mathbf{P}^{-1/2}\Big) = \exp\Big(-\frac{\gamma}{2}\log\mathbf{P}\Big) = \mathbf{P}^{-\gamma/2}, \tag{17}$$

which proves the stated interpolation; the same expression gives geodesic extrapolation for $\gamma > 1$. □

Idealized recovery of a latent angular distortion. The calibration objective does not generally identify a domain transform because a class-imbalanced or non-isotropic semantic mixture can itself make the marginal moment anisotropic. The following proposition states explicit sufficient conditions under which LGC approximately recovers an inverse angular distortion. Let latent directions satisfy $\mathbf{u} \in \mathbb{S}^{d-1}$ and $\mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \mathbf{I}/d$. For a shared-concentration vMF mixture, this isotropy follows when the class directions form a weighted tight frame, $\sum_c \pi_c \boldsymbol{\nu}_c \boldsymbol{\nu}_c^\top = \mathbf{I}/d$: rotational symmetry gives $\mathbb{E}[\mathbf{u}\mathbf{u}^\top \mid c] = a\mathbf{I} + b\boldsymbol{\nu}_c\boldsymbol{\nu}_c^\top$, and the trace constraint then yields the marginal $\mathbf{I}/d$.

Proposition 3 (Approximate inverse-distortion recovery)**.** *Let* $\mathbf{B} \succ 0$, $\mathbf{z} = T_{\mathbf{B}}(\mathbf{u})$, *and* $\bar{r} = \mathrm{tr}(\mathbf{B}^2)/d$. *Suppose that, for some* $0 \le \delta < 1$,

$$(1-\delta)\bar{r} \le \mathbf{u}^\top\mathbf{B}^2\mathbf{u} \le (1+\delta)\bar{r} \quad \textit{almost surely}. \tag{18}$$

Then the observed marginal moment $\mathbf{M}_{\mathbf{z}} = \mathbb{E}[\mathbf{z}\mathbf{z}^\top]$ *obeys*

$$\frac{\mathbf{B}^2}{(1+\delta)\,\mathrm{tr}(\mathbf{B}^2)} \preceq \mathbf{M}_{\mathbf{z}} \preceq \frac{\mathbf{B}^2}{(1-\delta)\,\mathrm{tr}(\mathbf{B}^2)}, \tag{19}$$

and consequently

$$\sqrt{(1-\delta)\,\mathrm{tr}(\mathbf{B}^2)}\,\mathbf{B}^{-1} \preceq \mathbf{M}_{\mathbf{z}}^{-1/2} \preceq \sqrt{(1+\delta)\,\mathrm{tr}(\mathbf{B}^2)}\,\mathbf{B}^{-1}. \tag{20}$$

Thus, as the normalization factor concentrates, the calibrated map approaches the inverse distortion up to a positive scale, which is immaterial to $T_{\mathbf{A}}$.

Proof of Proposition 3. Since $\mathbf{z} = \mathbf{B}\mathbf{u}/\sqrt{\mathbf{u}^\top\mathbf{B}^2\mathbf{u}}$,

$$\mathbf{M}_{\mathbf{z}} = \mathbf{B}\,\mathbb{E}\left[\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{B}^2\mathbf{u}}\right]\mathbf{B}. \tag{21}$$

Applying the scalar bounds in Eq. (18) to the positive-semidefinite matrix $\mathbf{B}\mathbf{u}\mathbf{u}^\top\mathbf{B}$ and using $\mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \mathbf{I}/d$ gives Eq. (19). Order reversal under inversion followed by operator monotonicity of the positive square root gives Eq. (20). Finally, $T_{c\mathbf{A}} = T_{\mathbf{A}}$ for every $c > 0$. □

Proof of Theorem 1. Let $\mathbf{u} = T_{\mathbf{A}}(\mathbf{z})$. By assumption, the transformed feature has class-conditional density

$$q_c(\mathbf{u}) = C_d(\kappa) \exp\big(\kappa \boldsymbol{\nu}_c^\top \mathbf{u}\big). \tag{22}$$

The angular map $T_{\mathbf{A}}$ is bijective on the sphere, with inverse $T_{\mathbf{A}^{-1}}$. The induced density on the original sphere is therefore

$$p_c(\mathbf{z}) = q_c(T_{\mathbf{A}}(\mathbf{z}))J_{\mathbf{A}}(\mathbf{z}), \tag{23}$$

where $J_{\mathbf{A}}(\mathbf{z})$ is the surface-measure Jacobian. This Jacobian depends on $\mathbf{z}$ and $\mathbf{A}$ but not on the class label. With equal class priors, all class-independent terms cancel in the Bayes decision rule, giving

$$\arg\max_c \log p_c(\mathbf{z}) = \arg\max_c \boldsymbol{\nu}_c^\top T_{\mathbf{A}}(\mathbf{z}). \tag{24}$$

If $\boldsymbol{\nu}_c = T_{\mathbf{A}}(\boldsymbol{\mu}_c)$, then

$$\boldsymbol{\nu}_c^\top T_{\mathbf{A}}(\mathbf{z}) = \frac{\mathbf{z}^\top \mathbf{A}^\top \mathbf{A}\boldsymbol{\mu}_c}{\sqrt{\mathbf{z}^\top \mathbf{A}^\top \mathbf{A}\mathbf{z}}\sqrt{\boldsymbol{\mu}_c^\top \mathbf{A}^\top \mathbf{A}\boldsymbol{\mu}_c}}, \tag{25}$$

which proves Eq. (2). □

vMF background. For a unit vector $\mathbf{z} \in \mathbb{S}^{d-1}$, the von Mises–Fisher density is

$$p(\mathbf{z} \mid \boldsymbol{\mu}, \kappa) = C_d(\kappa) \exp\big(\kappa\, \boldsymbol{\mu}^\top \mathbf{z}\big), \qquad \boldsymbol{\mu} \in \mathbb{S}^{d-1}. \tag{26}$$

Up to a class-independent constant, its log-likelihood is proportional to cosine similarity. This is why a cosine classifier with normalized CLIP features can be read as a vMF discriminant whose class directions are initialized by text prototypes.

Angular change of variables. Let $\mathbf{A}$ be any positive definite matrix and define

$$T_{\mathbf{A}}(\mathbf{z}) = \frac{\mathbf{A}\mathbf{z}}{\|\mathbf{A}\mathbf{z}\|_2}, \qquad \mathbf{z} \in \mathbb{S}^{d-1}. \tag{27}$$

Then $T_{\mathbf{A}}$ maps the sphere to itself and is a bijection with inverse $T_{\mathbf{A}^{-1}}$. If the transformed angular variable $\mathbf{u} = T_{\mathbf{A}}(\mathbf{z})$ follows a class-conditional vMF density,

$$q_c(\mathbf{u}) = C_d(\kappa) \exp\big(\kappa\, \boldsymbol{\nu}_c^\top \mathbf{u}\big), \qquad \mathbf{u} \in \mathbb{S}^{d-1}, \tag{28}$$

where $\boldsymbol{\nu}_c = T_{\mathbf{A}}(\boldsymbol{\mu}_c)$ is the transformed class direction, then the induced density on the original sphere is

$$p_c(\mathbf{z}) = q_c(T_{\mathbf{A}}(\mathbf{z}))\, J_{\mathbf{A}}(\mathbf{z}), \qquad J_{\mathbf{A}}(\mathbf{z}) = \frac{|\det \mathbf{A}|}{\|\mathbf{A}\mathbf{z}\|_2^d}, \tag{29}$$

where $J_{\mathbf{A}}$ is the surface-measure Jacobian of the angular map. Crucially, this Jacobian depends on $\mathbf{z}$ and $\mathbf{A}$ but not on the class label. Therefore, for equal class priors π_c,

$$\arg\max_c \{\log p_c(\mathbf{z}) + \log \pi_c\} = \arg\max_c \boldsymbol{\nu}_c^\top T_{\mathbf{A}}(\mathbf{z}). \tag{30}$$

The Jacobian thus cancels in classification. This is the statistical meaning of Eq. (7): whitening followed by re-projection defines angular coordinates on the sphere, and vMF scoring is the usual discriminant in those coordinates.

Consistency of the geometry estimator. STRA estimates the uncentered second moment by

$$\widehat{\mathbf{M}}_t = \frac{1}{t}\sum_{s=1}^{t} \mathbf{z}_s \mathbf{z}_s^\top. \tag{31}$$

For a stationary ergodic target stream with finite second moment, the strong law gives

$$\widehat{\mathbf{M}}_t \xrightarrow{a.s.} \mathbf{M}. \tag{32}$$

Because the fixed ridge $\lambda > 0$ keeps the spectrum bounded away from zero, the matrix power map is continuous on the positive definite cone, and therefore

$$\big(\widehat{\mathbf{M}}_t + \lambda \mathbf{I}\big)^{-\gamma/2} \xrightarrow{a.s.} \big(\mathbf{M} + \lambda \mathbf{I}\big)^{-\gamma/2}. \tag{33}$$

Thus the streaming moment estimator consistently estimates the corresponding population power-calibration operator under stationarity. This statement alone does not identify a true domain transform: semantic class structure and class priors may also make the marginal moment anisotropic. Proposition 3 gives explicit sufficient conditions for approximate inverse-distortion recovery in the unregularized population limit; the finite ridge used by STRA trades a small asymptotic bias for numerical stability. Separately, in the idealized case where pseudo-label weights equal the true class responsibilities and prototype blending/decay are removed, STRA aggregates $\sum_{s\in c}\mathbf{z}_s$ in the original coordinates and then applies $T_{\mathbf{A}}$. Equivalently, in transformed angular coordinates $\mathbf{u}_s = T_{\mathbf{A}}(\mathbf{z}_s)$, this is a norm-weighted vMF direction estimator, $\Pi(\sum_{s\in c}\|\mathbf{A}\mathbf{z}_s\|_2\mathbf{u}_s)$, rather than the unweighted MLE $\Pi(\sum_{s\in c}\mathbf{u}_s)$. The practical STRA update keeps text-prior blending and decay as stability devices against finite-stream pseudo-label noise.

Relation to shared-covariance Gaussian classifiers. With $\mathbf{A} = (\widehat{\mathbf{M}}_t + \lambda\mathbf{I})^{-\gamma/2}$, the STRA spherical score before the shared scale κ^* can be written as

$$s_\gamma(\mathbf{z}, \boldsymbol{\mu}_c) = \cos(T_{\mathbf{A}}(\mathbf{z}), T_{\mathbf{A}}(\boldsymbol{\mu}_c)) = \frac{\mathbf{z}^\top \mathbf{M}_\gamma \boldsymbol{\mu}_c}{\sqrt{\mathbf{z}^\top \mathbf{M}_\gamma \mathbf{z}}\sqrt{\boldsymbol{\mu}_c^\top \mathbf{M}_\gamma \boldsymbol{\mu}_c}}, \qquad \mathbf{M}_\gamma = (\widehat{\mathbf{M}}_t + \lambda\mathbf{I})^{-\gamma}. \tag{34}$$

A shared-covariance Gaussian classifier uses the LDA discriminant

$$g_c(\mathbf{z}) = \mathbf{z}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\boldsymbol{\mu}_c^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log \pi_c, \tag{35}$$

where $\boldsymbol{\Sigma}$ is the pooled within-class centered covariance and π_c is the class prior. Equation (34) shows the algebraic connection: $\gamma = 1$ uses a Mahalanobis-style metric in the bilinear numerator, followed by angular normalization on the sphere.

B Complete Results for Both Backbones

This section reports the complete results for both CLIP backbones, complementing the compact two-block tables in Section 4.2. Tables 8–10 give the full CLIP-ViT-B/16 comparison and Tables 11–13 give the corresponding CLIP-RN50 comparison, in both cases for natural distribution shift, fine-grained categorization, and corruption robustness, in that order. Unlike the main-text tables, which list only the methods evaluated on all three tasks, here every baseline named in Section 4.1 is included whenever its numbers are available, so the two backbones have different coverage: the ViT-B/16 tables reproduce the full comparison used in the literature, whereas the RN50 tables cover the subset of baselines run under our matched batch-size-one protocol. Best and second-best entries are marked *within each table*, so a method that is best among the smaller main-text subset need not be best here. ✓/× = with/without backpropagation, and "–" denotes methods that use labeled target data or do not perform test-time adaptation. All STRA results are averaged over five random test-stream orders.

C Additional Online Dynamics

Figure 6 complements the ViT-B/16 task-level online curves in Figure 4 with the corresponding RN50 results. Figures 7 and 8 further break down the online trajectories at the dataset level for ViT-B/16 and RN50, respectively. In these figures, Task 1 denotes natural distribution shift, Task 2 denotes fine-grained classification, and Task 3 denotes corruption robustness; shaded regions indicate the interquartile range across five random stream orders.

D Additional Geometry Spectrum Diagnostics

Figures 9–11 provide additional spectrum diagnostics beyond the task-level ViT-B/16 visualization in the main text. They report per-dataset spectra for ViT-B/16, per-dataset spectra for RN50, and task-level spectra for RN50, respectively.

E Additional Implementation Details

This section summarizes the fixed configuration used by STRA. Following the style of recent closed-form CLIP-TTA work, the main text keeps the experimental protocol compact; for reproducibility,

Table 8: Complete Top-1 accuracy (%) on the natural distribution shift benchmark (ImageNet + IN-A/V/R/S) using CLIP ViT-B/16. Best results are in bold and second-best results are underlined.

Method	BP-free	IN	IN-A	IN-V	IN-R	IN-S	**Shift Avg.**	**All Avg.**
CLIP	–	66.73	47.87	60.86	73.98	46.09	57.20	59.11
CoOp	–	71.51	49.71	64.20	75.21	47.99	59.28	61.72
CoCoOp	–	71.02	50.63	64.07	76.18	48.75	59.91	62.13
Tip-Adapter	×	70.75	51.04	63.41	77.76	48.88	60.27	62.37
TPT	×	68.98	54.77	63.45	77.06	47.94	60.81	62.44
DiffTPT	×	70.30	55.68	65.10	75.00	46.80	60.65	62.58
C-TPT	×	68.50	51.60	62.70	76.00	47.90	59.55	61.34
DMN	×	72.25	58.28	65.17	78.55	53.20	63.80	65.49
DPE	×	71.91	59.63	65.44	80.40	52.26	64.43	65.93
TPS	×	70.38	59.21	63.80	77.49	49.57	62.52	64.09
DynaPrompt	×	69.61	56.17	64.67	78.17	48.22	61.81	63.37
B2TPT	×	69.57	55.26	65.40	78.64	49.53	62.21	63.68
MTA	✓	70.08	58.06	64.24	78.33	49.61	62.56	64.06
TDA	✓	69.51	60.11	64.67	80.24	50.54	63.89	65.01
ZERO	✓	69.31	59.61	64.16	77.22	48.40	62.35	63.74
AWT	✓	71.32	60.33	65.15	80.64	51.60	64.43	65.81
RA-TTA	✓	70.58	59.21	64.16	79.68	50.83	63.47	64.89
BCA	✓	70.22	61.14	64.90	80.72	50.87	64.41	65.57
TCA	✓	68.88	50.13	62.10	77.11	48.95	59.57	61.43
Dota	✓	70.68	61.19	64.41	81.17	51.33	64.53	65.76
SCA	✓	71.75	60.33	65.38	80.85	52.50	64.77	66.16
ADAPT	✓	70.91	63.32	64.64	80.66	53.13	65.44	66.53
Ours	✓	**72.91**	**64.58**	**65.77**	**81.50**	**53.73**	**66.40**	**67.70**

Table 9: Complete Top-1 accuracy (%) on the fine-grained categorization benchmark (10 datasets) using CLIP ViT-B/16. Best results are in bold and second-best results are underlined.

Method	BP-free	Aircraft	Caltech	Cars	DTD	EuroSAT	Flowers	Food101	Pets	SUN397	UCF101	Avg.
CLIP	–	23.67	93.35	65.48	44.27	42.01	67.44	83.65	88.25	62.59	65.13	63.58
CoOp	–	18.47	93.70	64.51	41.92	46.39	68.71	85.30	89.14	64.15	66.55	63.88
CoCoOp	–	22.29	93.79	64.90	45.45	39.23	70.85	83.97	90.46	66.89	68.44	64.63
TPT	×	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT	×	25.60	92.49	67.01	47.00	43.13	70.10	87.23	88.22	65.74	68.22	65.47
C-TPT	×	24.00	93.60	65.80	46.00	43.20	69.80	83.70	88.20	64.80	65.70	64.48
DMN	×	**30.03**	**95.38**	67.96	55.85	59.43	74.49	85.08	92.04	70.18	72.51	70.30
TPS	×	26.27	94.56	67.00	53.80	42.11	71.69	84.78	87.82	68.25	71.18	66.75
DPE	×	28.95	94.81	67.31	54.20	55.79	75.07	86.17	91.14	70.07	70.44	69.40
HisTPT	×	26.90	94.50	69.20	48.90	49.70	71.20	**89.30**	89.10	67.20	70.10	67.61
DynaPrompt	×	24.33	94.32	67.65	47.96	42.28	69.95	85.42	88.28	66.32	68.72	65.52
MTA	✓	25.32	94.13	66.36	45.59	38.71	68.26	84.95	88.22	64.98	68.11	64.46
TDA	✓	23.91	94.24	67.28	47.40	58.00	71.42	86.14	88.63	67.62	70.66	67.53
ZLaP	✓	25.40	93.10	65.60	48.60	55.60	73.50	86.90	87.10	67.40	71.50	67.47
ZERO	✓	25.21	93.66	68.04	46.12	34.33	67.68	86.53	87.75	65.03	67.77	64.21
BCA	✓	28.59	94.69	66.86	53.49	56.63	73.12	85.97	90.43	68.41	67.59	68.58
OGA	✓	23.20	93.60	68.10	47.90	54.20	69.20	85.60	89.40	67.90	71.40	67.05
TCA	✓	24.87	93.63	65.33	46.16	**70.43**	73.33	85.31	89.53	65.92	72.38	68.69
Dota	✓	25.59	94.32	69.48	47.87	57.65	74.67	87.02	91.69	69.70	72.06	69.01
TT-RAA	✓	25.38	94.08	66.42	47.99	66.12	72.68	86.09	89.83	67.69	71.29	68.76
ADAPT	✓	27.99	93.87	69.08	55.38	58.88	75.27	85.32	92.12	70.31	73.88	70.21
SCA	✓	28.50	94.85	68.49	**57.09**	57.16	76.09	86.09	91.44	70.27	73.43	70.34
Ours	✓	27.33	94.71	**70.14**	56.49	60.39	**78.83**	87.01	**92.79**	**71.33**	**75.26**	**71.43**

Table 10: Complete Top-1 accuracy (%) on the corruption robustness benchmark (15 ImageNet-C corruptions at severity 5) using CLIP ViT-B/16. Best results are in bold and second-best results are underlined.

Method	Blur				Weather				Digital				Noise			Avg.
	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brig.	Cont.	Elas.	Pix.	JPEG	Gauss.	Shot	Impu.	
CLIP	24.25	15.71	24.46	22.60	33.08	31.06	37.61	55.62	17.11	13.43	33.04	33.70	13.25	14.16	13.48	25.50
TPT	27.56	15.48	26.16	26.94	36.74	34.28	39.38	60.22	16.96	15.64	**40.74**	37.90	10.64	11.94	10.92	27.43
DiffTPT	25.63	16.96	26.74	25.40	35.99	34.57	39.83	59.01	17.32	17.16	38.43	35.47	12.97	13.60	13.21	27.49
TDA	26.53	17.91	27.35	25.90	36.50	34.84	40.53	58.57	20.16	16.62	35.65	36.69	15.42	16.46	16.03	28.34
DMN	26.06	17.19	26.61	25.23	34.81	33.48	38.93	58.70	19.38	15.40	35.32	36.49	14.33	15.33	14.69	27.46
SCA	26.31	18.13	27.15	25.14	35.70	34.09	40.62	60.25	19.96	16.46	37.27	37.57	15.77	16.94	15.99	28.49
ADAPT	26.30	18.01	27.31	25.54	36.19	34.67	40.96	60.29	19.95	16.09	37.44	37.22	15.76	16.84	15.90	28.56
Ours	**28.10**	**19.75**	**28.81**	**27.22**	**38.39**	**36.92**	**42.97**	**61.18**	**21.13**	**19.40**	38.97	**37.99**	**17.88**	**18.90**	**18.92**	**30.43**

Table 11: Complete Top-1 accuracy (%) on the natural distribution shift benchmark (ImageNet + IN-A/V/R/S) using CLIP RN50. Best results are in bold and second-best results are underlined.

Method	BP-free	IN	IN-A	IN-V	IN-R	IN-S	**Shift Avg.**	**All Avg.**
CLIP	–	58.16	21.83	51.41	56.15	33.37	40.69	44.18
CoOp	–	63.33	23.06	55.40	56.60	34.67	42.43	46.61
CoCoOp	–	62.81	23.32	55.72	57.74	34.48	42.82	46.81
TPT	×	60.74	26.67	54.70	59.11	35.09	43.89	47.26
DiffTPT	×	60.80	31.06	55.80	58.80	37.10	45.69	48.71
C-TPT	×	60.20	23.40	54.70	58.00	35.10	42.80	46.28
DMN	×	63.87	28.57	56.12	61.44	39.84	46.49	49.97
TPS	×	62.27	29.80	**60.04**	55.49	35.74	45.27	48.67
DPE	×	63.41	30.15	56.72	**63.72**	40.03	47.66	50.81
DynaPrompt	×	61.56	27.84	55.12	60.63	35.64	44.81	48.16
BCA	✓	61.81	30.35	56.58	62.89	38.04	46.97	49.93
TDA	✓	61.35	30.29	55.54	62.58	38.12	46.63	49.58
Dota	✓	61.82	30.81	55.27	62.81	37.52	46.60	49.65
SCA	✓	63.11	30.64	56.51	62.85	39.33	47.33	50.49
ADAPT	✓	62.16	33.08	55.97	62.69	40.21	47.99	50.82
Ours	✓	**64.39**	**34.68**	57.07	63.38	**40.68**	**48.95**	**52.04**

Table 12: Complete Top-1 accuracy (%) on the fine-grained categorization benchmark (10 datasets) using CLIP RN50. Best results are in bold and second-best results are underlined.

Method	BP-free	Aircraft	Caltech	Cars	DTD	EuroSAT	Flowers	Food101	Pets	SUN397	UCF101	Avg.
CLIP	–	15.66	85.88	55.70	40.37	23.69	61.75	73.97	83.57	58.80	58.84	55.82
CoOp	–	15.12	86.53	55.32	37.29	26.20	61.55	75.59	87.00	58.15	59.05	56.18
CoCoOp	–	14.61	87.38	56.22	38.53	28.73	65.57	76.20	**88.39**	59.61	57.10	57.23
TPT	×	17.58	87.02	58.46	40.84	28.33	62.69	74.88	84.49	61.46	60.82	57.66
DiffTPT	×	17.60	86.89	**60.71**	40.72	41.04	63.53	**79.21**	83.40	62.72	62.67	59.85
C-TPT	×	17.00	86.90	56.50	42.20	27.80	65.20	74.70	84.10	61.00	59.70	57.51
DMN	×	**22.77**	90.14	60.02	50.41	48.72	67.93	76.70	86.78	64.39	65.34	63.32
TPS	×	18.30	89.80	59.40	48.40	24.30	68.20	76.20	84.40	62.70	64.30	59.60
DPE	×	19.80	90.83	59.26	50.18	41.67	67.60	77.83	85.97	64.23	61.98	61.94
BCA	✓	19.89	89.70	58.13	48.58	42.12	66.30	77.19	85.58	63.38	63.51	61.44
TDA	✓	17.61	89.70	57.78	43.74	42.11	68.74	77.75	86.18	62.53	64.18	61.03
Dota	✓	18.06	88.84	58.72	45.80	47.15	68.53	78.61	87.33	63.89	65.08	62.20
SCA	✓	20.82	**91.03**	59.73	53.31	43.33	71.09	76.51	86.15	64.50	66.40	63.29
ADAPT	✓	20.10	88.72	59.94	52.60	**49.70**	68.33	76.06	87.33	65.24	65.77	63.38
Ours	✓	20.56	89.88	60.56	**54.37**	44.48	**71.81**	78.28	88.02	**65.87**	**66.67**	**64.05**

Table 13: Complete Top-1 accuracy (%) on the corruption robustness benchmark (15 ImageNet-C corruptions at severity 5) using CLIP RN50. Best results are in bold and second-best results are underlined.

Method	Blur				Weather				Digital				Noise			Avg.
	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brig.	Cont.	Elas.	Pix.	JPEG	Gauss.	Shot	Impu.	
CLIP	9.54	3.40	7.46	12.62	12.29	15.72	22.08	41.69	6.24	4.67	11.01	14.24	2.43	3.07	2.52	11.27
TPT	8.02	2.74	5.34	10.97	10.59	12.92	16.17	35.67	4.45	3.73	11.56	16.68	1.43	1.94	1.42	9.58
DiffTPT	10.50	3.90	8.62	12.74	11.31	15.03	19.64	37.73	4.98	4.74	13.58	16.56	2.69	3.59	2.08	11.18
TDA	9.84	4.40	7.38	13.74	13.74	17.16	23.76	44.16	7.00	5.79	11.24	15.26	2.54	3.26	2.72	12.13
SCA	**11.88**	4.96	9.28	14.91	15.45	19.10	25.78	46.42	8.42	6.91	13.63	17.26	3.13	3.85	3.41	13.63
ADAPT	10.54	4.44	8.57	14.34	13.85	17.84	24.56	45.67	7.76	5.85	11.96	15.86	2.91	3.77	2.92	12.72
Ours	11.58	**5.75**	**9.87**	**16.47**	**16.33**	**20.67**	**28.04**	**47.37**	**8.55**	**8.65**	**14.23**	**17.66**	**3.89**	**4.65**	**4.07**	**14.52**

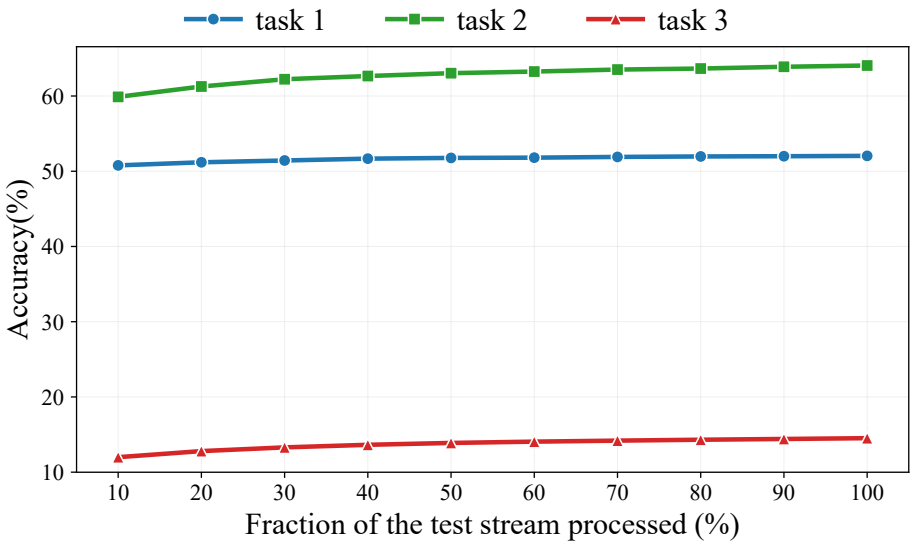


Figure 6: Task-level online adaptation dynamics of STRA with CLIP RN50.

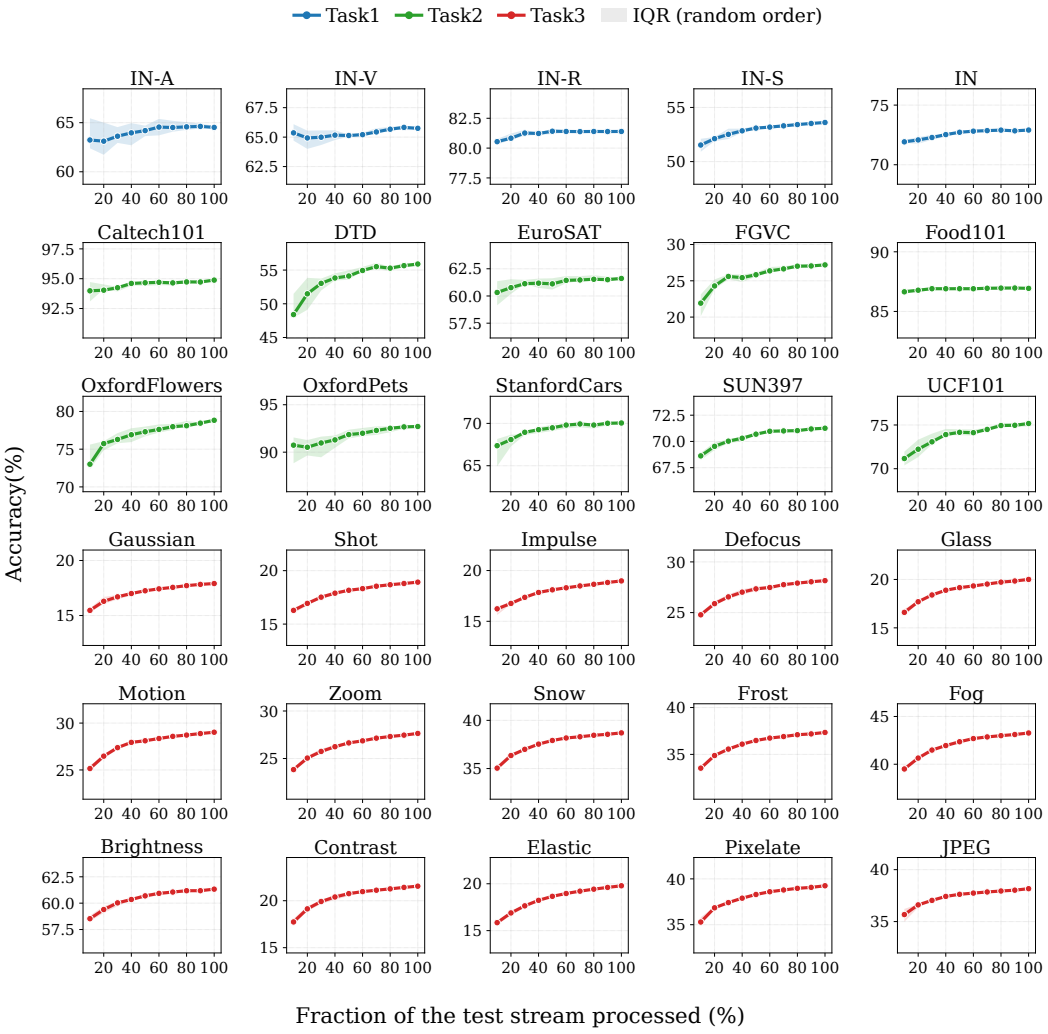


Figure 7: Per-dataset online adaptation dynamics of STRA with CLIP ViT-B/16.

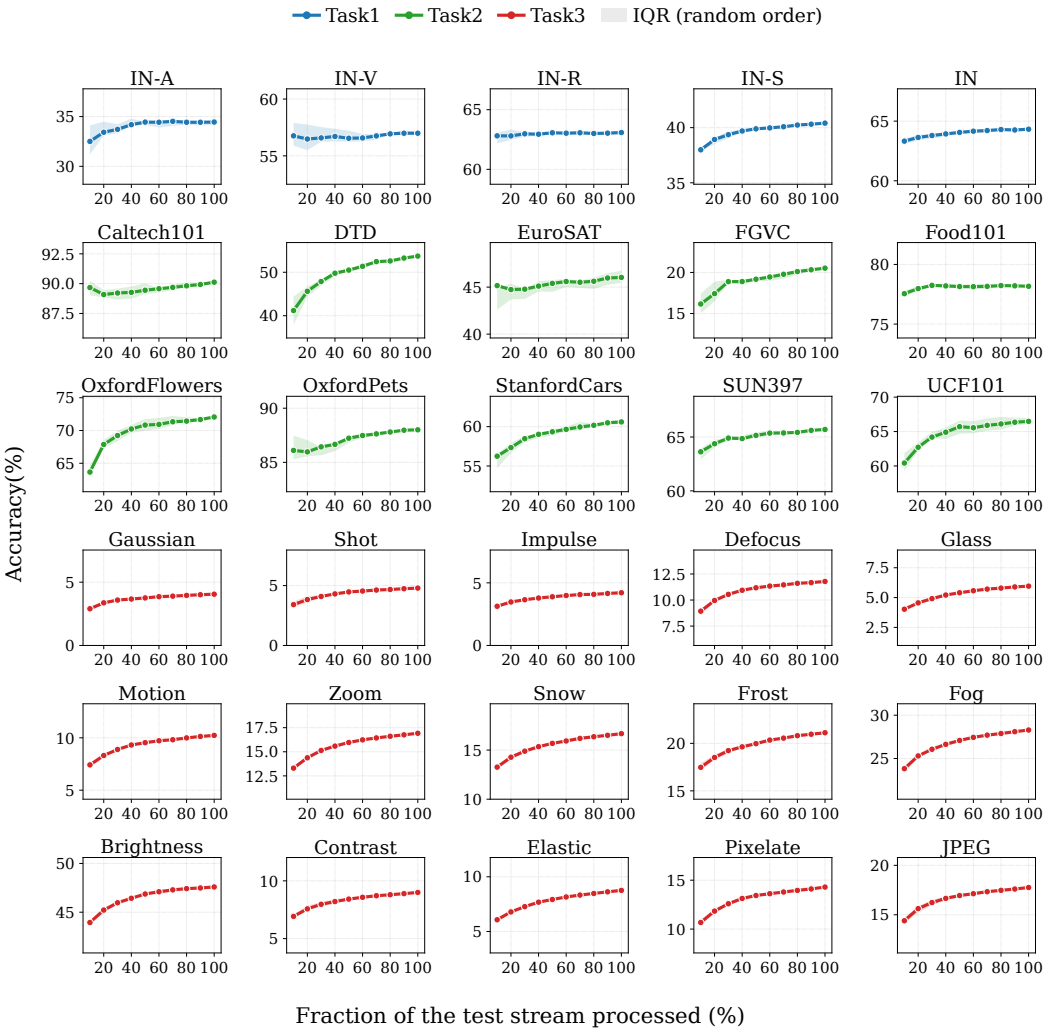


Figure 8: Per-dataset online adaptation dynamics of STRA with CLIP RN50.

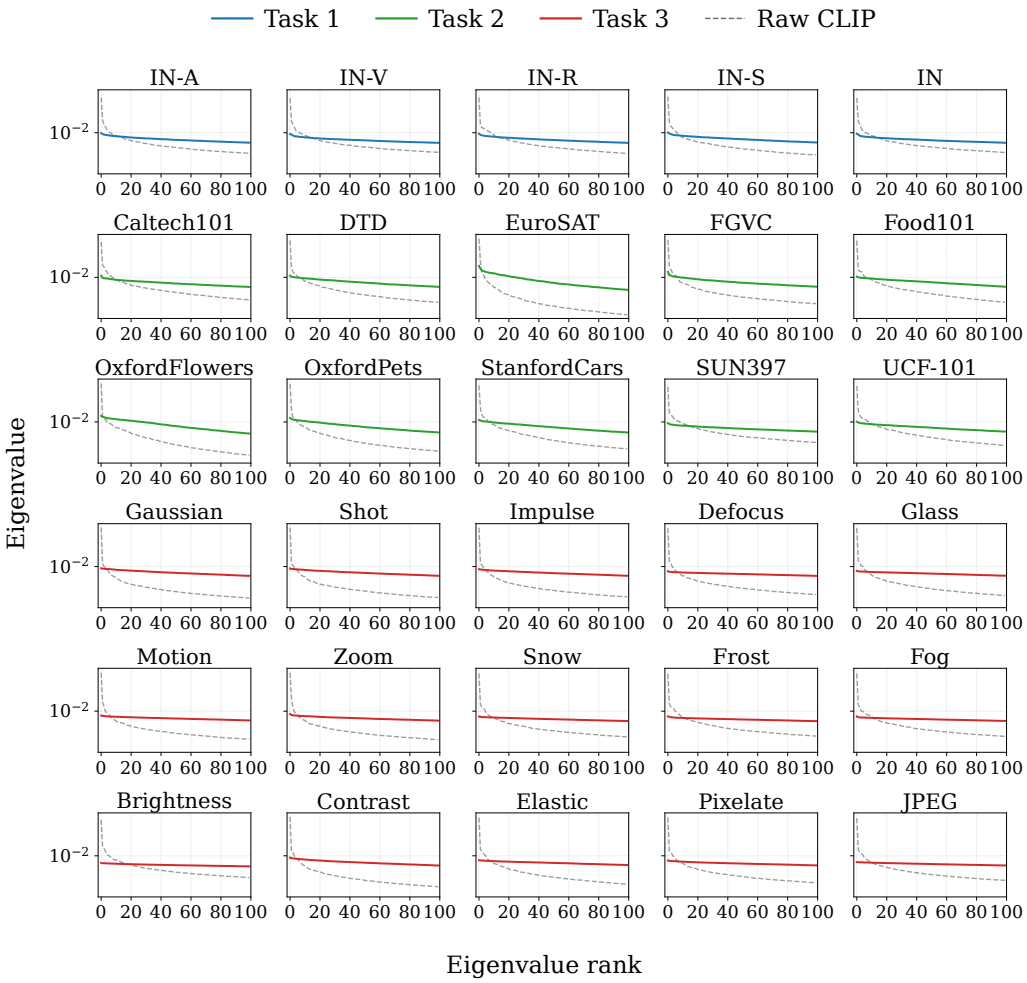


Figure 9: Per-dataset geometry spectrum diagnostics with CLIP ViT-B/16.

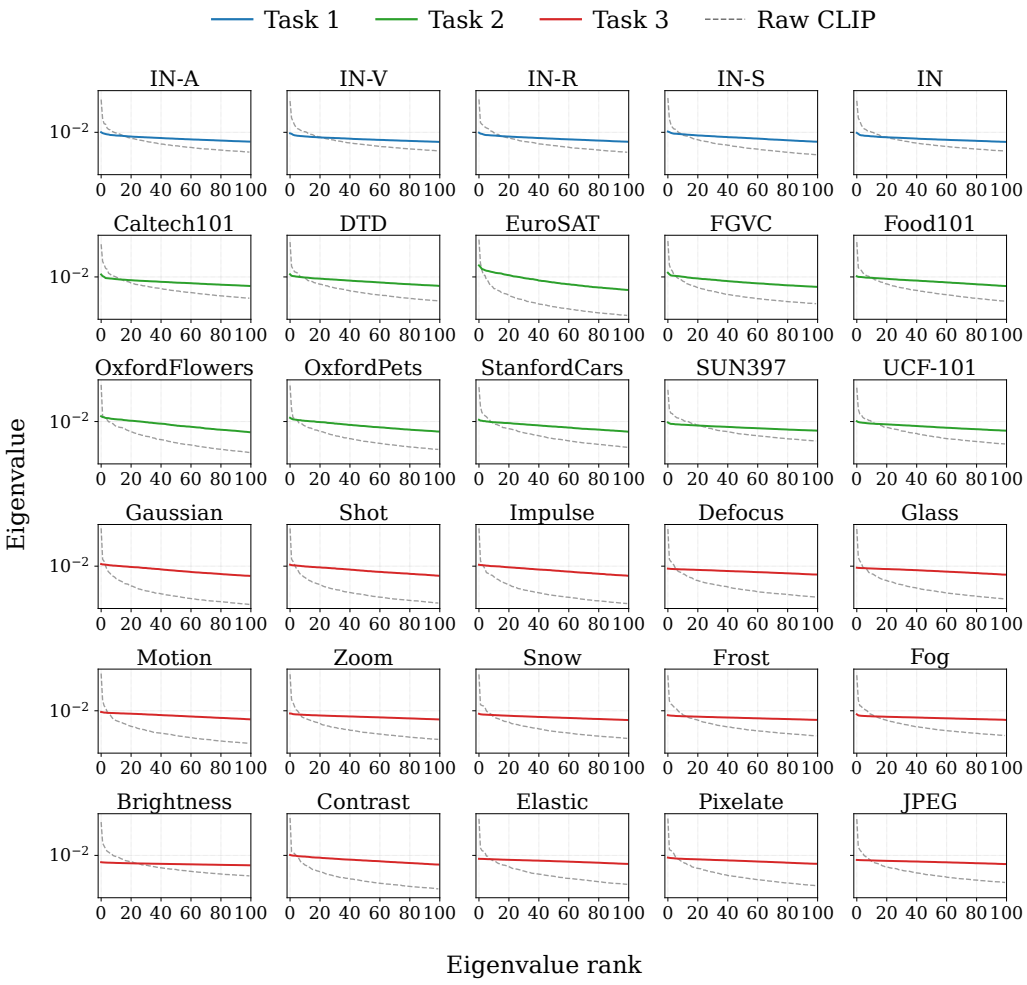


Figure 10: Per-dataset geometry spectrum diagnostics with CLIP RN50.

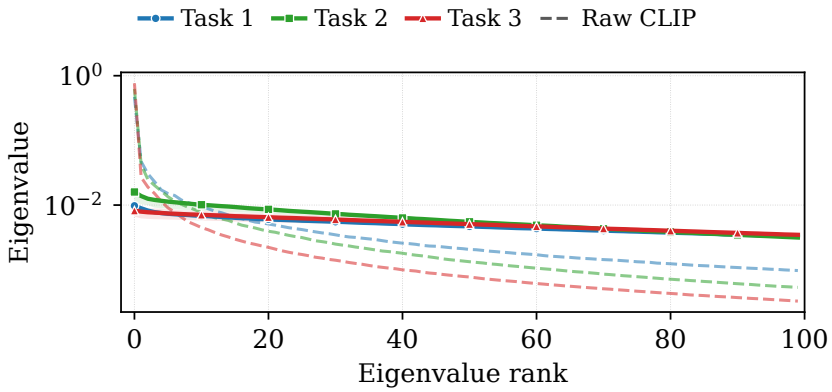


Figure 11: Task-level geometry spectrum diagnostics with CLIP RN50.

we list all implementation constants here. All remaining quantities are fixed globally across datasets, corruption types, stream orders, and backbones. They implement numerical stabilization, pseudo-label smoothing, lightweight prototype blending, entropy calibration, and amortized refreshes rather than validation-set tuning or dataset-specific model selection. Unless otherwise noted, each main benchmark number reported for STRA is the mean over five random test-stream orders. For the non-i.i.d. stream-order analysis in Table 7, the random row uses the same five-seed averages, class-blocked uses five shuffled class-block orders, and the entropy-sorted easy-to-hard/hard-to-easy orders are evaluated with seed 0 because the ordering is determined by frozen CLIP entropy.

Rationale for the fixed configuration. The hyperparameters in Table 14 are chosen as role-based defaults rather than selected per benchmark. The default $\gamma = 1$ is theoretically distinguished: by Proposition 2, it gives the unique SPD minimizer $(\widehat{\mathbf{M}} + \lambda \mathbf{I})^{-1/2}$ of the pre-projection marginal-moment calibration objective. Values $0 < \gamma < 1$ interpolate from the original CLIP geometry toward this solution along the log-Euclidean geodesic, while $\gamma > 1$ extrapolates beyond it. This choice does not assert that $\gamma = 1$ minimizes downstream classification error; indeed, severe corruptions in Task 3 can prefer extrapolation because the final spherical re-projection and class discrimination are not optimized by the marginal-moment surrogate. We nevertheless retain $\gamma = 1$ as a principled dataset-independent configuration and do not tune it per benchmark. The ridge $\lambda = 10^{-3}$ stabilizes the matrix power and makes the empirical moment positive definite. For the prototype update, $\alpha = 0.90$ makes the data-induced spherical direction dominant while retaining a small frozen-text anchor against pseudo-label drift, and $\rho = 0.999$ gives a long-memory online estimate suitable for batch-size-one streams. The truncation parameter $\tau_0 = 0.1$ removes low-probability softmax tails while preserving ambiguity among the most plausible classes. For the readout, $\kappa^* = 100$ follows the CLIP cosine-logit scale, $\beta = -0.5$ smoothly favors the adapted branch when it has lower entropy than the zero-shot branch, and $\rho_H = 0.99$ stabilizes the online entropy calibration. The prototype and whitening refresh intervals amortize computation while keeping the streaming statistics updated. The sensitivity analysis in Figure 5 shows broad stable regions for α, β, and γ, indicating that STRA is not driven by a narrow hyperparameter choice.

Table 14: Fixed STRA configuration used for all datasets and backbones.

Quantity	Role	Setting
γ	Geometry path; exact pre-projection moment calibration at 1	1
λ	Ridge for $(\widehat{\mathbf{M}} + \lambda \mathbf{I})^{-\gamma/2}$	10^{-3}
α	Data/text prototype blending coefficient	0.90
ρ	Class-conditional EMA decay for first moments	0.999
τ_0	Base threshold for truncated soft labels	0.1
κ^*	vMF/cosine logit scale	100
β	Entropy-fusion slope in the gating rule	-0.5
ρ_H	EMA decay for branch-entropy calibration	0.99
$\tau^{(t)}$	Branch-calibration temperature	estimated online from entropy EMAs
Prototype refresh	Prototype recomputation	confident sample, every 10 samples, or t=1
Confidence trigger	Normalized zero-shot entropy threshold	$H^{\text{zs}}/\log C < 0.5$
Whitening refresh	Amortized eigendecomposition interval	every 100 samples
Minimum class evidence	Fallback-to-text threshold	0.5 accumulated soft-label mass

F Five-Seed Variability for ViT-B/16

This section reports the seed-level values underlying the ViT-B/16 main results. Each entry is Top-1 accuracy (%) for one random stream order.

G Five-Seed Variability for RN50

This section reports the seed-level values underlying the RN50 appendix results. Each entry is Top-1 accuracy (%) for one random stream order. As in Appendix F, we report the mean and standard deviation across orders as two separate summary rows.

Table 15: Five-order variability on ImageNet and natural distribution shifts with CLIP ViT-B/16.

Order	IN	IN-A	IN-V	IN-R	IN-S
1	72.89	64.33	66.30	81.56	53.62
2	72.81	64.82	65.53	81.64	53.61
3	72.95	64.38	65.56	81.52	53.82
4	72.92	64.70	65.73	81.44	53.72
5	72.99	64.69	65.71	81.33	53.89
Mean	72.91	64.58	65.77	81.50	53.73
Std.	0.06	0.19	0.28	0.11	0.11

Table 16: Five-order variability on fine-grained datasets with CLIP ViT-B/16.

Order	Aircraft	Caltech	Cars	DTD	EuroSAT	Flowers	Food101	Pets	SUN397	UCF101
1	27.27	94.65	70.07	55.91	60.35	79.09	86.98	92.97	71.36	74.91
2	27.24	94.81	70.18	56.56	60.64	79.01	87.02	92.56	71.43	74.94
3	27.90	94.97	69.95	57.03	60.33	78.52	87.08	92.67	71.30	75.60
4	27.27	94.44	70.35	56.03	60.53	78.68	86.93	92.94	71.35	75.52
5	26.94	94.69	70.14	56.91	60.10	78.85	87.06	92.80	71.22	75.31
Mean	27.33	94.71	70.14	56.49	60.39	78.83	87.01	92.79	71.33	75.26
Std.	0.31	0.18	0.13	0.45	0.19	0.21	0.05	0.16	0.07	0.29

Table 17: Five-order variability on ImageNet-C at severity level 5 with CLIP ViT-B/16.

Order	Blur				Weather				Digital				Noise		
	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brig.	Cont.	Elas.	Pix.	JPEG	Gauss.	Shot	Impu.
1	28.13	19.69	28.79	27.38	38.44	36.90	43.07	61.37	21.02	19.40	38.94	38.00	17.86	18.93	18.82
2	28.18	19.74	28.80	27.13	38.36	36.89	42.93	61.11	21.03	19.43	38.96	37.84	17.87	18.93	18.94
3	27.98	19.75	28.81	27.20	38.38	36.99	42.90	61.12	21.20	19.34	38.96	38.07	17.89	18.84	18.91
4	28.07	19.70	28.81	27.13	38.32	36.83	43.03	61.17	21.19	19.34	38.90	38.06	17.90	18.92	19.00
5	28.15	19.84	28.84	27.25	38.44	37.00	42.92	61.13	21.23	19.49	39.09	37.98	17.85	18.90	18.92
Mean	28.10	19.75	28.81	27.22	38.39	36.92	42.97	61.18	21.13	19.40	38.97	37.99	17.88	18.90	18.92
Std.	0.07	0.05	0.02	0.09	0.05	0.07	0.07	0.10	0.09	0.06	0.06	0.08	0.02	0.03	0.06

Table 18: Five-order variability on ImageNet and natural distribution shifts with CLIP RN50.

Order	IN	IN-A	IN-V	IN-R	IN-S
1	64.50	35.12	57.08	63.50	40.60
2	64.45	34.84	57.16	63.25	40.56
3	64.56	34.54	56.90	63.53	40.63
4	64.17	34.48	57.04	63.25	40.89
5	64.28	34.44	57.18	63.37	40.73
Mean	64.39	34.68	57.07	63.38	40.68
Std.	0.14	0.26	0.10	0.12	0.12

Table 19: Five-order variability on fine-grained datasets with CLIP RN50.

Order	Aircraft	Caltech	Cars	DTD	EuroSAT	Flowers	Food101	Pets	SUN397	UCF101
1	20.55	89.78	60.88	53.90	45.89	71.99	78.21	87.95	65.93	66.75
2	20.46	89.78	60.96	54.61	43.44	71.74	78.27	88.25	66.00	66.93
3	21.03	90.06	60.22	54.73	44.21	72.27	78.28	88.25	65.84	67.20
4	20.67	89.98	60.22	53.84	45.14	71.38	78.33	88.17	65.52	66.30
5	20.10	89.82	60.55	54.79	43.73	71.70	78.33	87.49	66.05	66.19
Mean	20.56	89.88	60.56	54.37	44.48	71.81	78.28	88.02	65.87	66.67
Std.	0.30	0.12	0.32	0.41	0.91	0.30	0.05	0.29	0.19	0.38

Table 20: Five-order variability on ImageNet-C at severity level 5 with CLIP RN50.

Order	Blur				Weather				Digital				Noise		
	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brig.	Cont.	Elas.	Pix.	JPEG	Gauss.	Shot	Impu.
1	11.55	5.67	9.85	16.41	16.43	20.68	27.87	47.37	8.56	8.62	14.18	17.52	3.86	4.63	4.06
2	11.59	5.79	9.79	16.57	16.16	20.76	28.02	47.48	8.63	8.65	14.25	17.72	3.90	4.69	4.14
3	11.57	5.82	9.91	16.36	16.40	20.68	28.07	47.35	8.52	8.70	14.24	17.66	3.92	4.63	3.99
4	11.64	5.69	9.86	16.57	16.34	20.61	28.09	47.29	8.56	8.63	14.21	17.70	3.88	4.69	4.07
5	11.55	5.79	9.91	16.45	16.32	20.61	28.15	47.36	8.48	8.65	14.29	17.71	3.87	4.60	4.09
Mean	11.58	5.75	9.87	16.47	16.33	20.67	28.04	47.37	8.55	8.65	14.23	17.66	3.89	4.65	4.07
Std.	0.03	0.06	0.05	0.09	0.09	0.05	0.10	0.06	0.05	0.03	0.04	0.07	0.02	0.04	0.05

H Full Stream-Order Sensitivity Results

This section expands the task-level stream-order results in Table 7 to individual datasets and corruption types. All entries are Top-1 accuracy (%) using CLIP ViT-B/16. The random row corresponds to the five-seed averages used in the main benchmark. Class-blocked results are averaged over five seeds, where labels are used only to construct the diagnostic stream and both the class order and within-class sample order are shuffled. Easy-to-hard and hard-to-easy orders are obtained by sorting samples by frozen CLIP predictive entropy.

Table 21: Per-dataset accuracy (%) on the natural distribution-shift benchmark under different test-time stream orders. Dataset names follow Table 1.

Stream order	IN	IN-A	IN-V	IN-R	IN-S	Avg.
Class-blocked	71.37	63.70	65.25	79.36	52.00	66.34
Easy-to-Hard	73.08	64.98	66.10	81.66	53.71	67.91
Hard-to-Easy	71.94	63.67	65.14	80.61	52.11	66.70
Random (Ours)	72.91	64.58	65.77	81.50	53.73	67.70

Table 22: Per-dataset accuracy (%) on the fine-grained categorization benchmark under different test-time stream orders. Dataset names follow Table 2.

Stream order	Aircraft	Caltech	Cars	DTD	EuroSAT	Flowers	Food101	Pets	SUN397	UCF101	Avg.
Class-blocked	26.88	92.81	65.70	55.37	55.22	78.64	82.88	90.41	69.85	73.21	69.10
Easy-to-Hard	26.55	94.77	70.00	56.62	60.48	81.00	86.82	92.78	70.84	77.35	71.72
Hard-to-Easy	26.67	92.78	67.78	53.49	55.32	78.54	85.82	90.98	69.21	73.97	69.46
Random (Ours)	27.33	94.71	70.14	56.49	60.39	78.83	87.01	92.79	71.33	75.26	71.43

Table 23: Per-corruption accuracy (%) on ImageNet-C at severity level 5 under different test-time stream orders. Corruption groups follow Table 3.

Stream order	Blur				Weather				Digital				Noise			Avg.
	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brig.	Cont.	Elas.	Pix.	JPEG	Gauss.	Shot	Impu.	
Class-blocked	28.07	20.14	28.84	27.65	38.06	36.80	42.51	59.25	21.58	19.90	38.71	37.48	18.13	19.13	19.10	30.36
Easy-to-Hard	29.11	21.04	30.15	28.87	39.49	37.93	43.94	61.90	22.47	20.64	40.27	38.96	18.49	19.67	19.75	31.51
Hard-to-Easy	27.36	19.31	27.89	26.42	37.68	36.17	41.94	60.19	20.76	18.90	37.85	37.22	17.77	18.64	18.54	29.78
Random (Ours)	28.10	19.75	28.81	27.22	38.39	36.92	42.97	61.18	21.13	19.40	38.97	37.99	17.88	18.90	18.92	30.43

I Ablation of the Label-Free Geometry Estimator

Table 24 isolates the estimation of the second-moment statistic $\mathbf{G}$ while keeping the prototype update and the fusion rule fixed. Unlike the coupled variant, which updates $\mathbf{G}$ only from pseudo-label-confident samples, STRA accumulates $\mathbf{G}$ from all test features before spherical vMF scoring.

Proposition 1 shows that confidence selection changes the population geometry by a covariance term unless confidence is direction-independent; it also reduces the effective sample size. The label-free estimator avoids this confidence-conditioned target and consistently improves over the coupled variant by 0.57%, 0.35%, and 0.61% on Tasks 1–3, respectively.

Table 24: Ablation of the marginal-geometry estimator in STRA with CLIP ViT-B/16.

Geometry statistic	Task 1	Task 2	Task 3
Coupled $\mathbf{G}$	67.13	71.08	29.82
Label-free $\mathbf{G}$ (ours)	**67.70**	**71.43**	**30.43**

J Additional Stress Tests

Severe class imbalance. We construct a controlled closed-set imbalance stream from ImageNet-C Gaussian noise at severity 5. This setting preserves STRA's ImageNet-1000 label space, imposes a strong corruption that lowers zero-shot confidence, and starts from balanced per-class counts so that the imbalance is reproducible. Classes are split into head (20%), medium (30%), and tail (50%) groups. All head samples are retained, while medium and tail classes are capped at $1/\sqrt{50}$ and $1/50$ of the average head-class size, respectively. As shown in Table 25, STRA exceeds the best baseline by 1.35 points overall and by 1.40, 0.77, and 1.20 points on head, medium, and tail classes, respectively.

Table 25: Top-1 accuracy (%) under severe class imbalance on ImageNet-C Gaussian noise, severity 5.

Method	Overall	Head	Medium	Tail
SCA	15.32	15.37	15.19	14.80
ADAPT	15.20	15.14	15.52	15.00
STRA	**16.67**	**16.77**	**16.29**	**16.20**

Confusable classes. We further analyze Stanford Cars because its visually similar categories make frozen CLIP pseudo-labels particularly noisy. We define the top 20% of classes with the highest frozen-CLIP error rates as a confusable subset. This method-independent split is used only for post-hoc stratified evaluation; all methods adapt on the same standard unlabeled stream. Table 26 shows that STRA improves the confusable subset by 6.05 points over SCA while remaining competitive on the other classes. This robustness is consistent with STRA's geometry–semantics decoupling: the marginal geometry is estimated from all unlabeled samples, whereas noisy pseudo-labels only enter the text-anchored prototype update.

Table 26: Top-1 accuracy (%) on the frozen-CLIP-defined confusable subset of Stanford Cars.

Method	Overall	Confusable 20%	Other classes
SCA	68.49	29.07	78.32
ADAPT	68.19	26.90	78.49
STRA	**70.14**	**35.12**	**78.77**

ViT-L/14 scalability and refresh interval. Table 27 evaluates ImageNet-C Gaussian noise at severity 5 using CLIP ViT-L/14. STRA with $R = 10$ reaches 34.12%, outperforming SCA and ADAPT by 2.21 and 1.75 points while using less total time and peak memory close to SCA. Increasing R from 10 to 1000 lowers latency from 24.50 to 22.07 ms/sample with only a 0.02-point accuracy change, confirming that the periodic factorization is effectively amortized in practice.

K Broader Impacts

STRA aims to improve the reliability and efficiency of vision–language recognition under distribution shift by adapting from unlabeled test streams without backpropagation or raw-feature caching. Potential positive impacts include lower computational cost for robust visual recognition, easier deployment

Table 27: ViT-L/14 scalability on ImageNet-C Gaussian noise, severity 5.

Method	Accuracy ↑	Total time ↓	Latency ↓	Peak memory ↓
SCA	31.91	1477.1 s	29.54 ms	**0.910 GB**
ADAPT	32.37	5950.8 s	119.02 ms	1.336 GB
STRA ($R = 10$)	**34.12**	1224.8 s	24.50 ms	0.928 GB
STRA ($R = 100$)	34.10	1196.2 s	23.93 ms	0.928 GB
STRA ($R = 1000$)	34.10	**1103.5 s**	**22.07 ms**	0.928 GB

in resource-constrained settings, and reduced need for labeled target-domain data. At the same time, stronger test-time robustness does not by itself guarantee safe deployment. The method could be used in privacy-sensitive or high-stakes visual decision systems where errors, over-confidence, or distribution-specific failures may have harmful consequences. Because STRA builds on a frozen CLIP model, it may also inherit biases from the pretrained representation and from the prompts, datasets, and class taxonomies used at inference time; online adaptation under class imbalance or non-stationary streams may further amplify such biases. Responsible use should therefore include task-specific validation, calibration checks, privacy and fairness audits, monitoring for failure modes under shift, and human oversight or abstention mechanisms in high-stakes applications. The paper releases an adaptation algorithm and does not introduce new datasets, scraped data, or pretrained foundation models.