DECOUPLING GEOMETRY AND SEMANTICS FOR TEST-TIME ADAPTATION OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) improves model robustness to distribution shifts by adapting to unlabeled test data during inference. For vision-language models (VLMs) such as CLIP, our analysis reveals that distribution shifts alter two structures: the global second-order geometry of visual representations and the class-wise alignment between visual and textual semantics. However, existing VLM-TTA methods either focus primarily on prediction-level adaptation or estimate target statistics through prediction-dependent class assignments, without explicitly separating these two adaptation targets. Thus, erroneous class assignments can jointly distort geometric and semantic estimates under domain shifts. Motivated by this observation, we propose Geometry-Semantics Decoupled Test-Time Adaptation (GeoSem-TTA), a closed-form Gaussian-LDA framework that decouples global geometric adaptation from class-wise semantic adaptation. Specifically, GeoSem-TTA recursively tracks target geometry from reliability-weighted class-independent second-order observations, while adapting class-wise semantic locations through reliability-gated evidence accumulation and coupled centroid synthesis. The resulting statistics directly reconstruct a target-aware Gaussian Linear Discriminant Analysis classifier without parameter updates, backpropagation, or a growing feature bank. Experiments across cross-domain, cross-dataset, and OOD benchmarks demonstrate consistent improvements over strong VLM-TTA baselines, while running 2.3$\times$ faster than DLAE under the same setting.

1 INTRODUCTION

Vision-Language Models (VLMs), such as CLIP (Radford et al., 2021), have exhibited strong zero-shot recognition ability through large-scale image-text contrastive pre-training. However, in real-world deployment, variations in imaging conditions, visual styles, and scene distributions often induce distribution shifts, which disrupt the pretrained visual-semantic alignment and lead to substantial performance degradation. To address this issue, Test-Time Adaptation (TTA) (Wang et al., 2021; Niu et al., 2022) dynamically adjusts the model using only unlabeled test data, providing an effective way to improve its adaptability to unseen distributions. Nevertheless, what should be adapted to effectively accommodate such distribution shifts remains insufficiently understood.

To answer this question, we revisit how CLIP performs zero-shot recognition. CLIP predicts a class by measuring the similarity between visual features and class-wise textual representations. Under distribution shifts, the reliability of these similarity-based predictions is affected by two complementary changes in the representation space: the **global geometry of visual representations** and the **class-wise alignment between visual and textual semantics**. Our statistical analysis reveals that distribution shifts can substantially alter both structures. As shown in Figure 1(a), different target distributions exhibit pronounced variations in their second-order feature statistics, indicating shifts in the global geometry of visual representations. Meanwhile, Figure 1(b) shows clear changes in the similarity between target visual class centers and their corresponding textual representations, revealing shifted class-wise cross-modal alignment. These observations suggest that effective TTA should account for both global geometric adaptation and class-wise semantic adaptation.

However, existing VLM test-time adaptation methods remain limited in addressing these two types of changes. Some optimization-based approaches (Shu et al., 2022; Feng et al., 2023; Xiao et al., 2025; Samadh et al., 2023) update prompts or model parameters through entropy minimization or

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.

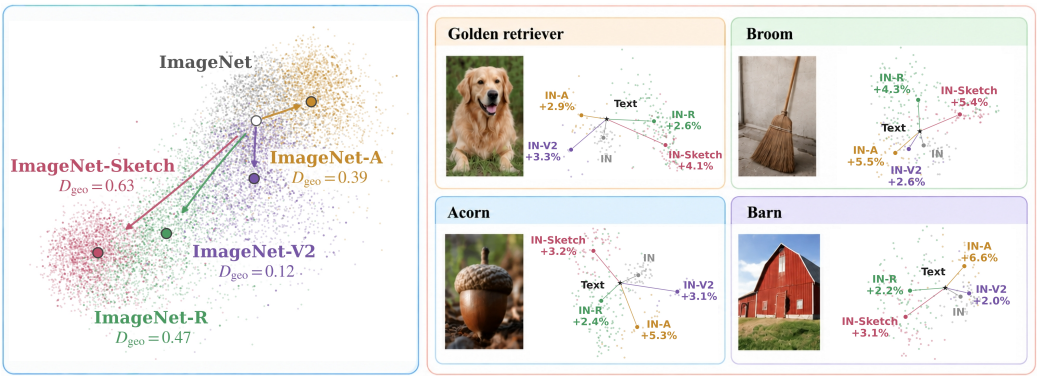


Figure 1: **Distribution shifts alter both global feature geometry and class-wise visual–text alignment. Left:** target domains exhibit distinct second-order geometric shifts from ImageNet. **Right:** representative classes exhibit heterogeneous visual–text alignment shifts across target domains; annotations report the percentage change in visual–text distance relative to ImageNet.

distribution alignment, but do not explicitly characterize the target-induced deformation of the representation geometry underlying CLIP similarity. Memory- or prototype-based approaches (Karmanov et al., 2024; Zhang et al., 2024a; Wu et al., 2026; Chen et al., 2025; Luo et al., 2026) instead construct feature caches or class prototypes from test samples and use them to refine the original CLIP predictions. Their adaptation is largely centered on class-related representations and prediction refinement, without explicitly characterizing changes in the global geometry of target visual representations. Recently, distribution- or statistics-based approaches (Guan et al., 2025; Dai & Yang, 2025; Zhou et al., 2025; Cui et al., 2025; Zhang et al., 2025) further exploit first- and second-order statistics or probabilistic modeling to characterize target distributions. However, when both class centroids and shared covariance are updated using predicted class assignments, one incorrect prediction can corrupt both semantic alignment and global geometry.

To address these limitations, we propose Geometry-Semantics Decoupled Test-Time Adaptation (GeoSem-TTA), which decouples global geometric adaptation from class-wise semantic adaptation. Guided by our representation-level observations, we first estimate the global target geometry directly from streaming visual features. Each test feature contributes a reliability-weighted second-order observation, from which the target precision matrix is recursively updated using the Sherman–Morrison formula. Since the geometric direction is determined by the visual feature itself rather than its predicted class, erroneous class assignments do not directly distort the estimated geometry. In parallel, we adapt class-wise semantics by evolving visual centroids around the pretrained text anchors. Candidate sparsification and cosine-based gating filter unreliable class evidence, while complementary selective and empirical visual statistics are accumulated to construct robust target centroids. The adapted precision matrix and class centroids are then combined through a Gaussian discriminant rule with Mahalanobis bias calibration, yielding a new test-time decision boundary rather than a residual correction to the original CLIP logits. Importantly, GeoSem-TTA performs adaptation entirely through compact recursively updated statistics, requiring neither model-parameter updates nor backpropagation or a growing feature bank, thereby enabling efficient single-pass streaming inference with constant memory. Our main contributions are summarized as follows:

- We reveal that test-time distribution shifts simultaneously alter the global geometry of visual features and the class-wise alignment between visual and textual semantics, identifying them as two complementary adaptation targets for VLM Test-Time Adaptation.
- We propose GeoSem-TTA, which adapts global geometry through recursive precision tracking and class-wise semantics through noise-gated centroid evolution. The two estimates are unified by a closed-form Gaussian discriminant rule with Mahalanobis bias calibration, constructing a standalone adapted decision boundary without model-parameter updates, backpropagation, or a growing feature bank.
- Extensive evaluations across diverse cross-domain, cross-dataset, and OOD benchmarks demonstrate that GeoSem-TTA achieves strong robustness under diverse distribution shifts, while its closed-form and bank-free design enables constant-memory single-pass adaptation with over 1,100 feature-level adaptation steps/s.

2 PRELIMINARIES AND PROBLEM DEFINITION

2.1 STREAMING TEST-TIME ADAPTATION FOR VISION-LANGUAGE MODELS

Let $\mathcal{Y} = \{1, \dots, K\}$ denote the label space. Given a test image X_t, a pre-trained VLM such as CLIP extracts an L2-normalized visual feature $\mathbf{x}_t \in \mathbb{R}^d$. The normalized text representations of all classes form a frozen classifier $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_K] \in \mathbb{R}^{d \times K}$, yielding the zero-shot prediction:

$$p_t^{(k)} = \frac{\exp(\mathbf{x}_t^\top \mathbf{m}_k / \tau)}{\sum_{j=1}^{K} \exp(\mathbf{x}_t^\top \mathbf{m}_j / \tau)}. \tag{1}$$

In practical online deployment, test data often arrive continuously and must be processed immediately under limited latency and storage budgets. Meanwhile, distribution shifts can make the fixed visual–text matching of CLIP unreliable. We therefore consider single-pass streaming TTA, where unlabeled samples arrive sequentially and each X_t must be predicted without source data, ground-truth labels, or future samples. Previously processed samples are not revisited.

2.2 REPRESENTATION SHIFTS AND DECOUPLED ADAPTATION FORMULATION

As shown in Figure 1, distribution shifts affect two complementary structures of CLIP representations. Target visual features exhibit domain-dependent changes in their global second-order geometry, which we characterize by the target second-order moment:

$$\mathbf{G}^\star = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\mathrm{tar}}}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{d \times d}. \tag{2}$$

Meanwhile, class-wise visual centroids shift differently relative to the frozen text anchors. For class k, we denote its target visual centroid as:

$$\boldsymbol{\mu}_k^\star = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\mathrm{tar}}}[\mathbf{x} \mid y = k], \qquad \boldsymbol{\Phi}^\star = [\boldsymbol{\mu}_1^\star, \dots, \boldsymbol{\mu}_K^\star]. \tag{3}$$

whose displacement from the corresponding text anchor m_k reflects the class-wise visual–text semantic shift. Importantly, the severities of these two shifts are not necessarily correlated across domains, indicating that global geometry and class-wise semantics reflect distinct aspects of representation drift. Detailed analyses are provided in Appendix B.

The two structures also differ in how they can be estimated. The second-order observation $\mathbf{x}\mathbf{x}^\top$ is class-independent, whereas estimating $\boldsymbol{\mu}_k^\star$ requires class-related evidence. Under domain shifts, erroneous class assignments may therefore corrupt class-wise semantic statistics; see Appendix C for an analysis of error reinforcement. Critically, reusing the same assignments to estimate global geometry unnecessarily propagates their errors to both structures. We thus estimate global geometry from class-independent observations and class-wise semantics from reliability-filtered class evidence.

Accordingly, we formulate streaming TTA as the online estimation of these two complementary target structures. Since fixed cosine similarity to frozen text anchors cannot explicitly account for either shift, we reconstruct a target-aware classifier directly from the estimated geometric and semantic statistics. This raises the remaining question of how the two structures can be jointly translated into an efficient decision rule for streaming inference.

3 GEOMETRY-SEMANTICS DECOUPLED TEST-TIME ADAPTATION

To jointly incorporate the two target structures while remaining tractable for streaming adaptation, we adopt Gaussian Linear Discriminant Analysis (LDA) (Hastie & Tibshirani, 1996; Anderson, 1958; Friedman, 1989). Its decision rule is parameterized by a shared second-order geometric component and class-specific means, matching the geometric–semantic decomposition identified above. Moreover, its classifier admits a closed-form solution, enabling direct reconstruction from target statistics without parameter optimization.

3.1 GAUSSIAN-LDA CLASSIFIER FROM TARGET STATISTICS

Given a target feature x, classification seeks the class with the highest posterior probability. According to Bayes' rule, this is equivalent to:

$$\hat{y} = \arg\max_k p(y = k \mid \mathbf{x}) = \arg\max_k \left[\log p(\mathbf{x} \mid y = k) + \log P(y = k)\right], \tag{4}$$

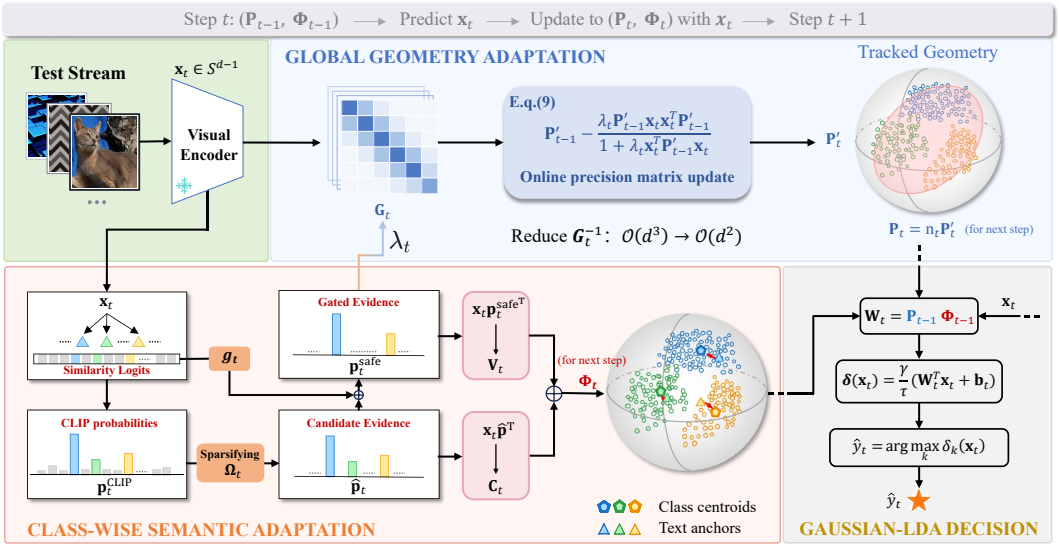


Figure 2: **Overview of GeoSem-TTA.** Global geometry and class-wise semantics are updated from each incoming representation through complementary second-order and reliability-gated semantic statistics. At step t, the historical states $\mathbf{P}_{t-1}$ and $\mathbf{\Phi}_{t-1}$ form the Gaussian-LDA classifier for predicting $\mathbf{x}_t$, after which evidence from $\mathbf{x}_t$ updates the states to $\mathbf{P}_t$ and $\mathbf{\Phi}_t$ for subsequent prediction.

where the marginal $p(x)$ is omitted because it is identical across classes. The remaining problem is therefore to characterize the class-conditional target distribution $p(x \mid y = k)$.

To this end, we adopt Gaussian LDA and model $p(\mathbf{x}|y = k) = \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Sigma})$, where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ denotes the class-specific semantic location, while the shared covariance $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ parameterizes the global geometric component of the decision rule. This parameterization naturally separates the class-specific semantic structure from the shared geometric structure identified in Section 2.

Under the uniform prior $P(y = k) = 1/K$, the prior term in Eq. (4) is also identical across classes. Expanding the Gaussian log-likelihood gives:

$$\begin{aligned}\log p(\mathbf{x}|y=k) &= -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) \\ &= \underbrace{-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}}_{\text{Class-Independent Constants}} + \boldsymbol{\mu}_k^T\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k. \end{aligned} \tag{5}$$

Substituting Eq. (5) into Eq. (4), all class-independent terms can be removed without affecting the class ranking. The remaining terms define the Gaussian-LDA discriminant score:

$$\delta_k(\mathbf{x}) \triangleq \boldsymbol{\mu}_k^\top\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k. \tag{6}$$

such that $\hat{y} = \arg\max_{k\in\{1,\dots,K\}} \delta_k(x)$. To expose the geometric and semantic components of the resulting classifier, we define the precision matrix $\mathbf{P} = \mathbf{\Sigma}^{-1}$ and collect all class means into $\mathbf{\Phi} = [\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K] \in \mathbb{R}^{d\times K}$. The discriminant scores can then be written as a linear classifier:

$$\mathbf{W} = \mathbf{P}\mathbf{\Phi}, \qquad \mathbf{b} = -\frac{1}{2}\operatorname{diag}\left(\mathbf{\Phi}^\top\mathbf{P}\mathbf{\Phi}\right). \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{d\times K}$ and $\mathbf{b} \in \mathbb{R}^K$. Accordingly, $\boldsymbol{\delta}(\mathbf{x}) = \mathbf{W}^\top\mathbf{x} + \mathbf{b}$. The resulting classifier is thus fully determined by the global geometric metric $\mathbf{P}$ and the class-wise semantic locations $\mathbf{\Phi}$.

3.2 Global Geometric Adaptation via Class-Independent Precision Tracking

We first estimate the geometric component from the second-order structure of the incoming target representations. Each observation $\mathbf{x}_t\mathbf{x}_t^\top$ is naturally class-independent, providing a direct way to characterize global geometry.

For target representations with a nonzero mean, the uncentered second-order moment satisfies $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathrm{Cov}(\mathbf{x}) + \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top$, and is therefore not identical to the centered covariance assumed by Gaussian LDA. Rather than estimating that covariance directly, we use the uncentered second-order structure as a tractable, class-assignment-independent surrogate for the global geometric metric $\mathbf{P}$.

Given the target stream $\{\mathbf{x}_t\}$, we accumulate a reliability-weighted second-order statistic as:

$$\mathbf{G}_t = \sum_{\tau=1}^{t} \lambda_\tau \mathbf{x}_\tau \mathbf{x}_\tau^\top = \mathbf{G}_{t-1} + \lambda_t \mathbf{x}_t \mathbf{x}_t^\top, \qquad \mathbf{G}_0 = \mathbf{0}. \tag{8}$$

where $\lambda_t \in [0, 1]$ is the sample-level reliability defined in Section 3.3. Here, $\mathbf{x}_t\mathbf{x}_t^\top$ determines the geometric observation itself, while λ_t controls its contribution based on the reliability of $\mathbf{x}_t$.

The Gaussian-LDA classifier in Section 3.1 operates on an inverse geometric metric. We therefore convert the accumulated second-order statistic into a regularized inverse form and maintain the auxiliary matrix $\mathbf{P}'_t = (\mathbf{G}_t + \gamma\mathbf{I})^{-1}$, where $\gamma > 0$ regularizes the inverse estimation. This auxiliary form is particularly suitable for streaming adaptation because each incoming sample introduces only a rank-one change to $\mathbf{G}_t$. Since each incoming sample induces an exact rank-one increment $\lambda_t \mathbf{x}_t \mathbf{x}_t^\top$ to this surrogate statistic, its regularized inverse admits an exact update through the Sherman–Morrison (Sherman & Morrison, 1950) identity:

$$\mathbf{P}'_t = \mathbf{P}'_{t-1} - \frac{\lambda_t \mathbf{P}'_{t-1}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{P}'_{t-1}}{1 + \lambda_t \mathbf{x}_t^\top \mathbf{P}'_{t-1}\mathbf{x}_t}, \qquad \mathbf{P}'_0 = \gamma^{-1}\mathbf{I}. \tag{9}$$

This recursion updates the inverse directly, reducing each geometric update to $O(d^2)$ computation and avoiding repeated inversion of a $d \times d$ matrix; see Appendix D.2 for the proof.

The auxiliary matrix $\mathbf{P}'_t$ corresponds to the inverse of the accumulated statistic $\mathbf{G}_t$, whose scale increases as target observations are incorporated. The geometric metric should instead reflect the elapsed-step scale of the accumulated state. We therefore define $n_t = \max(1, t)$ and construct the scale-aligned geometric state as

$$\mathbf{P}_t = n_t \mathbf{P}'_t \equiv \left(\frac{1}{n_t}\mathbf{G}_t + \frac{\gamma}{n_t}\mathbf{I}\right)^{-1}, \qquad n_t = \max(1, t). \tag{10}$$

Here, the factor n_t aligns the inverse statistic with the elapsed-step scale of the accumulated state. For prediction at step t, the classifier uses the historical geometric state $\mathbf{P}_{t-1} = n_{t-1}\mathbf{P}'_{t-1}$, where $n_{t-1} = \max(1, t-1)$, thereby preserving strict predict-before-update causality. At initialization, $\mathbf{P}_0 = \mathbf{P}'_0 = \gamma^{-1}\mathbf{I}$, corresponding to an isotropic geometric metric.

3.3 Semantic Adaptation via Reliability-Gated Centroid Evolution

With the global geometric adaptation established, we next update the class-wise semantic state $\mathbf{\Phi}_t$ using evidence from the current representation $\mathbf{x}_t$. Unlike global geometry, estimating class centroids requires associating each incoming representation with plausible semantic classes. Under distribution shifts, such associations can be uncertain, creating a natural trade-off between reliable but selective evidence and broader but potentially noisy evidence. We therefore construct complementary class-wise evidence to balance semantic reliability and target coverage.

Candidate and Reliability-Gated Evidence. For each incoming representation $\mathbf{x}_t$, we first restrict its semantic assignment to a small set of plausible classes. Let $p_t^{\max} = \max_k P_{\mathrm{CLIP}}(y = k \mid \mathbf{x}_t)$. We retain the minimal candidate set Ω_t whose cumulative probability reaches the adaptive threshold $\theta_t = (1 - p_t^{\max})\alpha + p_t^{\max}$, where the candidate-set coefficient α is fixed globally to 0.1. Renormalizing the CLIP scores within Ω_t gives the candidate evidence:

$$\hat{p}_t^{(k)} = \frac{\exp(\mathbf{x}_t^\top \mathbf{m}_k/\tau) \cdot \mathbb{I}(k \in \Omega_t)}{\sum_{\ell \in \Omega_t} \exp(\mathbf{x}_t^\top \mathbf{m}_\ell/\tau)}. \tag{11}$$

This candidate distribution suppresses weakly related classes while retaining multiple plausible assignments, thereby providing broader class-wise coverage than a hard pseudo-label.

Candidate probabilities characterize the relative preference among plausible classes, while reliable semantic adaptation additionally requires sufficient absolute visual–text agreement. We therefore

evaluate each candidate through its similarity to the corresponding frozen text anchor:

$$g_t^{(k)} = \mathbb{I}\left(\mathbf{x}_t^\top \mathbf{m}_k \geq \rho\right), \tag{12}$$

where ρ is the similarity threshold, and $\mathbb{I}(\cdot)$ is the indicator function. Combining the candidate distribution with this gate yields the reliability-filtered evidence:

$$\mathbf{p}_t^{\text{safe}} = \hat{\mathbf{p}}_t \odot \mathbf{g}_t, \qquad \lambda_t = \sum_{k=1}^{K} p_{t,k}^{\text{safe}} \in [0, 1]. \tag{13}$$

Here, $\mathbf{p}_t^{\text{safe}}$ retains class associations with stronger semantic support, while λ_t summarizes sample-level reliability and also weights the geometric update in Section 3.2.

Complementary Evidence Accumulation. The candidate and reliability-filtered distributions provide complementary information. The former offers broader target coverage, whereas the latter emphasizes reliable semantic associations. We accumulate them separately over the target stream:

$$\mathbf{C}_t = \mathbf{C}_{t-1} + \mathbf{x}_t \hat{\mathbf{p}}_t^\top, \qquad \mathbf{V}_t = \mathbf{V}_{t-1} + \mathbf{x}_t (\mathbf{p}_t^{\text{safe}})^\top, \tag{14}$$

where $\mathbf{C}_t = [\mathbf{c}_{1,t}, \ldots, \mathbf{c}_{K,t}]$ and $\mathbf{V}_t = [\mathbf{v}_{1,t}, \ldots, \mathbf{v}_{K,t}]$, with $\mathbf{C}_0 = \mathbf{V}_0 = \mathbf{0}$. Accordingly, $\mathbf{C}_t$ captures broader empirical class-wise directions from the target stream, while $\mathbf{V}_t$ retains more selectively supported semantic directions.

Coupled Centroid Synthesis. We first use the selective evidence in $\mathbf{V}_t$ to adapt the frozen text anchor toward reliably supported target regions. For each class k, let $\tilde{\mathbf{v}}_{k,t} = \frac{\mathbf{v}_{k,t}}{\|\mathbf{v}_{k,t}\|_2}$ when $\|\mathbf{v}_{k,t}\|_2 > 0$. The adapted semantic anchor is defined as:

$$\mathbf{m}_{k,t}^{\text{adapt}} = (1 - \eta)\mathbf{m}_k + \eta \tilde{\mathbf{v}}_{k,t}. \tag{15}$$

where $\mathbf{m}_{k,t}^{\text{adapt}} = \mathbf{m}_k$ when no reliability-filtered evidence has been accumulated for class k. This preserves the pre-trained semantic prior while incorporating reliable target-specific displacement.

Reliable evidence provides a stable adaptation direction, but its selective nature can limit target coverage, particularly for difficult classes or at early streaming stages. We therefore complement the adapted anchor with the broader empirical direction accumulated in $\mathbf{C}_t$. Let $\tilde{\mathbf{c}}_{k,t} = \frac{\mathbf{c}_{k,t}}{\|\mathbf{c}_{k,t}\|_2}$ when $\|\mathbf{c}_{k,t}\|_2 > 0$. The updated target semantic location is then synthesized as:

$$\boldsymbol{\phi}_{k,t} = (1 - \omega)\mathbf{m}_{k,t}^{\text{adapt}} + \omega \tilde{\mathbf{c}}_{k,t}, \qquad \boldsymbol{\Phi}_t = [\boldsymbol{\phi}_{1,t}, \ldots, \boldsymbol{\phi}_{K,t}]. \tag{16}$$

For classes with no accumulated candidate evidence, we set $\boldsymbol{\phi}_{k,t} = \mathbf{m}_{k,t}^{\text{adapt}}$.

3.4 Streaming Inference with the Gaussian-LDA Discriminant

At step t, the historical states $\mathbf{P}_{t-1}$ and $\boldsymbol{\Phi}_{t-1}$ summarize previous observations and are therefore jointly available before predicting $\mathbf{x}_t$. Classification weights $\mathbf{W}_t$ and quadratic Mahalanobis bias $\mathbf{b}_t$ evaluate as:

$$\mathbf{W}_t = \mathbf{P}_{t-1}\boldsymbol{\Phi}_{t-1}, \qquad \mathbf{b}_t = -\frac{1}{2}\operatorname{diag}\left(\boldsymbol{\Phi}_{t-1}^\top \mathbf{W}_t\right). \tag{17}$$

The final classification logit vector and label prediction are evaluated in closed form:

$$\boldsymbol{\delta}(\mathbf{x}_t) = \frac{\gamma}{\tau}\left(\mathbf{W}_t^\top \mathbf{x}_t + \mathbf{b}_t\right), \qquad \hat{y}_t = \arg\max_k \delta_k(\mathbf{x}_t), \tag{18}$$

where τ is the CLIP temperature and γ is the ridge parameter for geometric adaptation. The positive factor γ/τ preserves class ranking while aligning the initial discriminant scale with zero-shot CLIP. At initialization, $\mathbf{P}_0 = \gamma^{-1}\mathbf{I}$ and $\boldsymbol{\Phi}_0 = \mathbf{M}$, recovering the zero-shot CLIP ranking up to a class-independent offset. The historical precision $\mathbf{P}_{t-1}$ and coupled centroids $\boldsymbol{\Phi}_{t-1}$ then define the Gaussian-LDA discriminant used to predict $\mathbf{x}_t$. Detailed analyses are provided in Appendix D.

4 Experiments

We evaluate GeoSem-TTA under single-pass streaming TTA with CLIP ViT-B/16 and ResNet-50. Experiments assess its accuracy–efficiency trade-off against existing VLM-TTA methods, generalization across challenging target streams, and contributions of individual components. Detailed experimental setups and robustness analyses are provided in Appendices E.5, E.7 and E.8.

Table 1: Top-1 classification accuracy (%) on ImageNet and four OOD benchmarks under streaming TTA on CLIP ViT-B/16 (full ResNet-50 results are detailed in Appendix Table 9).

Method	ImageNet	-A	-V2	-R	-Sketch	OOD Avg.	Total Avg.
CLIP-ViT-B/16 (Radford et al., 2021)	66.73	47.87	60.86	73.98	46.09	57.20	59.11
CoOp (16-shot) (Zhou et al., 2022b)	71.51	49.71	64.20	75.21	47.99	59.28	61.72
TPT (Shu et al., 2022)	68.98	54.77	63.45	77.06	47.94	60.81	62.44
DiffTPT (Feng et al., 2023)	70.30	55.68	65.10	75.00	46.80	60.52	62.28
TDA (Karmanov et al., 2024)	69.51	60.11	64.67	80.24	50.54	63.89	65.01
DPE (Zhang et al., 2024a)	71.91	59.63	65.44	80.40	52.26	64.43	65.93
DLAE (Wu et al., 2026)	**72.41**	64.06	65.50	80.43	53.06	65.76	67.09
SCA (Guan et al., 2025)	71.75	60.33	65.38	80.85	52.50	64.77	66.16
BCA (Zhou et al., 2025)	70.22	61.14	64.90	80.72	50.87	64.16	65.37
ADAPT (Zhang et al., 2025)	70.91	63.32	64.64	80.66	**53.13**	65.44	66.53
GeoSem-TTA (Ours)	72.18	**64.51**	**65.52**	**81.20**	52.55	**65.94**	**67.19**

Table 2: Accuracy–efficiency comparison under the streaming TTA protocol on ImageNet.

Method	BP-Free	Throughput	Latency	Peak Memory	Total Avg.
CLIP (Radford et al., 2021)	✓	35,729.6 steps/s	0.03 ms	1020.5 MB	59.11
TPT (Shu et al., 2022)	×	1.7 steps/s	586.54 ms	24,102.1 MB	62.44
DiffTPT (Feng et al., 2023)	×	0.4 steps/s	2364.30 ms	24,102.9 MB	62.28
TDA (Karmanov et al., 2024)	✓	69.2 steps/s	14.45 ms	1192.1 MB	65.01
DPE (Zhang et al., 2024a)	×	499.8 steps/s	2.00 ms	1222.9 MB	65.93
DLAE (Wu et al., 2026)	×	502.8 steps/s	1.99 ms	1196.4 MB	67.09
SCA (Guan et al., 2025)	✓	665.5 steps/s	1.50 ms	1187.0 MB	66.16
GeoSem-TTA (Ours)	✓	**1165.0 steps/s**	**0.86 ms**	**1096.0 MB**	**67.19**

4.1 Quantitative Comparison with Existing Methods

Out-of-Distribution Performance. As shown in Table 1, GeoSem-TTA achieves the best OOD average of 65.94% and total average of 67.19%, ranking first on ImageNet-A, ImageNet-V2, and ImageNet-R. As illustrated in Figure 1, these OOD domains induce changes in both the global geometry of visual representations and the class-wise visual–text alignment. By explicitly adapting these two complementary structures, GeoSem-TTA maintains strong performance across diverse distribution shifts, supporting the effectiveness of the proposed geometry–semantic decoupling.

Computational Efficiency. As shown in Table 2, GeoSem-TTA achieves the lowest adaptation latency of **0.86 ms** and the highest throughput of 1,165.0 steps/s, running **2.3**× faster than DLAE with only 1,096.0 MB peak memory. This efficiency stems from its closed-form adaptation: global geometry is updated through recursive rank-one precision tracking, while class-wise semantics are maintained using compact streaming statistics, eliminating iterative optimization and feature-bank retrieval. These results demonstrate strong adaptation with lightweight online computation; detailed profiling and protocol analysis are provided in Appendix E.1.

4.2 Generalization and Streaming Robustness

Cross-Domain Generalization. As shown in Table 3, GeoSem-TTA achieves the highest mean accuracy on all four domain-generalization benchmarks, yielding an overall average of 79.81%, 2.20 points above the strongest reproduced baseline SCA. The consistent gains across OfficeHome, PACS, VLCS, and TerraIncognita support the proposed decoupling across heterogeneous visual domains, where global geometric and class-wise semantic shifts need not follow a common pattern.

Cross-Dataset Generalization. Table 4 further evaluates GeoSem-TTA on ten Cross-Dataset benchmarks spanning substantially different label spaces and visual tasks. GeoSem-TTA achieves a competitive 71.11% average, with particularly strong performance on Flowers102 and UCF101. These results extend beyond domain shifts within established domain-generalization benchmarks, supporting the transferability of the proposed adaptation across distinct recognition datasets.

Table 3: **Cross-domain generalization with CLIP ViT-B/16.** Top-1 accuracy (%) across four benchmarks; full domain-level results are provided in Appendix Table 15.

Method	OfficeHome	PACS	VLCS	TerraInc.	Mean
CLIP (Radford et al., 2021)	82.35	96.12	82.34	34.17	73.75
TPT (Shu et al., 2022)	77.20	93.23	74.57	32.45	69.36
CLIPArTT (Hakim et al., 2025)	77.56	93.95	80.06	34.28	71.46
TDA (Karmanov et al., 2024)	85.26	96.91	81.31	46.66	77.54
DPE (Zhang et al., 2024a)	85.45	96.91	82.22	44.94	77.38
SCA (Guan et al., 2025)	85.74	96.79	81.32	46.59	77.61
ADAPT (Zhang et al., 2025)	85.69	96.78	79.43	38.66	72.64
DLAE (Wu et al., 2026)	85.17	96.74	78.35	29.90	70.79
GeoSem-TTA (Ours)	**86.47**	**96.93**	**87.52**	**48.30**	**79.81**

Table 4: **Cross-dataset generalization under single-pass streaming TTA on CLIP ViT-B/16.** Top-1 classification accuracy (%) is reported across ten heterogeneous recognition datasets.

Method	Cal.	DTD	Cars	Euro.	Air.	Flow.	Pets	UCF	Food	SUN	Avg.
CLIP-ViT-B/16 (Radford et al., 2021)	93.35	44.27	65.48	42.01	23.67	67.44	88.25	65.13	83.65	62.59	63.58
CoOp (16-shot) (Zhou et al., 2022b)	93.70	41.92	64.51	46.39	18.47	68.71	89.14	66.55	85.30	64.15	63.88
TPT (Shu et al., 2022)	94.16	47.75	66.87	42.44	24.78	68.98	87.79	68.04	84.67	65.50	65.10
DiffTPT (Feng et al., 2023)	92.49	47.00	67.01	43.13	25.60	70.10	88.22	62.67	**87.23**	65.74	65.47
TDA (Karmanov et al., 2024)	94.24	47.40	67.28	58.00	23.91	71.42	88.63	70.66	86.14	67.62	67.53
DPE (Zhang et al., 2024a)	94.81	54.20	67.31	55.79	28.95	75.07	91.14	70.44	86.17	70.07	69.40
DLAE (Wu et al., 2026)	**95.33**	**58.86**	**69.99**	61.30	**32.58**	77.50	91.88	73.46	87.05	**70.83**	**71.88**
SCA (Guan et al., 2025)	94.85	57.09	68.49	57.16	28.50	76.09	91.44	73.43	86.09	70.27	70.34
BCA (Zhou et al., 2025)	94.69	53.49	66.86	56.63	28.59	73.12	90.43	67.59	85.97	68.41	68.59
ADAPT (Zhang et al., 2025)	94.48	55.20	68.19	**68.19**	28.95	75.56	**92.01**	70.66	83.81	70.57	70.76
GeoSem-TTA (Ours)	95.25	56.15	67.42	59.75	29.43	**79.62**	91.96	**74.97**	86.68	69.87	71.11

Imbalanced-Stream Robustness We further evaluate GeoSem-TTA under long-tailed test streams following the L-TTA construction (Wang et al., 2026), with imbalance ratios $\kappa \in \{10, 20, 50\}$ and five random seeds. As shown in Table 5, GeoSem-TTA achieves the best average Accuracy and Macro-F1 at all three imbalance levels; its Accuracy remains stable as imbalance increases, whereas Macro-F1 exhibits a relatively larger decline under severe imbalance. This indicates that the decoupled streaming statistics preserve overall recognition under skewed class frequencies without tying global geometric observations to class-specific assignments, although performance across classes is more affected as imbalance grows. Detailed dataset-level results are provided in Appendix Table 11.

Table 5: **ImageNet-family performance under class-imbalanced streams** Average Accuracy and Macro-F1 across five benchmarks.

	$\kappa = 10$		$\kappa = 20$		$\kappa = 50$	
Method	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
CLIP	56.93	53.08	56.69	51.87	55.37	47.84
TPT	60.44	56.40	60.62	55.85	59.49	51.99
TDA	64.82	60.50	64.25	58.46	63.51	56.24
DPE	65.63	60.96	64.85	59.76	64.58	56.81
SCA	66.09	61.48	65.94	59.65	66.07	57.10
DLAE	65.58	61.06	65.67	59.53	66.12	57.46
GeoSem (Ours)	**66.54**	**61.75**	**66.54**	**60.06**	**66.86**	**57.68**

Long-Stream Stability. Figure 3 further evaluates prolonged adaptation over the ten Cross-Dataset benchmarks in Table 4, retaining dataset-specific states across ten consecutive rounds. GeoSem-TTA retains 70.88% macro accuracy at T10 while sustaining 1,188 adaptation steps/s, compared with 68.69% and 647 steps/s for DLAE. The stable accuracy and throughput indicate that the decoupled geometric and semantic statistics remain effective as adaptation history accumulates, supporting sustained streaming adaptation through fixed-size states without an expanding memory.

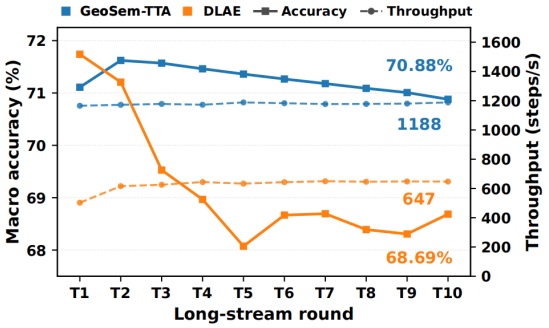


Figure 3: **Long-stream stability and efficiency.** Macro accuracy and throughput across ten rounds.

Table 6: **Component-wise ablation on the Cross-Dataset suite.** Average Top-1 accuracy (%) across ten datasets on ViT-B/16; Δ denotes absolute gain over global geometry adaptation.

Variant	Bias $\mathbf{b}_t$	Cent. $\mathbf{\Phi}_t$	Zero $\mathbf{V}_0$	Gate λ_t	Avg.	Δ
1. Global Geometry Adaptation	–	–	–	–	68.66	–
2. + Bias ($\mathbf{b}_t$)	✓	–	–	–	70.38	+1.72
3. + Empirical Mean ($\mathbf{C}_t$)	✓	✓	–	–	70.67	+2.01
4. + Zero-Init ($\mathbf{V}_0 = \mathbf{0}$)	✓	✓	✓	–	71.03	+2.37
5. Complete GeoSem-TTA	✓	✓	✓	✓	**71.11**	**+2.45**

Table 7: **Cross-Dataset accuracy (%) with CLIP ResNet-50.** Best and second-best results are shown in **bold** and underline, respectively.

Method	Cal.	DTD	Cars	Euro.	Air.	Flow.	Pets	UCF	Food	SUN	Avg.
CLIP	85.88	40.37	55.70	23.69	15.66	61.75	83.57	58.84	73.97	58.80	55.82
DPE	90.83	50.18	59.26	41.67	19.80	67.60	85.97	61.98	77.83	64.23	61.93
DLAE	90.91	**55.56**	**61.98**	50.04	23.34	69.63	**87.68**	64.76	**78.76**	**65.43**	64.81
SCA	**91.03**	53.31	59.73	43.33	20.82	71.09	86.15	66.40	76.51	64.50	63.29
GeoSem-TTA (Ours)	90.91	53.01	58.46	**63.94**	21.78	**72.07**	86.08	**66.77**	77.68	64.91	**65.56**

4.3 Ablation Studies

Effect of Each Component. As shown in Table 6, global geometry adaptation alone achieves a competitive 68.66% Cross-Dataset average, slightly above BCA (68.59%). Introducing the Gaussian-LDA quadratic bias provides the largest additional gain, raising the average to 70.38%. Incorporating the empirical evidence $\mathbf{C}_t$, zero-initializing the selective visual state $\mathbf{V}_0$, and applying reliability modulation further improve the accuracy to 70.67%, 71.03%, and 71.11%, respectively. The consistent gains support the complementary roles of geometric calibration, broader empirical coverage, selective visual accumulation, and reliability weighting in the complete GeoSem-TTA formulation. The larger gain from the quadratic bias further highlights the importance of calibrating the shared decision geometry, while the subsequent gains reflect complementary refinement of class-wise semantic evidence. Detailed dataset-level results are provided in Appendix Table 8.

Impact of Different Visual Backbone. As shown in Table 7, GeoSem-TTA achieves the highest average accuracy of 65.56% across the ten Cross-Dataset benchmarks with CLIP ResNet-50. The improvement is particularly pronounced on EuroSAT, where GeoSem-TTA reaches 63.94%, outperforming DLAE and SCA by 13.90 and 20.61 points, respectively. Notably, this behavior persists despite the change in feature dimensionality and representation structure between the two encoders. The consistent performance after replacing ViT-B/16 with ResNet-50 indicates that the proposed geometry–semantic adaptation transfers across different visual representations rather than relying on the architectural properties of a specific backbone family.

5 Conclusion

In this paper, we revisit test-time adaptation for vision-language models from the perspective of representation shift. Our analysis shows that distribution shifts affect both the global geometry of visual representations and the class-wise alignment between visual and textual semantics, motivating their treatment as complementary adaptation targets. Based on this observation, we propose Geometry-Semantics Decoupled Test-Time Adaptation (GeoSem-TTA), a closed-form Gaussian-LDA framework that decouples global geometric adaptation from class-wise semantic adaptation through recursive precision tracking and reliability-gated centroid evolution. The resulting target statistics directly reconstruct the test-time decision boundary using fixed-size streaming states, without iterative parameter optimization or a growing feature bank. Experiments across OOD, cross-domain, cross-dataset, and challenging streaming settings demonstrate robust adaptation across diverse distribution shifts, while GeoSem-TTA achieves 1,165 adaptation steps/s and runs 2.3$\times$ faster than DLAE under the same protocol. The shared-geometric formulation remains less effective for diffuse texture-dominated categories, suggesting multimodal geometric modeling as a direction for future work. Overall, our results highlight geometry–semantic decoupling as an effective and efficient principle for adapting vision-language models under distribution shift.

AI USE STATEMENT

In this work, generative AI assistance was utilized solely for LaTeX typesetting formatting and minor sentence-level grammatical polishing. All research formulations, mathematical derivations, algorithmic logic, experimental pipelines, and empirical evaluations were conceived, conducted, and verified independently by the authors. The authors accept full responsibility for the final content and scientific integrity of this work.

ETHICS STATEMENT

All datasets used are public academic benchmarks and involve no human subjects or private data. Our method aims to improve the reliability of vision-language models under distribution shift, and we are not aware of additional ethical risks. As with all test-time adaptation methods, deployment in safety-critical settings should be accompanied by independent monitoring and fallback mechanisms.

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility of our work, the theoretical derivations, mathematical proofs, and supporting analyses are detailed in Section 2, Section 3, Appendix C, and Appendix D. The complete source code, PyTorch implementation, dataset configuration files, and evaluation scripts for the reported experiments will be open-sourced upon paper acceptance.

REFERENCES

Theodore Wilbur Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 1958.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *European Conference on Computer Vision*, volume 8694 of *Lecture Notes in Computer Science*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 295–305, 2022.

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

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

Shuang Cui, Jinglin Xu, Yi Li, Xiongxin Tang, Jiangmeng Li, Jiahuan Zhou, Fanjiang Xu, Fuchun Sun, and Hui Xiong. Bayestta: Continual-temporal test-time adaptation for vision-language models via gaussian discriminant analysis. *arXiv preprint arXiv:2507.08607*, 2025. doi: 10.48550/arXiv.2507.08607.

Qiyuan Dai and Sibei Yang. Free on the fly: Enhancing flexibility in test-time adaptation with online em. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9538–9548, 2025.

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

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 129062–129093, 2024. doi: 10.52202/079017-4099.

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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178. IEEE Computer Society, 2004. doi: 10.1109/CVPR.2004.383.

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2704–2714, 2023.

Jerome H Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, 132(2):581–595, 2024.

Zenghao Guan, Yucan Zhou, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Gustavo Adolfo Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. CLIPArTT: Adaptation of CLIP to new domains at test time. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7092–7101. IEEE, 2025. doi: 10.1109/WACV61041.2025.00689.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. DOTA: Distributional test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 142772–142798, 2025.

Trevor Hastie and Robert Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):155–176, 1996.

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, 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, 2021b.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 2427–2440, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Ye-Hsin Ku, Jason Wang, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171, 2024.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, 2023.

Ansh Khurana, Sujoy Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. SITA: Single image test-time adaptation. In *NeurIPS Workshop on Distribution Shifts*, 2021.

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

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 6028–6039, 2020.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. TTN: A domain-shift aware batch normalization in test-time adaptation. In *International Conference on Learning Representations (ICLR)*, 2023.

Wei Luo, Yangfan Ou, Jin Deng, Zeshuai Deng, Xiquan Yan, Zhiquan Wen, and Mingkui Tan. ProtoDCS: Towards robust and efficient open-set test-time adaptation for vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2026. doi: 10.1109/TCSVT.2026.3717446.

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

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

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 (ICML)*, pp. 16888–16905, 2022.

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 (ICLR)*, 2023.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

Alec Radford, Jong Wook Kim, Aditya 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 (ICML)*, pp. 8748–8763, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, pp. 5389–5400, 2019.

Jameel Abdul Samadh, Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Shahbaz Khan, and Salman H. Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

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.

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

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.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10506–10518, 2019.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7201–7211, 2022.

Xucong Wang, Zhe Zhao, Zekun Wang, Xiaofeng Cao, Xu Wang, Di Wu, Pengkun Wang, and Yang Wang. Long-tailed test-time adaptation for vision-language models. In *International Conference on Learning Representations (ICLR)*, pp. 46492–46530, 2026.

Haoyan Wu, Yahao Liu, Yinjie Lei, Lixin Duan, and Wen Li. Dynamic logits adjustment and exploration for test-time adaptation in vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3143–3152, 2026.

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

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees G Snoek. DynaPrompt: Dynamic test-time prompt tuning. In *International Conference on Learning Representations (ICLR)*, 2025.

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

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12924–12933, 2024b. doi: 10.1109/CVPR52733.2024.01228.

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, 2022a.

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 (ECCV)*, pp. 493–510, 2022b. doi: 10.1007/978-3-031-19833-5_29.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. BoostAdapter: Improving vision-language test-time adaptation via regional bootstrapping. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 15430–15452, 2024c.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28718–28728, 2024d.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 (IJCV)*, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29999–30009, 2025.

APPENDIX

CONTENTS

A Related Work **15**

A.1 Prompt-Based Test-Time Adaptation 15

A.2 Memory- and Prototype-Based Test-Time Adaptation 15

A.3 Distribution- and Statistics-Based Test-Time Adaptation 15

B Analysis of Representation Shifts **16**

B.1 Robustness Across Visual Backbones 16

B.2 Complementary Geometry and Semantic Shifts 17

B.3 Shift-Sensitive Projection Diagnostics 17

C Entropy-Gradient Sensitivity under Erroneous Predictions **18**

C.1 Exact Log-Ratio Derivation of the Entropy Gradient 18

C.2 Local Error Reinforcement in Streaming Adaptation 19

D Detailed Mathematical Derivations for GeoSem-TTA **19**

D.1 Gaussian LDA Log-Posterior Expansion and Mahalanobis Bias 20

D.2 Sherman–Morrison Rank-1 Update for Uncentered Second-Order Statistics 20

D.3 Time-Step Scale Alignment and Boundary Continuity 21

D.4 Class-Assignment Independence of the Second-Order Statistic 21

D.5 Complete End-to-End Streaming Algorithm 22

E Extended Experiments and Implementation Details **22**

E.1 Detailed Computational Efficiency and Hardware Overhead Profiling 22

E.2 Complete 10-Dataset Breakdown for Component-Wise Ablations 23

E.3 Boundary Analysis on Diffuse Non-Object Texture Manifolds (DTD) 23

E.4 Complete ResNet-50 Benchmark Evaluations 23

E.5 Corruption Robustness Evaluation 24

E.6 Robustness to Class-Imbalanced Test Streams 24

E.7 Hyperparameter Sensitivity and Stream-Order Robustness 26

E.8 Benchmark and Evaluation Setup 27

E.9 Detailed Cross-Domain Evaluation 27

A Related Work

A.1 Prompt-Based Test-Time Adaptation

Parameter-efficient fine-tuning (PEFT) originally adapted VLMs like CLIP (Radford et al., 2021; Jia et al., 2021) via prompt learning on labeled datasets (Zhou et al., 2022b;a; Khattak et al., 2023; Zhang et al., 2022b; Gao et al., 2024). Inspired by unsupervised test-time objectives (Wang et al., 2021; Liang et al., 2020; Zhang et al., 2022a; Niu et al., 2022; 2023; Wang et al., 2022), TPT (Shu et al., 2022) optimizes continuous prompts by minimizing predictive entropy across augmented views, while DiffTPT (Feng et al., 2023) enriches view diversity via diffusion models. CLIPArTT (Hakim et al., 2025) further adapts CLIP to new target domains at test time by adjusting its representation space without requiring labeled target data. Subsequent variants maintain online prompt buffers (Xiao et al., 2025), align distribution statistics (Samadh et al., 2023), or explore distribution-guided shifts (Zhang et al., 2024b). However, these methods primarily adapt prompts or model parameters through prediction-based objectives, without explicitly modeling the global geometric changes of target visual representations. Moreover, their reliance on gradient-based optimization increases the computational cost of streaming adaptation, while erroneous predictions can reinforce subsequent updates under unfavorable local gradient alignment. These limitations motivate adaptation mechanisms that can characterize target representation shifts without per-sample parameter optimization.

A.2 Memory- and Prototype-Based Test-Time Adaptation

To bypass iterative optimization, feature-level TTA methods adjust representations directly during inference (Iwasawa & Matsuo, 2021; Chen et al., 2022; Khurana et al., 2021; Lim et al., 2023). For VLMs, cache-based TTA maintains memory banks of past visual features or prototype residuals. TDA (Karmanov et al., 2024) establishes dynamic positive and negative key-value caches, DPE (Zhang et al., 2024a) evolves dual visual-textual prototypes, while MCP (Chen et al., 2025) leverages multi-cache structures with negative sample reflection. Concurrently, DLAE (Wu et al., 2026) combines dynamic logit adjustment with an exploratory cache, and ProtoDCS (Luo et al., 2026) explores Gaussian Mixture Model (GMM) separation for open-set updates, complemented by non-parametric memory structures (Zhang et al., 2024c;d; Farina et al., 2024). Together, these methods show that target-side memories and class-related representations provide an effective alternative to repeated prompt optimization. Their adaptive states, however, are mainly organized around stored features, class prototypes, or prediction refinement, rather than an explicit separation between global target geometry and class-wise semantic adaptation. This distinction becomes important when distribution shifts affect both global geometry and class-wise visual–text alignment.

A.3 Distribution- and Statistics-Based Test-Time Adaptation

Beyond feature and prototype memories, recent VLM-TTA methods increasingly model the target distribution through accumulated statistics or probabilistic estimation. SCA (Guan et al., 2025) caches first- and second-order target statistics and constructs an adapted classifier through ridge regression. Probabilistic approaches further characterize class-conditional target distributions: BCA (Zhou et al., 2025) performs Bayesian adaptation of class statistics, while BayesTTA (Cui et al., 2025) uses Gaussian discriminant modeling for continual-temporal adaptation. FreeTTA (Dai & Yang, 2025) and DOTA (Han et al., 2025) further estimate target distributions through online probabilistic adaptation. More recently, ADAPT (Zhang et al., 2025) performs backpropagation-free test-time adaptation through probabilistic Gaussian alignment. These methods mark an important shift from prediction-level correction toward explicit modeling of target distributions.

However, while these approaches explicitly accumulate target statistics, they generally do not explicitly separate global geometric adaptation from class-wise semantic adaptation according to their different estimation requirements, with target statistics often estimated using zero-shot prediction posteriors or pseudo-label responsibilities. Under severe distribution shifts, erroneous initial predictions can therefore affect both geometric and semantic estimates when the same class-related evidence is used for both. Furthermore, existing methods may rely on historical knowledge banks (Zhang et al., 2025) or retain static text anchors during adaptation (Dai & Yang, 2025). In contrast, GeoSem-TTA decouples global geometric adaptation from class-wise semantic adaptation: preci-

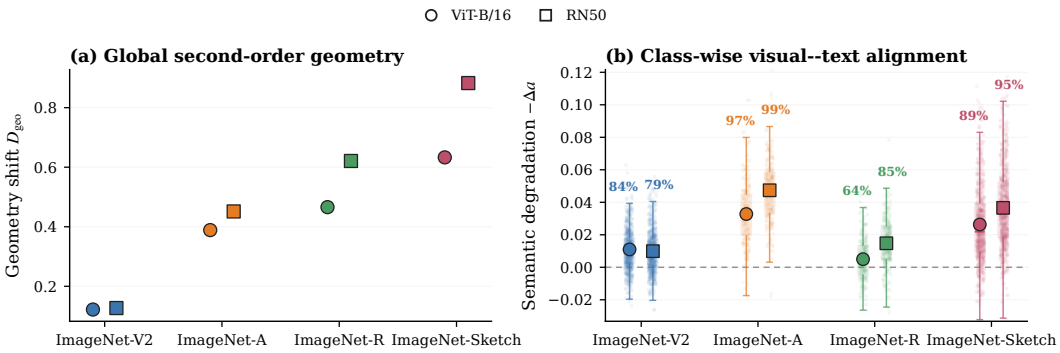


Figure 4: **Robustness of the representation-level shifts.** **(a)** Full-data second-order discrepancy D_{geo} under ViT-B/16 and RN50. **(b)** Paired class-wise semantic degradation $-\Delta a$ relative to ImageNet; percentages denote the fraction of shared classes with decreased alignment. Both effects persist across backbones.

sion updates accumulate class-independent second-order observations with scalar reliability weighting, while noise-gated centroid evolution adapts class-wise semantics around text priors. This separation distinguishes GeoSem-TTA from prior prompt-, memory-, and distribution-based adaptation: rather than adapting a single prediction mechanism or jointly estimated target state, it treats global geometry and class-wise semantics as complementary adaptation targets with different sources of evidence.

B Analysis of Representation Shifts

Figure 1 motivates GeoSem-TTA from two representation-level observations: target shifts alter both global second-order feature geometry and class-wise visual–text alignment. Here we further examine the robustness, complementarity, and low-dimensional structure of these effects. Unless otherwise stated, all statistics are computed from the same frozen feature caches used in the main analysis; ground-truth labels are used only for this offline diagnostic and never by GeoSem-TTA adaptation.

B.1 Robustness Across Visual Backbones

We first test whether the two representation-level effects in Figure 4 persist across different visual backbones. Figure 4(a) reports the full-data second-order discrepancy

$$D_{\text{geo}}(D) = \frac{\|\mathbf{S}_D - \mathbf{S}_{\text{IN}}\|_F}{\|\mathbf{S}_{\text{IN}}\|_F}, \qquad \mathbf{S}_D = \frac{1}{N_D}\sum_i \mathbf{x}_i \mathbf{x}_i^\top,$$

for both ViT-B/16 and RN50. Across both backbones, the domain ordering is preserved: ImageNet-V2 exhibits the weakest geometric shift, followed by ImageNet-A and ImageNet-R, while ImageNet-Sketch exhibits the strongest. Thus, the observed domain-dependent geometric differences persist across the two visual backbones.

Figure 4(b) evaluates the complementary class-wise semantic effect. For each class k shared with ImageNet, we compute

$$a_{k,D} = \cos\left(\boldsymbol{\mu}_{k,D}^{\text{GT}}, \mathbf{m}_k\right), \qquad \Delta a_{k,\mathcal{D}} = a_{k,\mathcal{D}} - a_{k,\text{IN}},$$

where $\mu_{k,\mathcal{D}}^{\text{GT}}$ is the ground-truth visual class center and $\mathbf{m}_k$ is the corresponding frozen CLIP text anchor. The mean paired semantic degradation $-\Delta a$ is positive for every target domain under both ViT-B/16 and RN50, with all bootstrap confidence intervals above zero. For ViT-B/16, 83.7%, 97.0%, 64.5%, and 89.3% of the shared classes decrease in alignment on ImageNet-V2, ImageNet-A, ImageNet-R, and ImageNet-Sketch, respectively; the corresponding RN50 fractions are 78.7%, 99.0%, 85.0%, and 95.0%. Thus, both representation-level effects persist across the two visual backbones.

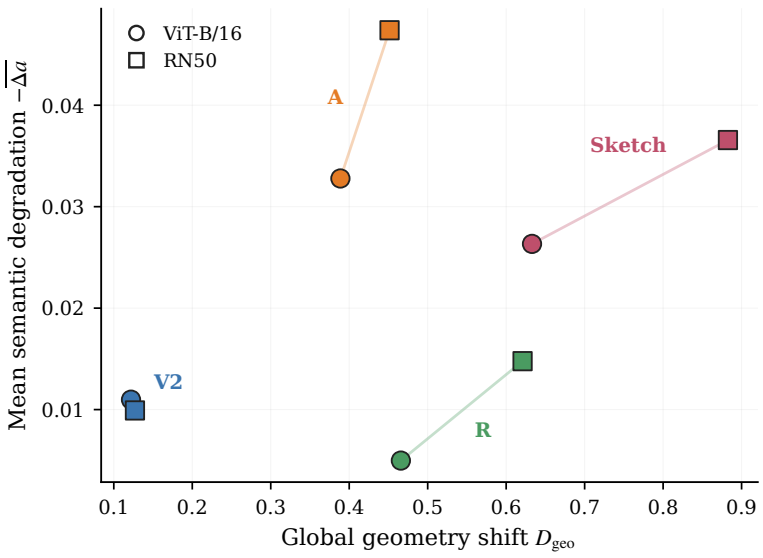


Figure 5: **Global geometric and class-wise semantic shifts exhibit distinct domain-dependent severities.** Each point reports D_{geo} and the mean paired semantic degradation $-\overline{\Delta a}$; circles and squares denote ViT-B/16 and RN50, respectively, and lines connect the same target domain across backbones. ImageNet-R exhibits a larger geometric shift than ImageNet-A under both backbones, whereas ImageNet-A exhibits substantially stronger semantic degradation.

B.2 Complementary Geometry and Semantic Shifts

The persistence of both effects does not imply that they measure the same notion of shift severity. Figure 5 directly compares the global second-order discrepancy D_{geo} with the mean paired semantic degradation $-\overline{\Delta a}$ for each target domain and backbone. The two diagnostics exhibit markedly different domain-dependent severity patterns.

A particularly informative comparison is ImageNet-A versus ImageNet-R. Under ViT-B/16,

$$D_{\text{geo}}(\text{R}) = 0.466 > D_{\text{geo}}(\text{A}) = 0.389,$$

whereas their semantic ordering is reversed:

$$-\overline{\Delta a}_{\text{A}} = 0.0328 > -\overline{\Delta a}_{\text{R}} = 0.0050.$$

The same reversal persists with RN50:

$$D_{\text{geo}}(\text{R}) = 0.621 > D_{\text{geo}}(\text{A}) = 0.451, \qquad -\overline{\Delta a}_{\text{A}} = 0.0474 > -\overline{\Delta a}_{\text{R}} = 0.0148.$$

Likewise, ImageNet-Sketch exhibits the largest global geometric discrepancy under both backbones, yet ImageNet-A exhibits the largest mean semantic degradation.

These observations show that stronger global second-order deformation does not necessarily entail stronger class-wise visual–text misalignment. We therefore treat the two diagnostics as *complementary aspects* of target-domain representation shift rather than interchangeable measurements of a single scalar severity. This empirical distinction motivates handling global geometric structure and class-wise semantic structure as separate adaptation targets in GeoSem-TTA.

B.3 Shift-Sensitive Projection Diagnostics

The two-dimensional visualization in Figure 1 is used only to expose the dominant structure of the global geometric shift; all reported D_{geo} values are computed in the original feature space. To quantify how much of the observed second-order deformation is represented by this low-dimensional view, we form the aggregate shift operator

$$\mathbf{A}_{\text{shift}} = \sum_{D \in \mathcal{T}} (\mathbf{S}_D - \mathbf{S}_{\text{IN}})^2.$$

where $\mathcal{T}$ contains the four target domains, and examine the cumulative energy captured by its leading eigendirections.

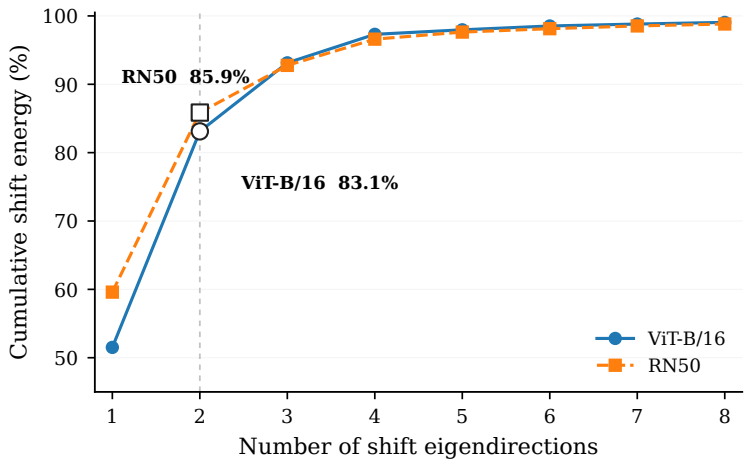


Figure 6: **Concentration of second-order shift energy.** Cumulative energy captured by the leading eigendirections of the aggregate shift operator A_{shift}. The top two directions capture 83.1% and 85.9% of the aggregate shift energy for ViT-B/16 and RN50, respectively, supporting the shift-sensitive low-dimensional visualization used for geometric analysis. The vertical axis is truncated for visibility.

As shown in Figure 6, the first two shift-sensitive directions capture 83.1% of the aggregate shift energy for ViT-B/16 and 85.9% for RN50. The cumulative energy further rises to approximately 93% with three directions and exceeds 96% with four. Thus, the two-dimensional shift-sensitive subspace captures the dominant aggregate deformation modes under both backbones, while the quantitative geometry statistics themselves remain full-dimensional.

Summary. Together, these analyses strengthen the representation-level motivation behind GeoSem-TTA in two respects: the geometric and semantic shifts persist across visual backbones, and the two diagnostics exhibit distinct domain-dependent severity patterns. Meanwhile, the dominant second-order deformation is strongly concentrated in a small number of shift-sensitive directions. These results support modeling global geometry and class-wise semantics as complementary structures during test-time adaptation.

C Entropy-Gradient Sensitivity under Erroneous Predictions

This section examines how entropy minimization behaves when an erroneous prediction dominates the predictive distribution. The analysis is intended to characterize a potential source of error reinforcement rather than to establish a universal divergence guarantee for arbitrary model parameterizations.

C.1 Exact Log-Ratio Derivation of the Entropy Gradient

Let $P_k(\boldsymbol{\Theta}) = P(y = k|\mathbf{x}_t; \boldsymbol{\Theta})$ denote the predicted probability for category $k \in \{1, \ldots, K\}$. For brevity, let $c = \hat{y}_t^{\text{wrong}}$ represent the dominant erroneous category under severe domain shifts. The predictive entropy loss is defined as:

$$\mathcal{L}_{\text{TTA}}(\mathbf{x}_t; \boldsymbol{\Theta}) = \mathcal{H}(P) = -\sum_{k=1}^{K} P_k(\boldsymbol{\Theta}) \ln P_k(\boldsymbol{\Theta}). \tag{19}$$

Evaluating the partial gradient of $\mathcal{L}_{\text{TTA}}$ with respect to adaptation parameters $\boldsymbol{\Theta}$:

$$\nabla_{\boldsymbol{\Theta}} \mathcal{L}_{\text{TTA}}(\mathbf{x}_t; \boldsymbol{\Theta}) = -\sum_{k=1}^{K} (1 + \ln P_k(\boldsymbol{\Theta})) \nabla_{\boldsymbol{\Theta}} P_k(\boldsymbol{\Theta}). \tag{20}$$

Because probability distributions normalize to unity ($\sum_{k=1}^{K} P_k(\boldsymbol{\Theta}) = 1$), differentiating both sides yields the identity $\sum_{k=1}^{K} \nabla_{\boldsymbol{\Theta}} P_k(\boldsymbol{\Theta}) = \mathbf{0}$. Substituting this identity into Equation (20) cancels the

constant 1:

$$\nabla_{\mathbf{\Theta}} \mathcal{L}_{\text{TTA}}(\mathbf{x}_t; \mathbf{\Theta}) = -\sum_{k=1}^{K} \ln P_k(\mathbf{\Theta}) \cdot \nabla_{\mathbf{\Theta}} P_k(\mathbf{\Theta}). \tag{21}$$

Using the normalization constraint again, the gradient of the dominant class c is expressed as the negative sum of the gradients of all non-dominant classes $k \neq c$:

$$\nabla_{\mathbf{\Theta}} P_c(\mathbf{\Theta}) = -\sum_{k \neq c} \nabla_{\mathbf{\Theta}} P_k(\mathbf{\Theta}). \tag{22}$$

Substituting Equation (22) into Equation (21) combines the dominant term with the non-dominant summation:

$$\begin{aligned}
\nabla_{\mathbf{\Theta}} \mathcal{L}_{\text{TTA}}(\mathbf{x}_t; \mathbf{\Theta}) &= -\ln P_c(\mathbf{\Theta}) \nabla_{\mathbf{\Theta}} P_c(\mathbf{\Theta}) - \sum_{k \neq c} \ln P_k(\mathbf{\Theta}) \nabla_{\mathbf{\Theta}} P_k(\mathbf{\Theta}) \\
&= -\ln P_c(\mathbf{\Theta}) \left(-\sum_{k \neq c} \nabla_{\mathbf{\Theta}} P_k(\mathbf{\Theta}) \right) - \sum_{k \neq c} \ln P_k(\mathbf{\Theta}) \nabla_{\mathbf{\Theta}} P_k(\mathbf{\Theta}) \\
&= \sum_{k \neq c} \left(\ln P_c(\mathbf{\Theta}) - \ln P_k(\mathbf{\Theta}) \right) \nabla_{\mathbf{\Theta}} P_k(\mathbf{\Theta}) \\
&= \sum_{k \neq c} \ln \left(\frac{P_c(\mathbf{\Theta})}{P_k(\mathbf{\Theta})} \right) \nabla_{\mathbf{\Theta}} P_k(\mathbf{\Theta}).
\end{aligned} \tag{23}$$

When the incorrect candidate c dominates the predictive mass ($P_c \gg P_k,\ \forall k \neq c$), the positive log-ratio weights satisfy $\ln(P_c/P_k) \gg 0$, emphasizing the contributions associated with non-dominant classes. In this dominant-error regime, the entropy-minimization descent direction tends to further favor the dominant prediction and can be approximated by the log-likelihood direction of the dominant error:

$$\nabla_{\mathbf{\Theta}} \mathcal{L}_{\text{TTA}}(\mathbf{x}_t; \mathbf{\Theta}) \sim -\nabla_{\mathbf{\Theta}} \ln P(\hat{y}_t^{\text{wrong}} | \mathbf{x}_t; \mathbf{\Theta}). \tag{24}$$

C.2 Local Error Reinforcement in Streaming Adaptation

Consider a streaming sequence where test instances are processed with discrete parameter updates:

$$\mathbf{\Theta}_{t+1} = \mathbf{\Theta}_t - \eta_t \nabla_{\mathbf{\Theta}} \mathcal{L}_{\text{TTA}}(\mathbf{x}_t; \mathbf{\Theta}_t) = \mathbf{\Theta}_t + \eta_t \alpha_t \nabla_{\mathbf{\Theta}} \ln P(\hat{y}_t^{\text{wrong}} | \mathbf{x}_t; \mathbf{\Theta}_t), \tag{25}$$

where $\alpha_t > 0$ summarizes the dominant log-ratio weighting derived above. Consider a subsequent target instance $\mathbf{x}$ processed after this update. A first-order Taylor expansion of its log-posterior for the same erroneous class c gives

$$\log P(c|\mathbf{x}; \mathbf{\Theta}_{t+1}) \approx \log P(c|\mathbf{x}; \mathbf{\Theta}_t) + \eta_t \alpha_t \nabla_{\mathbf{\Theta}} \log P(c|\mathbf{x}; \mathbf{\Theta}_t)^T \nabla_{\mathbf{\Theta}} \log P(c|\mathbf{x}_t; \mathbf{\Theta}_t). \tag{26}$$

The second term is positive whenever the two log-posterior gradients are positively aligned. Therefore, for target instances that induce locally aligned gradients, an update driven by an erroneous dominant prediction increases the local preference for the same erroneous class. Repeated exposure to such aligned instances can consequently reinforce prediction errors during streaming adaptation. This analysis characterizes a local feedback mechanism rather than a universal monotonic-divergence guarantee.

D Detailed Mathematical Derivations for GeoSem-TTA

In this appendix, we present the unabridged mathematical expansions for Gaussian LDA decision score derivation, Sherman-Morrison precision matrix updates, time-step scale alignment, and decoupled boundary properties.

D.1 Gaussian LDA Log-Posterior Expansion and Mahalanobis Bias

Let $\mathbf{x} \in \mathbb{S}^{d-1} \subset \mathbb{R}^d$ denote an L_2-normalized visual feature vector and $y \in \{1, \dots, K\}$ represent its category label. Under Bayes' rule, the posterior class probability is expressed as:

$$P(y = k|\mathbf{x}) = \frac{p(\mathbf{x}|y = k)P(y = k)}{\sum_{j=1}^{K} p(\mathbf{x}|y = j)P(y = j)}. \tag{27}$$

Defining the decision score as the natural logarithm of the unnormalized numerator in Equation (27):

$$\delta_k(\mathbf{x}) = \ln p(\mathbf{x}|y = k) + \ln P(y = k). \tag{28}$$

Assuming class-conditional densities follow multivariate Gaussian distributions $p(\mathbf{x}|y = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ with a shared covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ (Gaussian LDA):

$$p(\mathbf{x}|y = k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right). \tag{29}$$

Taking the logarithm yields:

$$\ln p(\mathbf{x}|y = k) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k). \tag{30}$$

Expanding the quadratic vector-matrix product:

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) = -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k. \tag{31}$$

Combining terms into the log-posterior decision score:

$$\delta_k(\mathbf{x}) = \underbrace{-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}_{\text{Class-Independent Constants}} + \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln P(y = k). \tag{32}$$

Because $-\frac{d}{2}\ln(2\pi)$, $-\frac{1}{2}\ln|\boldsymbol{\Sigma}|$, and $-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}$ are identical for all candidate classes $k \in \{1, \dots, K\}$, they do not alter the argmax ranking during classification. Under a uniform prior $P(y = k) = 1/K$, dropping class-independent terms isolates the exact LDA decision score:

$$\delta_k(\mathbf{x}) = \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k. \tag{33}$$

For streaming prediction at step t, only the states accumulated through step $t-1$ are available. Substituting the historical geometric state $\mathbf{P}_{t-1}$ and coupled centroid matrix $\boldsymbol{\Phi}_{t-1}$ into the Gaussian-LDA discriminant therefore yields the causal matrix-vector formulation in Equation (17) and Equation (18):

$$\mathbf{W}_t = \mathbf{P}_{t-1}\boldsymbol{\Phi}_{t-1} \in \mathbb{R}^{d\times K}, \qquad \mathbf{b}_t = -\frac{1}{2}\operatorname{diag}\left(\boldsymbol{\Phi}_{t-1}^T\mathbf{W}_t\right) \in \mathbb{R}^K, \qquad \boldsymbol{\delta}(\mathbf{x}_t) \propto \mathbf{W}_t^T\mathbf{x}_t + \mathbf{b}_t. \tag{34}$$

D.2 Sherman–Morrison Rank-1 Update for Uncentered Second-Order Statistics

Let

$$\mathbf{G}_t = \sum_{\tau=1}^{t} \lambda_\tau \mathbf{x}_\tau \mathbf{x}_\tau^T \in \mathbb{R}^{d\times d}$$

denote the reliability-weighted uncentered second-order statistic after processing $\mathbf{x}_t$. Accordingly,

$$\mathbf{G}_t = \mathbf{G}_{t-1} + \lambda_t \mathbf{x}_t \mathbf{x}_t^T, \qquad \mathbf{G}_0 = \mathbf{0}. \tag{35}$$

Adding ridge regularization $\gamma\mathbf{I}$, define the corresponding precision state as

$$\mathbf{P}'_t = (\mathbf{G}_t + \gamma\mathbf{I})^{-1}. \tag{36}$$

Setting $\mathbf{A}_t = \mathbf{G}_t + \gamma\mathbf{I}$, we have $\mathbf{P}'_t = \mathbf{A}_t^{-1}$. Using Equation (35), the regularized matrix satisfies

$$\mathbf{A}_t = \mathbf{A}_{t-1} + \lambda_t \mathbf{x}_t \mathbf{x}_t^T. \tag{37}$$

According to the Sherman–Morrison theorem (Sherman & Morrison, 1950), for any invertible matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and column vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{d\times 1}$:

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}. \tag{38}$$

Setting the rank-1 perturbation vectors as

$$\mathbf{u} = \sqrt{\lambda_t}\mathbf{x}_t \in \mathbb{R}^{d\times 1}, \qquad \mathbf{v}^T = \sqrt{\lambda_t}\mathbf{x}_t^T \in \mathbb{R}^{1\times d}, \tag{39}$$

where $\lambda_t \in [0, 1]$ is the sample reliability scalar, gives

$$\mathbf{P}'_t = (\mathbf{A}_{t-1} + \lambda_t\mathbf{x}_t\mathbf{x}_t^T)^{-1} = \mathbf{A}_{t-1}^{-1} - \frac{\mathbf{A}_{t-1}^{-1}(\sqrt{\lambda_t}\mathbf{x}_t)(\sqrt{\lambda_t}\mathbf{x}_t^T)\mathbf{A}_{t-1}^{-1}}{1 + (\sqrt{\lambda_t}\mathbf{x}_t^T)\mathbf{A}_{t-1}^{-1}(\sqrt{\lambda_t}\mathbf{x}_t)}. \tag{40}$$

Substituting $\mathbf{A}_{t-1}^{-1} = \mathbf{P}'_{t-1}$ and simplifying the scalar factors yields Equation 9:

$$\mathbf{P}'_t = \mathbf{P}'_{t-1} - \frac{\lambda_t\mathbf{P}'_{t-1}\mathbf{x}_t\mathbf{x}_t^T\mathbf{P}'_{t-1}}{1 + \lambda_t\mathbf{x}_t^T\mathbf{P}'_{t-1}\mathbf{x}_t}. \tag{41}$$

This completes the exact proof.

D.3 Time-Step Scale Alignment and Boundary Continuity

The auxiliary inverse state

$$\mathbf{P}'_t = (\mathbf{G}_t + \gamma\mathbf{I})^{-1} \tag{42}$$

is defined from the accumulated second-order statistic after processing $\mathbf{x}_t$. Because the magnitude of $\mathbf{G}_t$ grows with the elapsed stream length, we associate each post-update state with the elapsed-step factor

$$n_t = \max(1, t), \tag{43}$$

and define the corresponding scale-aligned geometric state as

$$\mathbf{P}_t = n_t\mathbf{P}'_t = \left(\frac{1}{n_t}\mathbf{G}_t + \frac{\gamma}{n_t}\mathbf{I}\right)^{-1}. \tag{44}$$

This is the state-aligned form of Equation 10: $\mathbf{G}_t$, $\mathbf{P}'_t$, and $\mathbf{P}_t$ all represent the geometry after processing $\mathbf{x}_t$. For prediction at step t, only the preceding state is available, so the classifier uses

$$\mathbf{P}_{t-1} = n_{t-1}\mathbf{P}'_{t-1}, \qquad n_{t-1} = \max(1, t-1), \tag{45}$$

which preserves strict predict-before-update causality.

Initial Boundary Continuity. Before the first test sample is processed, we initialize

$$\mathbf{G}_0 = \mathbf{0}, \qquad \mathbf{P}'_0 = \mathbf{P}_0 = \gamma^{-1}\mathbf{I}, \qquad \mathbf{\Phi}_0 = \mathbf{M}. \tag{46}$$

Hence, at the first prediction step,

$$\mathbf{W}_1 = \mathbf{P}_0\mathbf{\Phi}_0 = \gamma^{-1}\mathbf{M}, \qquad \mathbf{b}_1 = -\frac{1}{2\gamma}\mathbf{1}_K, \tag{47}$$

and the scaled score becomes

$$\boldsymbol{\delta}(\mathbf{x}_1) = \frac{1}{\tau}\mathbf{M}^T\mathbf{x}_1 - \frac{1}{2\tau}\mathbf{1}_K. \tag{48}$$

The class-independent offset preserves the zero-shot CLIP ranking, while $(\gamma/\tau)\gamma^{-1} = \tau^{-1}$ for any $\gamma > 0$.

D.4 Class-Assignment Independence of the Second-Order Statistic

The two adaptation tracks differ in how predictive evidence enters their sufficient statistics. For the geometric track, the rank-1 second-order observation

$$\mathbf{S}_t = \mathbf{x}_t\mathbf{x}_t^T \tag{49}$$

is determined solely by the incoming visual feature and contains no class index or pseudo-label assignment. Prediction-derived information enters Equation (9) only through the scalar reliability weight λ_t, which controls the contribution magnitude of $\mathbf{S}_t$ without assigning it to any category. In contrast, semantic evidence accumulation distributes evidence across classes through $\hat{\mathbf{p}}_t$

and $\mathbf{p}_t^{\text{safe}}$. GeoSem-TTA therefore decouples class-independent second-order geometry from class-specific first-order semantic adaptation: class-related evidence determines the semantic accumulation, while only the scalar reliability λ_t modulates the geometric update.

Algorithm 1 GeoSem-TTA Streaming Test-Time Adaptation Loop

Require: Streaming test features $\{\mathbf{x}_t\}_{t=1}^T$, static text classifier $\mathbf{M}$, parameters $\alpha, \gamma, \tau, \rho, \eta, \omega$.
Ensure: Predicted class labels $\{\hat{y}_t\}_{t=1}^T$.

Initialize $\mathbf{P}'_0 \leftarrow \gamma^{-1}\mathbf{I}_{d\times d}$, $\mathbf{P}_0 \leftarrow \gamma^{-1}\mathbf{I}_{d\times d}$, $\mathbf{V}_0 \leftarrow \mathbf{0}_{d\times K}$, $\mathbf{C}_0 \leftarrow \mathbf{0}_{d\times K}$, $\mathbf{\Phi}_0 \leftarrow \mathbf{M}$.
for $t = 1, \ldots, T$ **do**
  // Causal Gaussian-LDA classifier
  $\mathbf{W}_t \leftarrow \mathbf{P}_{t-1}\mathbf{\Phi}_{t-1}$, $\mathbf{b}_t \leftarrow -0.5 \cdot \mathrm{diag}(\mathbf{\Phi}_{t-1}^T \mathbf{W}_t)$
  // Predict before update
  $\boldsymbol{\delta}(\mathbf{x}_t) \leftarrow \frac{\gamma}{\tau}(\mathbf{W}_t^T \mathbf{x}_t + \mathbf{b}_t)$, $\hat{y}_t \leftarrow \arg\max_k \delta_k(\mathbf{x}_t)$
  Evaluate candidate evidence $\hat{\mathbf{p}}_t$ (Equation 11) and cosine gate $\mathbf{g}_t$ (Equation 12)
  // Reliability-filtered evidence
  $\mathbf{p}_t^{\text{safe}} \leftarrow \hat{\mathbf{p}}_t \odot \mathbf{g}_t$, $\lambda_t \leftarrow \sum_{k=1}^K p_{t,k}^{\text{safe}}$
  // Class-wise evidence accumulation
  $\mathbf{C}_t \leftarrow \mathbf{C}_{t-1} + \mathbf{x}_t \hat{\mathbf{p}}_t^T$, $\mathbf{V}_t \leftarrow \mathbf{V}_{t-1} + \mathbf{x}_t (\mathbf{p}_t^{\text{safe}})^T$
  Synthesize $\mathbf{\Phi}_t$ from $\mathbf{V}_t$ and $\mathbf{C}_t$ via Equation 16
  // Conditional precision update
  $\mathbf{P}'_t \leftarrow (\lambda_t > 0)\ ?\ \left(\mathbf{P}'_{t-1} - \frac{\lambda_t \mathbf{P}'_{t-1}\mathbf{x}_t\mathbf{x}_t^T\mathbf{P}'_{t-1}}{1+\lambda_t \mathbf{x}_t^T\mathbf{P}'_{t-1}\mathbf{x}_t}\right) : \mathbf{P}'_{t-1}$
  // Scale-aligned state for the next step
  $n_t \leftarrow \max(1, t)$, $\mathbf{P}_t \leftarrow n_t \mathbf{P}'_t$
end for

D.5 Complete End-to-End Streaming Algorithm

For completeness, the entire gradient-free, single-pass streaming test-time adaptation procedure of GeoSem-TTA is summarized in Algorithm 1.

E Extended Experiments and Implementation Details

E.1 Detailed Computational Efficiency and Hardware Overhead Profiling

To rigorously expose the underlying mathematical complexity and computational bottlenecks of different Test-Time Adaptation (TTA) paradigms, we introduce the **Streaming Feature-Level Adaptation Protocol**. This protocol evaluates the pure algorithmic adaptation and inference latency upon sample arrival, isolating the measurement from constant upstream overheads (e.g., disk I/O, CPU-bound multi-crop augmentation, and frozen visual encoder forward passes).

Establishing this protocol relies on three theoretical and empirical justifications:

- **Separation of Concerns**: In VLM-TTA, the pre-trained backbone (e.g., CLIP ViT-B/16) remains entirely frozen. Incorporating visual forward propagation (T_{Vision}) and data augmentation (T_{Aug}) introduces a heavy constant background noise (typically 70–90 ms) that masks the sub-millisecond algorithmic differences between competing adaptation mechanisms.
- **Hardware Artifact Elimination**: Raw image decoding latency is heavily bounded by host hardware heterogeneity (e.g., HDD vs. NVMe SSD, PCIe bandwidth). Using a unified pre-extracted feature stream eliminates exogenous hardware variables, ensuring absolute reproducibility.
- **Feature-Level Profiling Rationale**: Feature-level profiling isolates the adaptation and decision module from shared upstream costs such as image decoding and frozen visual encoding, allowing the computational differences among adaptation mechanisms to be measured directly.

As summarized in Table 2 in the main text, evaluations are conducted on a large-scale vocabulary setting ($d = 512$, $K = 1000$ classes) tracking the exact ImageNet scale geometry under strict single-pass streaming conditions (batch size 1), measured on a single NVIDIA RTX 3090 GPU (24GB VRAM).

Table 8: **Full dataset breakdown of the component-wise ablation study on CLIP ViT-B/16 across the ten Cross-Dataset benchmarks.**

Configuration	Cal.	DTD	Cars	Euro.	Air.	Flow.	Pets	UCF	Food	SUN	Avg.
1. Global Geometry Adaptation	94.20	54.12	66.30	54.10	26.50	74.20	90.10	72.10	85.50	69.50	68.66
2. + Bias ($\mathbf{b}_t$)	94.90	55.40	67.10	58.20	28.10	77.30	91.20	73.80	86.40	**71.40**	70.38
3. + Empirical Mean ($\mathbf{C}_t$)	95.00	55.80	67.20	58.90	28.70	77.80	91.50	74.20	86.50	71.10	70.67
4. + Zero-Init ($\mathbf{V}_0 = \mathbf{0}$)	95.20	56.10	67.40	59.60	29.30	79.50	91.90	74.80	86.60	70.00	71.03
5. Complete GeoSem-TTA	**95.25**	**56.15**	**67.42**	**59.75**	**29.43**	**79.62**	**91.96**	**74.97**	**86.68**	69.87	**71.11**

Table 9: **Full Top-1 classification accuracy (%) on ImageNet and four OOD benchmarks using CLIP ResNet-50**.

Method	ImageNet	-A	-V2	-R	-Sketch	OOD Avg.	Total Avg.
CLIP-ResNet-50	58.16	21.83	51.41	56.15	33.37	40.69	44.18
CoOp (16-shot)	63.33	23.06	55.40	56.60	34.67	42.43	46.61
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
TDA	61.35	30.29	55.54	62.58	38.12	46.63	49.58
DPE	63.41	30.15	56.72	**63.72**	40.03	47.66	50.81
DLAE	**64.10**	**34.80**	**57.13**	63.63	**40.68**	**49.06**	**52.06**
SCA	63.11	30.64	56.51	62.85	39.33	47.33	50.49
GeoSem-TTA (Ours)	63.59	34.35	56.78	63.22	39.34	48.42	51.46

Mechanism Attribution for Computational Bottlenecks. The measured costs reflect distinct adaptation mechanisms. TPT and DiffTPT require gradient-based prompt optimization, while TDA incurs cache construction and retrieval overhead; DPE and DLAE reduce this cost but still rely on iterative optimization. SCA avoids backpropagation but requires solving a discriminative linear system.

GeoSem-TTA instead maintains its geometric state through $O(d^2)$ Sherman–Morrison updates and its semantic state through compact streaming statistics. This closed-form formulation reduces feature-level adaptation latency to 0.86 ms (1165.0 steps/s) while retaining a constant memory footprint.

E.2 Complete 10-Dataset Breakdown for Component-Wise Ablations

Table 8 provides the complete dataset-level breakdown of the component-wise ablation study on the Cross-Dataset suite using CLIP ViT-B/16.

E.3 Boundary Analysis on Diffuse Non-Object Texture Manifolds (DTD)

On the Describable Textures Dataset (DTD (Cimpoi et al., 2014)), the visual categories are dominated by texture patterns rather than compact object-centric structures, providing a distinct boundary case for the shared-geometric modeling used by GeoSem-TTA. As reported in Table 4 and Table 7, GeoSem-TTA achieves 56.15% on ViT-B/16 and 53.01% on ResNet-50, respectively. Several competing methods obtain higher accuracy on DTD under both backbones, showing that diffuse texture-dominated categories remain a boundary case for the shared-geometric formulation despite its stronger aggregate Cross-Dataset performance.

E.4 Complete ResNet-50 Benchmark Evaluations

Table 9 provides the complete ImageNet-family evaluation using the pretrained CLIP ResNet-50 ($d = 1024$) backbone; the corresponding 10-dataset Cross-Dataset results are reported in Table 7 of the main text.

Table 10: **Corruption robustness with CLIP ViT-B/16.** Top-1 accuracy (%) averaged equally over 15 independently adapted corruption streams at severity 5.

Method	CIFAR-10-C	CIFAR-100-C	ImageNet-C	Mean
CLIP	60.20	35.20	25.50	40.30
TDA	<u>67.37</u>	38.24	27.67	44.43
DPE	66.31	39.70	<u>27.68</u>	<u>44.56</u>
SCA	66.75	38.11	26.69	43.85
ADAPT	64.50	**40.70**	27.20	44.13
GeoSem-TTA (Ours)	**68.08**	<u>39.73</u>	**28.83**	**45.55**

E.5 Corruption Robustness Evaluation

We further evaluate corruption robustness on CIFAR-10-C, CIFAR-100-C, and ImageNet-C using CLIP ViT-B/16. Each benchmark uses its native label space and corresponding text classifier (10, 100, and 1,000 classes, respectively). Following the common corruption setting, we evaluate all 15 corruption types at severity 5. Each corruption is treated as an independently adapted test stream: the adaptation state is re-initialized before every corruption, and the reported benchmark accuracy is the equally weighted average over the 15 corruption streams. All baselines are reproduced using the same backbone, benchmark, severity, corruption subset, and label-space settings while retaining their original adaptation protocols.

As shown in Table 10, GeoSem-TTA achieves the highest accuracy on CIFAR-10-C (68.08%) and ImageNet-C (28.83%), while ranking second on CIFAR-100-C (39.73%). It yields the best overall mean of 45.55%, outperforming the strongest competing mean of 44.56%. These results further demonstrate robust adaptation across common corruptions and label spaces of different scales.

E.6 Robustness to Class-Imbalanced Test Streams

Following the long-tailed test-stream construction of L-TTA (Wang et al., 2026), we further evaluate GeoSem-TTA under class-imbalanced streams using CLIP ViT-B/16. For a dataset with C classes, classes are ordered by their original cardinalities and randomly subsampled to form an exponentially decayed class distribution,

$$n_i = \max\left(\left\lfloor n_{\max}\left(\frac{1}{\kappa}\right)^{\frac{i}{C-1}}\right\rfloor, 1\right),$$

where $n_{\max}$ is the maximum class cardinality and $\kappa \in \{10, 20, 50\}$ denotes the imbalance ratio. Samples excluded by subsampling do not participate in prediction, online state updates, or evaluation. Each $(\mathrm{seed}, \mathrm{dataset}, \kappa)$ setting starts from a fresh adaptation state. GeoSem-TTA and the reproduced baselines use their fixed original configurations without long-tail-specific tuning. We use five independent random seeds and report Accuracy and Macro-F1.

ImageNet-Family Evaluation. We first examine whether GeoSem-TTA remains effective as the class distribution becomes progressively more imbalanced across the five ImageNet-family benchmarks. Table 5 summarizes the overall results, while Table 11 provides the complete dataset-level comparison at all three imbalance levels. Results for CLIP, TPT, TDA, and DPE follow the same L-TTA evaluation protocol (Wang et al., 2026); SCA, DLAE, and GeoSem-TTA are evaluated under the same stream construction using their fixed method-specific configurations.

Across the three imbalance levels, GeoSem-TTA achieves the highest overall Accuracy and Macro-F1 among the compared methods. More importantly, the two metrics exhibit different responses as imbalance increases: overall Accuracy changes only from 66.54% at $\kappa = 10$ to 66.86% at $\kappa = 50$, whereas Macro-F1 decreases from 61.75% to 57.68%. Thus, aggregate recognition accuracy remains stable as class imbalance increases, whereas performance across classes is more affected as imbalance grows. This distinction indicates that robustness in sample-weighted recognition should not be interpreted as insensitivity to class imbalance.

The dataset-level results in Table 11 further show that the aggregate behavior is supported across multiple target domains rather than by a single benchmark. GeoSem-TTA obtains the best Total Avg.

Table 11: **Detailed ImageNet-family results under increasing class imbalance.** Each dataset reports Accuracy (Acc.) and Macro-F1 (F1). OOD Avg. averages ImageNet-A, ImageNet-R, ImageNet-Sketch, and ImageNet-V2, while Total Avg. additionally includes ImageNet. Best and second-best results are shown in **bold** and underline, respectively.

(a) $\kappa = 10$

	ImageNet		-A		-R		-Sketch		-V2		OOD Avg.		Total Avg.	
Method	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
CLIP	63.75	58.94	46.84	43.71	72.27	70.40	42.17	40.74	59.63	51.62	55.23	51.62	56.93	53.08
TPT	65.53	60.36	52.79	48.08	77.33	74.28	45.22	43.60	61.34	55.70	59.17	55.42	60.44	56.40
TDA	69.10	63.89	60.00	54.30	81.23	78.00	47.55	45.77	66.21	60.53	63.75	59.65	64.82	60.50
DPE	70.16	64.19	60.31	54.41	80.58	77.42	49.32	45.45	**67.78**	**63.32**	64.50	60.15	65.63	60.96
SCA	71.83	67.57	64.26	58.07	81.04	78.02	**50.44**	**46.05**	62.87	57.69	64.65	59.96	66.09	61.48
DLAE	71.93	67.67	62.33	56.66	79.96	77.23	48.93	44.97	64.76	58.76	63.99	59.41	65.58	61.06
GeoSem-TTA (Ours)	**72.09**	**67.85**	**64.41**	**58.32**	**81.27**	**78.09**	50.10	45.39	64.85	59.09	**65.16**	**60.22**	**66.54**	**61.75**

(b) $\kappa = 20$

	ImageNet		-A		-R		-Sketch		-V2		OOD Avg.		Total Avg.	
Method	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
CLIP	63.36	57.33	46.89	41.68	72.57	69.30	42.25	40.45	58.36	50.60	55.02	50.51	56.69	51.87
TPT	65.22	60.70	53.30	47.11	77.76	72.82	46.04	43.79	60.78	54.81	59.47	54.63	60.62	55.85
TDA	68.92	63.36	60.40	53.38	81.60	75.94	45.65	43.46	64.70	56.17	63.09	57.24	64.25	58.46
DPE	67.41	63.53	60.05	53.87	81.18	75.87	**50.10**	**45.02**	**65.50**	**60.53**	64.21	**58.82**	64.85	59.76
SCA	71.66	64.73	64.66	57.55	81.44	76.15	49.83	43.32	62.11	56.51	64.51	58.38	65.94	59.65
DLAE	72.34	65.62	62.70	56.22	80.58	75.79	48.63	42.81	64.12	57.20	64.01	58.01	65.67	59.53
GeoSem-TTA (Ours)	**72.38**	**65.67**	**64.74**	**57.75**	**81.79**	**76.34**	49.62	42.89	64.19	57.65	**65.08**	58.66	**66.54**	**60.06**

(c) $\kappa = 50$

	ImageNet		-A		-R		-Sketch		-V2		OOD Avg.		Total Avg.	
Method	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
CLIP	63.18	53.70	44.74	36.26	71.52	62.60	39.36	36.08	58.03	50.58	53.41	46.38	55.37	47.84
TPT	65.29	56.57	51.13	40.84	76.77	66.30	42.99	41.28	61.27	54.95	58.04	50.84	59.49	51.99
TDA	68.07	62.06	58.97	49.12	80.89	70.04	44.87	41.49	64.76	58.50	62.37	54.79	63.51	56.24
DPE	68.04	**62.37**	60.21	47.46	80.76	69.96	48.07	**43.50**	**65.80**	**60.78**	63.71	55.42	64.58	56.81
SCA	71.32	60.53	65.36	55.21	82.23	73.27	49.22	39.88	62.20	56.63	64.75	56.24	66.07	57.10
DLAE	72.41	61.78	63.86	54.71	81.64	**73.68**	48.41	40.28	64.29	56.87	64.55	56.38	66.12	57.46
GeoSem-TTA (Ours)	**72.48**	62.04	**65.70**	**55.88**	**82.62**	73.54	**49.26**	39.73	64.25	57.21	**65.46**	**56.59**	**66.86**	**57.68**

Table 12: **Cross-Dataset robustness under increasing class imbalance.** Ten-dataset mean Accuracy (Acc.) and Macro-F1 (F1), reported as mean $\pm$ sample standard deviation over five independently sampled streams.

	$\kappa = 10$		$\kappa = 20$		$\kappa = 50$	
Method	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
GeoSem-TTA (Ours)	68.51 ± 0.50	64.87 ± 0.41	68.16 ± 0.11	62.94 ± 0.19	67.41 ± 0.84	59.67 ± 0.86

Accuracy and Macro-F1 at every imbalance level and the best OOD Avg. Accuracy throughout, while its OOD Avg. Macro-F1 remains best at $\kappa = 10$ and $\kappa = 50$ and second-best at $\kappa = 20$. At the same time, the per-dataset results expose clear boundary cases: DPE remains stronger on ImageNet-V2, and several Macro-F1 entries on ImageNet-Sketch favor competing methods as imbalance becomes severe. The results therefore support robustness to increasing class-frequency skew at the aggregate level without implying uniform dominance across every domain and metric.

Cross-Dataset Generalization under Class Imbalance. We next test whether the same imbalance response extends beyond the ImageNet family to the ten heterogeneous Cross-Dataset benchmarks. This experiment is used as a complementary generalization check rather than a method-wise long-tail comparison. As summarized in Table 12, GeoSem-TTA is evaluated with the same fixed configuration while only the class-frequency distribution of the incoming stream is varied.

Table 13: **Hyperparameter sensitivity across the ImageNet family.** One dataset-level adaptation parameter is varied at a time while the remaining parameters are fixed to the corresponding dataset configuration. Accuracy (%) is reported for each tested value; bold entries indicate the configuration used for that dataset.

(a) ImageNet-A

γ	Value	25	50	75	**100**	150	250	500
	Acc.	63.58	64.36	64.44	**64.51**	64.48	64.29	64.18
ρ	Value	0	0.10	0.20	**0.25**	0.30	0.50	1.00
	Acc.	64.39	64.39	64.38	**64.51**	64.36	63.55	63.55
η	Value	0.005	0.02	0.04	**0.06**	0.08	0.12	0.20
	Acc.	64.23	64.35	64.48	**64.51**	64.47	64.38	63.67
ω	Value	0	0.05	0.10	**0.15**	0.20	0.30	0.50
	Acc.	64.31	64.34	64.31	**64.51**	64.44	63.49	54.59

(b) ImageNet-V2

γ	Value	2.5k	5k	**10k**	12.5k	15k	25k	50k
	Acc.	65.34	65.41	**65.52**	65.46	65.46	65.51	65.42
ρ	Value	0	0.05	**0.10**	0.20	0.30	0.50	1.00
	Acc.	65.49	65.52	**65.52**	65.52	65.47	65.39	65.39
η	Value	0	0.0025	**0.005**	0.01	0.02	0.05	0.10
	Acc.	65.46	65.50	**65.52**	65.52	65.43	65.40	65.18
ω	Value	0	0.01	**0.02**	0.03	0.05	0.10	0.30
	Acc.	65.45	65.51	**65.52**	65.51	65.43	65.32	60.70

(c) ImageNet-R

γ	Value	100	**200**	300	400	500	750	1200
	Acc.	81.00	**81.20**	81.20	81.19	81.18	81.12	81.06
ρ	Value	0	0.10	0.20	**0.30**	0.40	0.60	1.00
	Acc.	81.18	81.18	81.18	**81.20**	80.42	80.42	80.42
η	Value	0.005	0.02	0.05	**0.07**	0.09	0.15	0.25
	Acc.	81.08	81.09	81.18	**81.20**	81.08	81.06	80.26
ω	Value	0	0.05	0.10	**0.15**	0.20	0.30	0.50
	Acc.	81.06	81.14	81.16	**81.20**	81.15	80.64	69.75

(d) ImageNet-Sketch

γ	Value	200	350	450	**550**	650	850	1200
	Acc.	52.12	52.44	52.52	**52.55**	52.53	52.46	52.36
ρ	Value	0	0.05	0.10	0.15	**0.20**	0.40	1.00
	Acc.	52.55	52.55	52.54	52.54	**52.55**	51.44	51.40
η	Value	0	0.005	0.01	0.02	**0.03**	0.05	0.10
	Acc.	52.53	52.53	52.49	52.54	**52.55**	52.51	52.19
ω	Value	0	0.05	0.10	**0.15**	0.20	0.30	0.50
	Acc.	51.71	52.12	52.47	**52.55**	52.30	51.08	43.97

(e) ImageNet

γ	Value	10k	30k	50k	**70k**	90k	110k	130k
	Acc.	72.07	72.14	72.18	**72.18**	72.18	72.17	72.18
ρ	Value	0	0.05	**0.10**	0.15	0.20	0.30	0.50
	Acc.	72.18	72.18	**72.18**	72.18	72.17	72.04	72.00
η	Value	0.01	0.03	0.05	**0.07**	0.09	0.12	0.15
	Acc.	72.04	72.05	72.10	**72.18**	72.11	71.81	71.56
ω	Value	0	0.05	**0.08**	0.10	0.15	0.20	0.30
	Acc.	72.05	72.15	**72.18**	72.14	71.69	71.19	69.38

The Cross-Dataset results reproduce the same metric-dependent trend on a substantially different collection of recognition tasks. As κ increases from 10 to 50, average Accuracy decreases from 68.51% to 67.41%, a change of only -1.09 points, whereas Macro-F1 decreases from 64.87% to 59.67%, or -5.20 points. Together with the ImageNet-family results, this provides consistent evidence that GeoSem-TTA preserves relatively stable overall recognition under increasingly imbalanced test streams, while performance across classes is more affected by severe class-frequency skew.

E.7 Hyperparameter Sensitivity and Stream-Order Robustness

Hyperparameter Sensitivity. We evaluate the sensitivity of GeoSem-TTA to the four dataset-level adaptation parameters γ, ρ, η, and ω across ImageNet and its four OOD variants using CLIP ViT-B/16. We adopt a one-at-a-time (OAT) protocol: for each dataset, one parameter is varied while the remaining parameters are fixed to the corresponding configuration used in evaluation. Table 13 reports the complete sweeps.

Across the ImageNet family, γ, ρ, and η remain stable around their adopted configurations, whereas ω exhibits the clearest boundary behavior: moderate coupling preserves performance, while excessively large values cause substantially larger degradation, particularly on ImageNet-A, ImageNet-R, and ImageNet-Sketch. The consistent patterns across the five benchmarks indicate that the selected dataset-level configurations lie within stable operating regions rather than isolated performance peaks.

Stream-Order Robustness. We additionally evaluate GeoSem-TTA over 10 random permutations of the test stream. The resulting variation remains limited to $\pm 0.12\%$, indicating stable performance across the evaluated stream orders.

Table 14: Detailed statistics of the 15 benchmark datasets evaluated in this work.

Benchmark	Dataset	Classes	Test Samples
Cross-Dataset	Caltech101 (Fei-Fei et al., 2004)	100	2,465
	DTD (Cimpoi et al., 2014)	47	1,692
	Stanford Cars (Krause et al., 2013)	196	8,041
	EuroSAT (Helber et al., 2019)	10	8,100
	FGVC Aircraft (Maji et al., 2013)	100	3,333
	Oxford Flowers (Nilsback & Zisserman, 2008)	102	2,463
	Oxford Pets (Parkhi et al., 2012)	37	3,669
	UCF101 (Soomro et al., 2012)	101	3,783
	Food101 (Bossard et al., 2014)	101	30,300
	SUN397 (Xiao et al., 2010)	397	19,850
OOD ImageNet	ImageNet (Deng et al., 2009)	1,000	50,000
	ImageNet-A (Hendrycks et al., 2021b)	200	7,500
	ImageNet-V2 (Recht et al., 2019)	1,000	10,000
	ImageNet-R (Hendrycks et al., 2021a)	200	30,000
	ImageNet-Sketch (Wang et al., 2019)	1,000	50,889

E.8 Benchmark and Evaluation Setup

For the primary Cross-Dataset and ImageNet OOD evaluations, we use 15 benchmarks comprising a 10-dataset Cross-Dataset suite and a 5-dataset ImageNet-family OOD suite, with detailed dataset specifications summarized in Table 14. We benchmark against Zero-Shot CLIP (Radford et al., 2021), 16-shot CoOp (Zhou et al., 2022b), prompt-tuning methods TPT (Shu et al., 2022) and DiffTPT (Feng et al., 2023), and cache/statistics-based methods TDA (Karmanov et al., 2024), DPE (Zhang et al., 2024a), SCA (Guan et al., 2025), BCA (Zhou et al., 2025), ADAPT (Zhang et al., 2025), and DLAE (Wu et al., 2026). All reproduced methods are evaluated under single-pass streaming TTA ($N = 1$, batch size one), using a single crop on the Cross-Dataset benchmarks and 63 random crops on the ImageNet OOD benchmarks. BCA and ADAPT use the online results reported in their original papers. Experiments use pre-trained CLIP ResNet-50 ($d = 1024$) and ViT-B/16 ($d = 512$) with the fixed CLIP inverse temperature $\tau^{-1} = 100$. The candidate-set coefficient is fixed globally to $\alpha = 0.1$ in all experiments. The adaptation parameters γ, ρ, η, and ω use dataset-specific configurations that are fixed throughout the corresponding evaluations, without condition-specific retuning for corruption severity, class imbalance, stream order, or stream duration. Their ImageNet-family configurations and sensitivity ranges are reported in Table 13. Here, ρ denotes the cosine-similarity gating threshold.

Benchmark Diversity. As summarized in Table 14, the evaluation suite spans diverse recognition regimes, from 10-class remote sensing to 397-class scene recognition, together with texture, fine-grained object, action, food, and ImageNet-family distribution shifts. The benchmarks also vary substantially in class cardinality and stream length, providing heterogeneous conditions for evaluating streaming adaptation.

E.9 Detailed Cross-Domain Evaluation

Table 15 provides the complete domain-level results corresponding to the benchmark-level comparison in Table 3, covering 16 evaluation domains across OfficeHome, PACS, VLCS, and TerraIncognita.

Domain-Level Behavior. The benchmark-level gains in Table 3 are supported by the domain-level results in Table 15, rather than by a single outlier. GeoSem-TTA leads all four OfficeHome domains and shows particularly consistent improvements on the more shifted VLCS domains, yielding benchmark means of 86.47% and 87.52%, respectively.

Table 15: **Detailed top-1 accuracy (%) across 16 evaluation domains of OfficeHome, PACS, VLCS, and TerraIncognita with CLIP ViT-B/16.** TPT and CLIPArTT use reported results; all other baselines are reproduced under our unified streaming protocol.

Dataset	Domain	CLIP	TPT	CLIPArTT	TDA	DPE	SCA	ADAPT	DLAE	GeoSem-TTA
OfficeHome	Art	85.83	75.76	73.84	86.07	86.11	86.20	86.16	86.03	**86.98**
	Clipart	71.75	63.08	63.54	73.01	73.08	74.04	73.22	72.94	**74.27**
	Product	90.56	84.07	85.23	91.17	91.71	91.94	92.23	90.83	**92.75**
	Real World	90.84	85.89	87.61	90.77	90.89	90.80	91.14	90.87	**91.88**
	Mean	84.75	77.20	77.56	85.26	85.45	85.74	85.69	85.17	**86.47**
PACS	Art	98.19	95.52	96.57	98.10	98.29	98.29	98.44	97.41	**98.63**
	Cartoon	98.98	94.77	96.00	99.02	99.02	98.98	99.06	98.93	**99.19**
	Photo	99.82	99.42	99.28	99.88	99.88	99.88	99.82	99.88	**99.94**
	Sketch	89.87	83.22	83.93	90.66	90.46	90.00	89.79	**90.74**	89.95
	Mean	96.72	93.23	93.95	96.91	96.91	96.79	96.78	96.74	**96.93**
VLCS	Caltech101	99.70	99.36	99.51	**99.90**	99.70	99.39	96.06	91.52	**99.90**
	LabelMe	72.46	54.88	67.96	70.15	71.87	70.41	69.17	71.05	**79.83**
	SUN09	72.22	67.30	68.68	73.10	73.27	74.62	75.86	78.08	**83.41**
	VOC2007	83.79	76.74	84.09	82.10	84.05	80.87	76.64	72.75	**86.92**
	Mean	82.04	74.57	80.06	81.31	82.22	81.32	79.43	78.35	**87.52**
TerraIncognita	L38	39.57	37.04	20.61	43.83	42.37	41.60	39.04	19.40	**44.27**
	L43	48.98	32.84	32.69	50.72	49.83	47.66	46.49	41.04	**50.75**
	L46	34.63	30.90	31.75	37.62	35.17	37.31	36.39	31.87	**45.20**
	L100	52.31	29.02	52.06	53.27	51.24	**59.79**	32.71	27.28	52.98
	Mean	43.87	32.45	34.28	46.66	44.94	46.59	38.66	29.90	**48.30**
Overall Mean		76.85	69.36	71.46	77.54	77.38	77.61	72.64	70.79	**79.81**

Behavior under Stronger Domain Variation. On the comparatively saturated PACS benchmark, GeoSem-TTA achieves a competitive 96.93% mean without uniformly leading every domain. TerraIncognita presents a harder setting, where GeoSem-TTA obtains the highest benchmark mean of 48.30% despite not leading L100. These results show that the aggregate cross-domain improvement persists across substantially different domain structures without implying uniform dominance on every constituent domain.