BEYOND FIXED GEOMETRY: RELIABLE TEST-TIME ADAPTATION OF VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Under test-time distribution shift, vision–language models can fail along two distinct reliability axes: the pretrained similarity geometry may misweight target-relevant feature directions, while augmented views may remain confidently wrong. We introduce a composite heteroscedastic reference model whose MAP rule exposes two distinct reliability roles: directional comparison precision and view-specific observation precision. Guided by these roles, GeoRel (Geometry–Reliability Factorized Test-Time Adaptation) learns a shared positive-semidefinite comparison geometry from weak cross-modal class evidence through discriminative attraction, separation, and redundancy conditioning, while ranking augmented observations through scale-free agreement between frozen-space query consistency and adaptive predictive certainty. EMA separately stabilizes noisy target-visual evidence used by the geometry-adapted prediction. Across natural-shift and cross-dataset benchmarks with RN50 and ViT-B/16, GeoRel achieves strong and broadly competitive performance. The resulting principle is simple: adapt the comparison geometry, but anchor view reliability in frozen space.

1 INTRODUCTION

Zero-shot vision–language models (VLMs), exemplified by CLIP (Radford et al., 2021), recognize images by comparing visual representations with textual class prototypes in a shared embedding space. Their broad transfer ability makes them attractive for deployment without task-specific training, but performance can deteriorate when the test distribution differs from pretraining. Test-time adaptation (TTA) addresses this setting by adapting the model from unlabeled test observations. Recent VLM-TTA methods modify prompts, maintain dynamic caches or prototypes, introduce lightweight adapters, or correct prediction logits (Shu et al., 2022; Karmanov et al., 2024; Zhang et al., 2024a; Farina et al., 2024; Liang et al., 2025; Maharana et al., 2025; Wu et al., 2026). Despite this diversity, the geometry used to compare visual and textual evidence is rarely treated as an explicit adaptation target.

This omission matters because the pretrained embedding space need not provide the most discriminative comparison rule after distribution shift. Classical metric learning shows that task supervision can reshape feature space to separate locally confusable classes (Weinberger & Saul, 2009; Kulis, 2012), while domain-adaptation work demonstrates that second-order feature structure can matter under distribution change (Sun et al., 2016; Sun & Saenko, 2016). More recently, CLIP image and text embeddings have been shown to exhibit structured, non-spherical and offset ellipsoidal geometry (Levi & Gilboa, 2025). These observations motivate a question largely orthogonal to adapting prompts or prototypes: when target evidence changes, should the comparison geometry itself remain fixed?

A separate difficulty arises from the augmented views widely used in test-time VLM adaptation. Entropy- or confidence-based selection can emphasize views that are decisive without being semantically faithful to the original query. Recent work has therefore explored view quality, calibration, consistency, and semantic anchors from several directions (Zanella & Ben Ayed, 2024; Yoon et al., 2024; Sharifdeen et al., 2025; Liang et al., 2025; Choi & Kim, 2026; Wang et al., 2026; Song et al., 2026). Collectively, these studies indicate that predictive confidence alone is an incomplete reliability signal under distribution shift: an augmented crop may receive a sharp prediction while suppressing foreground evidence or emphasizing an irrelevant region. This creates a second question that should be distinguished from geometry adaptation: which observations should be trusted?

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.

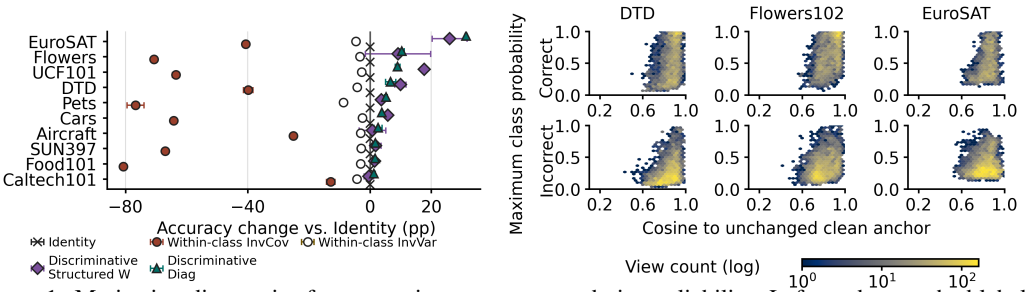

Figure 1: Motivation diagnostics for comparison geometry and view reliability. Left: under matched label-assisted splits, within-class InvCov/InvVar degrade accuracy relative to Identity, while discriminatively learned Structured W and Diagonal transforms reveal substantial task-aligned geometric headroom. Right: incorrect augmented views can remain highly confident.

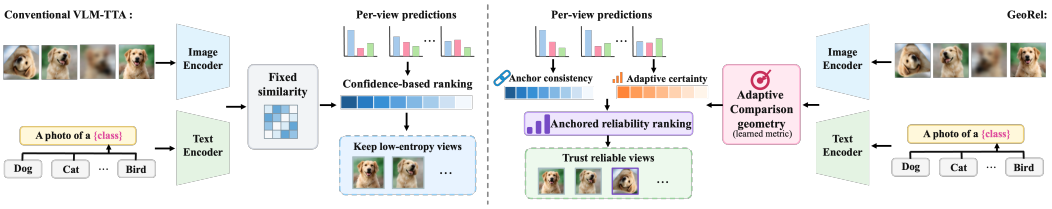

Figure 2: Conceptual decomposition underlying GeoRel. Conventional confidence-driven adaptation keeps comparison geometry fixed and scores augmented views using a single predictive signal. GeoRel separates discriminative geometry adaptation from anchored observation reliability.

Figure 1 makes the distinction concrete. Label-assisted covariance-based transformations fail to improve the pretrained comparison rule, whereas discriminatively learned transforms reveal substantial geometric headroom; meanwhile, incorrectly classified augmented views can remain highly confident. These observations motivate treating comparison geometry and observation reliability as separate adaptation problems.

We formalize this distinction with a composite heteroscedastic reference model that exposes a directional comparison role and an observation-weighting role. GeoRel (Geometry–Reliability Factorized Test-Time Adaptation) realizes the former through shared discriminative geometry adaptation and the latter through anchored ordinal reliability based on frozen-space consistency and adaptive certainty. The resulting asymmetry lets comparison geometry adapt while retaining a stable reference for judging augmented evidence (Figure 2).

Our contributions are:

- We formulate reliable VLM-TTA around two distinct structural roles: target comparison geometry and view-dependent observation reliability. This factorization provides a unified way to separate where class evidence is compared from which observations should influence prediction.
- We introduce GeoRel, which realizes these roles through shared discriminative geometry adaptation and anchored, scale-free reliability ranking. The geometry is learned from weak cross-modal class evidence, while view reliability combines frozen-space consistency with adaptive predictive certainty through ordinal agreement.
- We validate both mechanisms across natural-shift and cross-dataset benchmarks with RN50 and ViT-B/16. Controlled analyses show that the learned geometry improves discriminative separation, while anchored reliability selects cleaner multi-view evidence and provides complementary predictive information.

We discuss the broader VLM-TTA, calibration, and metric-learning literature in Appendix G.

2 Reference-Model Reliability Factorization for VLM-TTA

We now formalize the two structural roles introduced above. Let $\mathbf{x}_i \in \mathbb{R}^D$, $i \in \mathcal{I}$, denote the normalized feature of an augmented observation with class $y \in \{1, \ldots, C\}$, and let $\boldsymbol{\mu}_c$ denote the target-domain center of class c. We introduce the following composite heteroscedastic model as a

tractable *reference model* that exposes separate directional-comparison and observation-weighting roles in the Bayes rule.

Assumption 1 (Composite heteroscedastic reference model)**.** *For class c and view $i \in \mathcal{I}$,*

$$\mathbf{x}_i \mid y = c \sim \mathcal{N}\left(\boldsymbol{\mu}_c, \frac{\boldsymbol{\Sigma}_t}{\gamma_i}\right), \tag{1}$$

where $\boldsymbol{\Sigma}_t \succ 0$ is shared across classes and captures target-domain directional uncertainty, and $\gamma_i > 0$ is view-specific precision. Views are independent given y, and class priors are equal.

A standard Gaussian discriminant calculation gives the following decomposition.

Proposition 1 (Reference-model reliability factorization)**.** *Let $\mathbf{M}_t = \boldsymbol{\Sigma}_t^{-1}$ and $\Gamma = \sum_{i\in\mathcal{I}} \gamma_i$. Under Assumption 1, if class centers have equal $\mathbf{M}_t$-norm, the MAP rule reduces to*

$$\widehat{y} = \arg\max_c \boldsymbol{\mu}_c^\top \mathbf{M}_t \overline{\mathbf{x}}_\gamma, \qquad \overline{\mathbf{x}}_\gamma = \frac{1}{\Gamma}\sum_{i\in\mathcal{I}} \gamma_i \mathbf{x}_i. \tag{2}$$

Eq. (2) separates two reliability roles: $\mathbf{M}_t$ determines *where* class evidence is compared by weighting feature directions, while $\{\gamma_i\}$ determine *which* observations dominate inference. Appendix A.1 gives the full MAP rule and proof. In particular,

$$\boldsymbol{\mu}_c^\top \mathbf{M}_t \overline{\mathbf{x}}_\gamma = \sum_{i\in\mathcal{I}} \frac{\gamma_i}{\Gamma} \boldsymbol{\mu}_c^\top \mathbf{M}_t \mathbf{x}_i, \tag{3}$$

Thus, Proposition 1 establishes the structural decomposition: latent $\mathbf{M}_t = \boldsymbol{\Sigma}_t^{-1}$ controls directional comparison, while latent $\{\gamma_i\}$ control relative view influence.

From roles to operational surrogates. The factorization has two direct design consequences. GeoRel realizes the shared comparison role with one common geometry across image, text, and target-visual representations, and the observation-specific role with instance-level ordinal reliability. We intentionally retain a frozen-space consistency reference while allowing comparison geometry to adapt; Figure 3 summarizes this mapping.

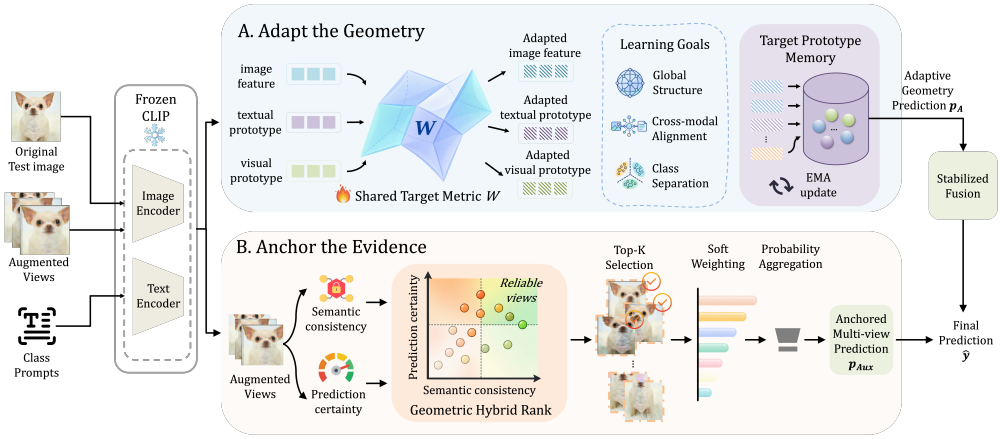

Figure 3: GeoRel operationalizes the two reference-model roles: shared PSD geometry adapts directional comparison, while anchored reliability ranks view influence. EMA separately stabilizes evidence for the primary prediction; selected views supply the auxiliary prediction, and decision-level fusion produces the final prediction.

3 METHOD

GeoRel realizes the comparison and observation-weighting roles in Eq. (2) through discriminative geometry and ordinal reliability, respectively. We retain ReTA's residual-prompt and positive-cache states and update rules (Liang et al., 2025); Appendix C gives their definitions. Let $\mathbf{x}_0$ denote the normalized frozen CLIP feature of the unaugmented query and $\{\mathbf{x}_i\}_{i\in\mathcal{I}}$ its N augmented candidates. With C classes and P prompts per class, $\widetilde{\mathbf{t}}_{c,p}$ is the residual-corrected text feature, $\widehat{\mathbf{v}}_c$ the cache-derived visual prototype (the ℓ_2-normalized mean of cached features for class c; Appendix C.1), and $\ell^{\text{cache}}(\cdot, c)$ the inherited cache logit correction. We write $\text{norm}(\mathbf{a}) = \mathbf{a}/\|\mathbf{a}\|_2$.

3.1 Discriminative Geometry Adaptation

A shared structured metric. We seek a task-discriminative geometry that realizes the comparison role of $\mathbf{M}_t$ using relations observable at test-time. Text semantics provide class identities, while the positive cache supplies pseudo-labeled target visual evidence; together they induce weak pairwise constraints. We parameterize the shared transform as

$$\mathbf{W} = \mathbf{\Delta} + \mathbf{U}\mathbf{U}^\top, \quad \mathbf{\Delta} = \mathrm{Diag}(\mathbf{d}), \quad \mathbf{U} \in \mathbb{R}^{D \times r_g}, \quad r_g \ll D, \tag{4}$$

where $\theta = (\mathbf{d}, \mathbf{U})$. The diagonal reweights individual directions, and the low-rank component captures correlated corrections. The parameterization balances flexibility with the limited evidence available online. A diagonal transform can reweight individual directions but cannot model correlated corrections, whereas a dense transform introduces quadratically many degrees of freedom. The diagonal-plus-low-rank form retains direct per-direction control while allocating a compact subspace to correlated changes, requiring only $O(D+Dr_g)$ parameters. This structure is therefore well matched to batch-size-one adaptation, where the target stream provides weak but continually accumulating class evidence. For any feature $\mathbf{q}$,

$$\phi_{\mathbf{W}}(\mathbf{q}) = \frac{\mathbf{W}\mathbf{q}}{\|\mathbf{W}\mathbf{q}\|_2}. \tag{5}$$

We apply this map to images, residual-corrected prompts, and cache-derived visual prototypes:

$$\mathbf{z}_i^{\mathrm{img}} = \phi_{\mathbf{W}}(\mathbf{x}_i), \quad \mathbf{z}_{c,p}^{\mathrm{txt}} = \phi_{\mathbf{W}}(\widetilde{\mathbf{t}}_{c,p}), \quad \mathbf{z}_c^{\mathrm{vis}} = \phi_{\mathbf{W}}(\widehat{\mathbf{v}}_c), \tag{6}$$

where $i \in \{0\} \cup \mathcal{I}$. All representations participate in the same comparisons; a shared transform preserves one common decision geometry, while separate modality transforms could improve individual representations but make their cross-modal relation inconsistent. Section 4.3 compares shared and modality-specific transforms at matched capacity. The class text prototype is

$$\mathbf{z}_c^{\mathrm{txt}} = \mathrm{norm}\left(\frac{1}{P}\sum_{p=1}^{P} \mathbf{z}_{c,p}^{\mathrm{txt}}\right). \tag{7}$$

The induced PSD metric $\mathbf{M}_\theta = \mathbf{W}^\top\mathbf{W} \succeq 0$ defines a generalized cosine:

$$\phi_{\mathbf{W}}(\mathbf{a})^\top \phi_{\mathbf{W}}(\mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{M}_\theta \mathbf{b}}{\sqrt{\mathbf{a}^\top \mathbf{M}_\theta \mathbf{a}}\sqrt{\mathbf{b}^\top \mathbf{M}_\theta \mathbf{b}}}. \tag{8}$$

The normalized class prototypes have unit norm in the transformed decision space, yielding the normalized comparison form used by the classifier. The generalized-cosine denominators remove variation in the raw $\mathbf{M}_\theta$-norm. Appendix A.2 details the metric properties and efficient evaluation.

Target-side discriminant metric learning. Following large-margin metric learning (Weinberger & Saul, 2009), we use matched text–visual classes and locally confusable text classes to constrain the geometry. For unit transformed prototypes, write $d_\theta^2(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2^2 = 2 - 2\mathbf{a}^\top\mathbf{b}$; dependence on θ is through the prototypes. The discriminant energy is

$$\begin{aligned} \mathcal{E}_{\mathrm{disc}} = &\frac{\lambda_a}{2|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t} d_\theta^2(\mathbf{z}_c^{\mathrm{txt}}, \mathbf{z}_c^{\mathrm{vis}}) \\ &+ \frac{\lambda_m}{Ck}\sum_{c=1}^{C}\sum_{c'\in\mathcal{N}_k(c)}\left[m - \frac{1}{2}d_\theta^2(\mathbf{z}_c^{\mathrm{txt}}, \mathbf{z}_{c'}^{\mathrm{txt}})\right]_+ . \end{aligned} \tag{9}$$

Here $\mathcal{C}_t$ contains cached classes, and $\mathcal{N}_k(c)$ contains the k most similar competing text prototypes. Attraction injects target visual evidence, local margin preserves class discrimination, and the two transfer the supervised pull–push principle to weak cross-modal evidence available online.

Redundancy conditioning. We additionally penalize correlated directions in the transformed class-text matrix $\mathbf{Z}^{\mathrm{txt}} \in \mathbb{R}^{C\times D}$:

$$\mathcal{R}_{\mathrm{red}} = \left\|\mathrm{offdiag}(\mathrm{Cov}(\mathbf{Z}^{\mathrm{txt}}))\right\|_F^2. \tag{10}$$

This follows the redundancy-reduction principle of DeCov and Barlow Twins (Cogswell et al., 2016; Zbontar et al., 2021). The online objective is

$$\mathcal{L}_{\mathrm{geom}} = \mathcal{E}_{\mathrm{disc}} + \lambda_d\mathcal{R}_{\mathrm{red}}, \qquad \mathcal{L}_{\mathrm{online}} = \mathcal{L}_{\mathrm{ReTA}} + \mathcal{L}_{\mathrm{geom}}. \tag{11}$$

Redundancy conditioning prevents the transform from relying on correlated prototype directions, while attraction and separation determine class-discriminative structure. The corresponding cosine-form implementation is listed in Appendix C.3.

3.2 ANCHORED RELIABILITY RANKING

Why an anchored pair of cues? Predictive certainty and query consistency answer different reliability questions. Certainty measures how decisively the evolving classifier interprets a view, whereas frozen-space consistency measures whether that view preserves evidence from the original query. Using both prevents view selection from depending entirely on the same adaptive classifier that ultimately consumes the selected observations. This creates a simple division of labor: certainty tracks current decision sharpness, while consistency anchors semantics independently of adaptation.

Consistency and certainty. For $i \in \mathcal{I}$,

$$q_i = \mathbf{x}_i^\top \mathbf{x}_0 \tag{12}$$

measures cosine consistency with the unaugmented query in frozen CLIP space. Using the untransformed residual-corrected text prototype

$$\bar{\mathbf{t}}_c = \mathrm{norm}\left(\frac{1}{P}\sum_{p=1}^{P}\widetilde{\mathbf{t}}_{c,p}\right), \tag{13}$$

the auxiliary classifier and entropy are

$$\mathbf{p}_i^{\mathrm{aux}} = \mathrm{softmax}_c\left[s\mathbf{x}_i^\top \bar{\mathbf{t}}_c + \ell^{\mathrm{cache}}(\mathbf{z}_i^{\mathrm{img}}, c)\right], \quad h_i = -\sum_{c=1}^{C} p_{i,c}^{\mathrm{aux}} \log p_{i,c}^{\mathrm{aux}}, \tag{14}$$

where s is the CLIP logit scale. Consistency and certainty fail in different ways: certainty can remain high on semantically poor augmentations, while frozen-space consistency supplies an adaptation-independent reference. Neither signal alone is treated as calibrated precision; ranking keeps only within-instance ordinal information, avoiding comparison across incompatible raw scales.

Scale-free relative evidence. Cosine consistency and entropy have different numerical scales. We map each signal to a within-instance normalized rank:

$$\mathcal{R}(b_i) = \frac{\mathrm{rank}_{\mathrm{asc}}(b_i) - 1}{\max\{N-1, 1\}}, \qquad r_i^Q = \mathcal{R}(q_i), \quad r_i^H = \mathcal{R}(-h_i). \tag{15}$$

These are empirical percentile positions. Because consistency and entropy have different and potentially drifting scales, GeoRel retains only their within-instance ordering. This gives a scale-free ordinal reliability representation; strictly increasing transformations of either signal preserve the ranks and resulting selection (Appendix A.3).

Agreement-sensitive fusion. With $\alpha = w_Q/(w_Q + w_H)$, $\beta = w_H/(w_Q + w_H)$, and $w_Q, w_H > 0$, the agreement-sensitive reliability score is

$$u_i = (r_i^Q + \epsilon)^\alpha (r_i^H + \epsilon)^\beta. \tag{16}$$

The geometric combination acts as a soft conjunction: a view receives a high score only when both ordinal cues provide support. For equal weights, setting $a = r_i^Q + \epsilon$ and $b = r_i^H + \epsilon$ gives

$$\frac{a+b}{2} - u_i = \frac{a+b}{2} - \sqrt{ab} = \frac{1}{2}(\sqrt{a} - \sqrt{b})^2 \geq 0. \tag{17}$$

At a fixed arithmetic mean, greater disagreement lowers the geometric score. Accordingly, u_i is an agreement-sensitive ordinal score, not a numerical estimate of γ_i. Section 4.4 tests whether this algebraic disagreement penalty matches empirical view risk.

3.3 STABILIZED PREDICTION

Cache-derived visual prototypes inherit noise from classifier-assigned pseudo-labels. We therefore use EMA as a separate stabilization mechanism for the target-visual evidence consumed by the primary branch. For each represented class, the EMA update is

$$\mathbf{e}_c^{(t)} = m_{\mathrm{ema}}\mathbf{e}_c^{(t-1)} + (1 - m_{\mathrm{ema}})\mathbf{z}_c^{\mathrm{vis},(t)}. \tag{18}$$

Under an idealized independent stationary-noise model with a fixed mean, its steady-state covariance is reduced by a factor $\frac{1-m_{\text{ema}}}{1+m_{\text{ema}}}$. Appendix A.4 gives the corresponding expression for temporally correlated observations. The class representative is

$$\mathbf{g}_c^{(t)} = \text{norm}\Big(\tau \mathbf{z}_c^{\text{txt}} + (1-\tau)\mathbf{e}_c^{(t)}\Big), \tag{19}$$

with text-only fallback before visual memory is available. Primary uses the clean query under adapted geometry; Auxiliary uses reliability-filtered augmented evidence. The two branches are intentionally asymmetric. Primary preserves the clean query and benefits from geometry adaptation and stabilized class evidence; Auxiliary exploits augmentation diversity only after reliability filtering. Their fusion therefore combines two different evidence sources rather than averaging replicas of the same predictor. The Primary prediction from the unaugmented query is

$$\mathbf{p}_A = \text{softmax}_c\left[s(\mathbf{z}_0^{\text{img}})^\top \mathbf{g}_c^{(t)} + \ell^{\text{cache}}(\mathbf{z}_0^{\text{img}}, c)\right]. \tag{20}$$

From reference weighting to decision-level aggregation. Let $\mathcal{S}_K = \text{TopK}_{i\in\mathcal{I}}(u_i, K)$ contain the K augmented views with the highest reliability scores, with $K = 5$ in the main experiments. The Auxiliary prediction is the reliability-weighted mixture of normalized per-view class probabilities:

$$a_i = \frac{\exp(u_i/\tau_r)}{\sum_{j\in\mathcal{S}_K}\exp(u_j/\tau_r)}, \qquad \mathbf{p}_{\text{Aux}} = \sum_{i\in\mathcal{S}_K} a_i \mathbf{p}_i^{\text{aux}}. \tag{21}$$

Thus, a_i plays a role analogous to the normalized relative influence γ_i/Γ in the reference rule, preserving the structural principle that relatively more reliable observations exert greater influence. Convex fusion gives the Final prediction:

$$\mathbf{p}_{\text{final}} = \lambda_{\text{mix}}\mathbf{p}_A + (1-\lambda_{\text{mix}})\mathbf{p}_{\text{Aux}}, \qquad \widehat{y} = \arg\max_c p_{\text{final},c}. \tag{22}$$

Branch fusion is an operational design choice; it allows geometry-adapted clean-query evidence and reliability-filtered multi-view evidence to remain complementary. Appendix C specifies initialization and the online update sequence.

4 Experiments

4.1 Experimental Setup

Benchmarks. Following standard VLM-TTA protocols (Shu et al., 2022; Karmanov et al., 2024; Zhang et al., 2024a; Liang et al., 2025), we evaluate ImageNet and its V2, A, R, and Sketch shifts, plus Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVC-Aircraft, SUN397, DTD, EuroSAT, and UCF101. We report top-1 accuracy and benchmark averages.

Protocol. We evaluate frozen CLIP-RN50 and ViT-B/16 encoders at batch size one, with one clean query and 63 augmented candidates per image. ImageNet variants use random resized crops; other datasets use AugMix (Hendrycks et al., 2020) (Appendix B). Each sample updates inherited ReTA parameters and $\mathbf{W}$ once; results average three seeds. Primary, Auxiliary (Aux.), and Final denote $\mathbf{p}_A$, $\mathbf{p}_{\text{Aux}}$, and $\mathbf{p}_{\text{final}}$ in Eqs. (20)–(22).

Baselines. We compare against zero-shot CLIP, prompt ensemble, CoOp, TPT, DiffTPT, TDA, ZERO, BoostAdapter, DPE, TPS, DMN-ZS, and ReTA (Zhou et al., 2022; Shu et al., 2022; Feng et al., 2023; Karmanov et al., 2024; Farina et al., 2024; Zhang et al., 2024a;b;c; Sui et al., 2025; Liang et al., 2025). Tables 1–4 retain ReTA's RN50 and ViT-B/16 baseline blocks and add GeoRel under the same protocol.

4.2 Main Results

Tables 1–4 report the standard natural-shift and cross-dataset evaluations. Relative to ReTA, GeoRel improves the overall/OOD natural-shift averages by $0.93/0.98$ pp with RN50 and $0.97/1.02$ pp with ViT-B/16, and improves the ten-dataset cross-dataset average by 1.31 and 0.83 pp, respectively. Across the 30 dataset–backbone evaluations, GeoRel achieves the highest accuracy in 24; the six exceptions are ImageNet-A/RN50, Food101/RN50, EuroSAT/RN50, Oxford Pets/ViT-B/16, Food101/ViT-B/16, and EuroSAT/ViT-B/16, where BoostAdapter, DiffTPT, or ReTA retain advantages. Notable gains appear on DTD, UCF101, StanfordCars, and ImageNet-R, while the remaining exceptions show that the benefit is not uniform. We therefore use the following analyses to test whether the intended geometry and reliability mechanisms explain these aggregate gains.

Table 1: Natural-shift top-1 accuracy (%) with CLIP-RN50. The best competing baseline in each column is **bold**. OOD Avg. excludes IN.

Method	IN	IN-V2	IN-A	IN-R	IN-Sketch	Avg.	OOD Avg.
CLIP	58.16	51.41	21.83	56.15	33.37	44.18	40.69
Ensemble	59.81	52.91	23.24	60.72	35.48	46.43	43.09
CoOp	63.33	55.40	23.06	56.60	34.67	46.61	42.43
TPT	60.74	54.70	26.67	59.11	35.09	47.26	43.89
DiffTPT	60.80	55.80	31.06	58.80	37.10	48.71	45.69
TDA	61.35	55.54	30.29	62.58	38.12	49.58	46.63
TPS	61.47	54.96	30.48	62.87	37.14	49.38	46.36
DMN-ZS	63.87	56.12	28.57	61.44	39.84	49.97	46.49
BoostAdapter	61.04	56.22	**35.52**	62.87	38.87	50.91	48.37
DPE	63.41	56.72	30.15	63.72	40.03	50.81	47.66
ReTA	63.90	56.34	34.42	64.10	40.54	51.86	48.85
GeoRel	**64.63**	**57.89**	35.44	**64.50**	**41.49**	**52.79**	**49.83**

Table 2: Natural-shift top-1 accuracy (%) with CLIP-ViT-B/16. The best competing baseline in each column is **bold**. OOD Avg. excludes IN.

Method	IN	IN-V2	IN-A	IN-R	IN-Sketch	Avg.	OOD Avg.
CLIP	66.73	60.86	47.87	73.98	46.09	59.11	57.20
Ensemble	68.34	61.88	49.89	77.65	48.24	61.20	59.42
CoOp	71.51	64.20	49.71	75.21	47.99	61.72	59.28
TPT	68.98	63.45	54.77	77.06	47.94	62.44	60.81
DiffTPT	70.30	65.10	55.68	75.00	46.80	62.28	60.52
TDA	69.51	64.67	60.11	80.24	50.54	65.01	63.89
TPS	70.19	64.73	60.08	80.27	49.95	65.04	63.76
ZERO	70.93	65.16	64.06	80.75	50.32	66.24	65.07
DMN-ZS	72.25	65.17	58.28	78.55	53.20	65.49	63.80
BoostAdapter	69.39	65.19	64.06	80.77	51.51	66.19	65.39
DPE	71.91	65.44	59.63	80.40	52.26	65.93	64.43
ReTA	72.18	65.52	64.22	81.31	53.21	67.29	66.07
GeoRel	**72.96**	**66.12**	**65.51**	**82.84**	**53.89**	**68.26**	**67.09**

Table 3: Cross-dataset top-1 accuracy (%) with CLIP-RN50. The best competing baseline in each column is **bold**. Avg. covers all ten datasets.

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	Avg.
CLIP	85.88	83.57	55.70	61.75	73.97	15.66	58.80	40.37	23.69	58.84	55.82
Ensemble	87.26	82.97	55.89	62.77	74.82	16.11	60.85	40.37	25.79	59.48	56.63
CoOp	86.53	87.00	55.32	61.55	75.59	15.12	58.15	37.29	26.20	59.05	56.18
TPT	87.02	84.49	58.46	62.69	74.88	17.58	61.46	40.84	28.33	60.82	57.66
DiffTPT	86.89	83.40	60.71	63.53	**79.21**	17.60	62.72	40.72	41.04	62.67	59.85
TDA	89.70	86.18	57.78	68.74	77.75	17.61	62.53	43.74	42.11	64.18	61.03
BoostAdapter	88.48	85.75	59.67	68.25	78.78	18.93	62.83	43.85	**44.40**	64.42	61.54
DPE	90.83	85.97	59.26	67.60	77.83	19.80	64.23	50.18	41.67	61.98	61.93
ReTA	90.35	86.90	61.11	70.12	77.46	22.62	65.11	52.46	39.64	66.18	63.20
GeoRel	**91.28**	**88.14**	**62.68**	**71.50**	77.80	**22.98**	**65.74**	**54.26**	43.88	**66.80**	**64.51**

Table 4: Cross-dataset top-1 accuracy (%) with CLIP-ViT-B/16. The best competing baseline in each column is **bold**. Avg. covers all ten datasets.

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	Avg.
CLIP	93.35	88.25	65.48	67.44	83.65	23.67	62.59	44.27	42.01	65.13	63.58
Ensemble	93.55	86.92	66.11	66.99	82.86	23.22	65.63	45.04	50.42	65.16	64.59
CoOp	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
TPT	94.16	87.79	66.87	68.98	84.67	24.78	65.50	47.75	42.44	68.04	65.10
DiffTPT	92.49	88.22	67.01	70.10	**87.23**	25.60	65.74	47.00	43.13	62.67	65.47
TDA	94.24	88.63	67.28	71.42	86.14	23.91	67.62	47.40	58.00	70.66	67.53
ZERO	94.14	87.20	68.48	66.82	84.58	24.42	66.90	45.86	42.09	68.57	64.91
BoostAdapter	94.77	89.51	69.30	71.66	87.17	27.45	68.09	45.69	**61.22**	71.93	68.68
DPE	94.81	91.14	67.31	75.07	86.17	28.95	70.07	54.20	55.79	70.44	69.40
ReTA	95.29	**92.37**	69.11	77.55	86.69	31.86	70.70	57.39	58.26	74.52	71.37
GeoRel	**96.15**	91.99	**70.34**	**77.95**	86.99	**32.19**	**71.43**	**59.57**	58.48	**76.92**	**72.20**

4.3 Evidence for Shared Discriminative Geometry

Sample-level discriminative separation. Eq. (9) encourages matched class evidence to contract while separating local confusers. We test whether this prototype-level objective increases sample-level correct-versus-confuser separation by isolating geometry adaptation and EMA readout on all 1,692 DTD samples with CLIP-ViT-B/16. Identity disables geometry adaptation, while $\tau = 1$ removes EMA readout. For comparison setting $b \in \{\mathrm{ref}, \mathrm{geo}\}$, let $f_b(\mathbf{x}_0)$ and $\mathbf{t}_c^{(b)}$ denote its clean-view image representation and class prototype. We define $m_b(\mathbf{x}_0) = \mathrm{sim}(f_b(\mathbf{x}_0), \mathbf{t}_y^{(b)}) - \max_{c \neq y} \mathrm{sim}(f_b(\mathbf{x}_0), \mathbf{t}_c^{(b)})$ and the comparison-margin shift $\Delta m = m_{\mathrm{geo}} - m_{\mathrm{ref}}$. Positive Δm indicates greater correct-versus-strongest-confuser separation under the learned comparison geometry.

Table 5: Geometry-adaptation and EMA intervention on DTD with CLIP-ViT-B/16. Margin shift Δm: $\times 10^{-3}$. Purity@5 is the auxiliary classifier's correctness rate among the five selected views; CWR@5 is the fraction of those five views that are both misclassified and highly certain ($r_i^H \geq 0.8$).

Condition	Primary Acc. (%)	Aux. Acc. (%)	Final Acc. (%)	Margin shift Δm ($\times 10^{-3}$)	Purity@5 (%)	CWR@5 (%)	ΔFinal Acc. vs. Full (pp)
Full	**58.39**	57.80	**59.57**	+2.11	**58.76**	30.47	–
No EMA read ($\tau = 1$)	58.27	**58.51**	58.92	**+2.30**	58.53	**30.34**	-0.65
No geometry adaptation (identity)	56.86	58.27	58.10	0.00	56.65	31.88	-1.47
No geometry adaptation, no EMA read	57.17	57.57	57.68	0.00	54.34	34.16	-1.89

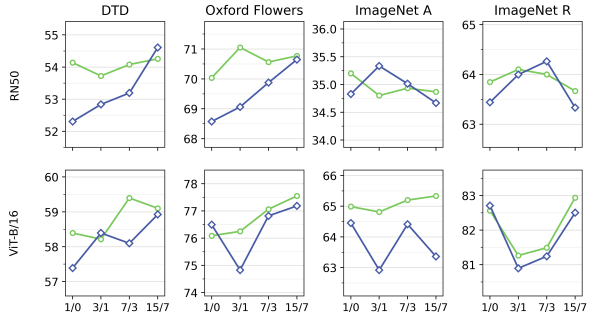

Figure 4: Parameter-matched geometry transforms. Green circles: Full-shared; blue diamonds: Separated. Horizontal: Shared rank / Separate rank at equal adapter parameter counts. Vertical: Final fused top-1 accuracy (%).

On DTD, enabling geometry adaptation (Table 5) produces a $+2.11 \times 10^{-3}$ comparison-margin shift and improves Primary/Final accuracy by 1.53/1.47 pp. Removing EMA readout reduces Final accuracy by 0.65 pp.

Cross-dataset geometry transfer. Table 6 extends the paired geometry-adaptation on/off intervention to six datasets with samples, views, and configuration fixed. The learned geometry increases Primary accuracy on all six datasets and yields a positive comparison-margin shift on five of six, with mean Primary and Final gains of 0.90 and 0.77 pp. Because the intervention changes only the geometry while holding samples, views, and the remaining configuration fixed, the paired comparison links the accuracy change directly to the altered comparison space.

Table 6: Paired geometry-adaptation on/off comparison across six datasets. Margin shift Δm: $\times 10^{-3}$; other entries: on minus off.

Dataset	ΔPrimary Acc. (pp)	ΔAux. Acc. (pp)	ΔFinal Acc. (pp)	Margin shift Δm ($\times 10^{-3}$)	ΔPurity@5 (pp)	ΔCWR@5 (pp)
DTD	**+1.53**	-0.47	**+1.47**	+2.11	**+2.11**	**-1.41**
Oxford Flowers	+0.97	+0.31	+0.68	+1.60	+1.15	+0.78
EuroSAT	+1.18	-0.06	+1.14	+0.30	+0.27	-0.07
Caltech101	+0.46	**+0.34**	+0.41	**+3.20**	+0.16	-0.36
Oxford Pets	+0.21	-0.15	-0.09	-0.50	-0.23	+0.04
UCF101	+1.05	+0.12	+1.03	+2.80	+1.44	-1.02

Shared geometry under matched capacity. Figure 4 tests the shared geometry against modality-specific transforms under matched parameter budgets. The shared geometry wins 25 of 32 dataset–backbone–budget comparisons by 0.57 pp on average and is less sensitive to the tested budgets in all eight panels. Figure 5 visualizes the joint reorganization of image, text, and cache-derived visual prototypes under $\mathbf{W}$.

4.4 Evidence for Anchored View Reliability and Branch Complementarity

Agreement tracks view reliability. Figure 6(a–b) tests this ordering on DTD, Oxford Flowers, and EuroSAT with ViT-B/16. At comparable certainty, consistency tracks view accuracy; after stratifying by mean evidence strength, larger consistency–certainty disagreement predicts lower accuracy. Across six datasets, high-consistency/high-certainty views are most accurate, with Q5–Q1 gaps of 7.2–22.3 pp. Selector replay improves Purity@5 over Random in 24/24 settings and lowers CWR@5 relative to certainty-only in 23/24 (Appendices D–E).

Complementary prediction paths. Figure 6(c–d) shows that Final rescues more errors than it introduces relative to either branch in all six comparisons. On EuroSAT/ViT-B/16, Final changes by

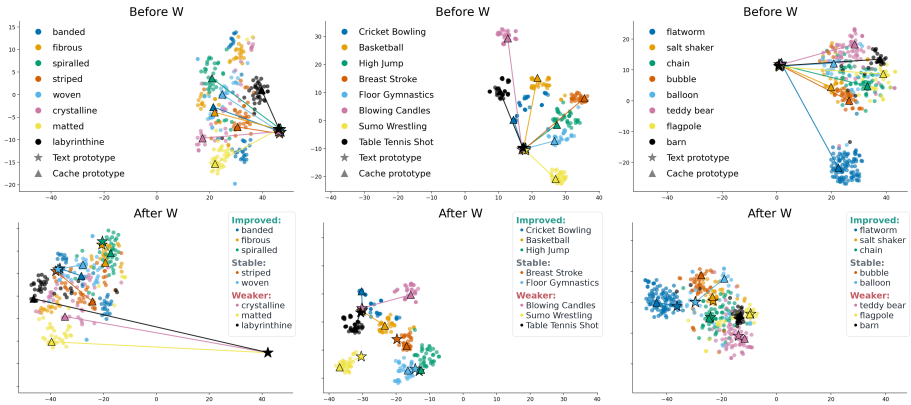

Figure 5: Cross-modal representations before and after shared transform $\mathbf{W}$ on DTD, UCF101, and ImageNet-A. Each dataset jointly embeds image, text, and cache-derived prototypes with one t-SNE fit and fixed axes; markers denote images, text prototypes, and visual prototypes.

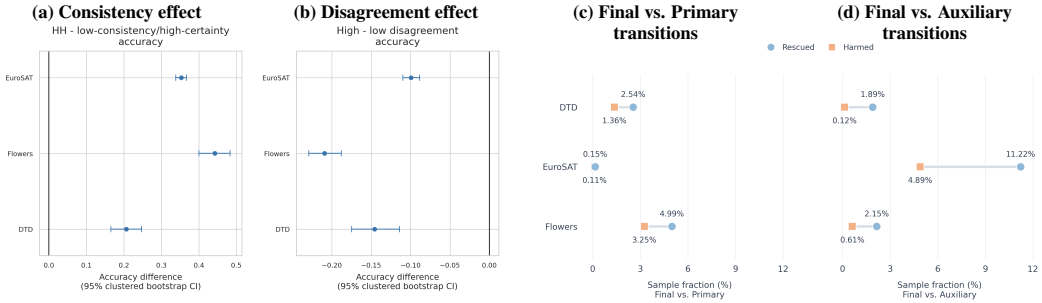

Figure 6: Anchored reliability and branch complementarity. (a) Consistency at high certainty; (b) disagreement effect after stratification; (c–d) prediction transitions. Bars show 95% clustered-bootstrap intervals.

only $+0.04$ pp versus Primary but exceeds Auxiliary by 6.33 pp, showing the value of the geometry-adapted query when multi-view evidence is weak. Extended analysis appears in Appendix E.3.

Together, these studies support the factorized design: geometry adaptation improves clean-query separation, while anchored agreement concentrates cleaner augmented evidence without relying on calibrated confidence. Their complementary error patterns support keeping the two prediction paths distinct.

5 Limitations and Conclusion

Limitations. The reference model is structural rather than a literal description of normalized CLIP features: augmented views are correlated, and the learned geometry and reliability scores are operational surrogates rather than estimators of latent reference-model quantities. GeoRel depends on pseudo-labeled target evidence; incorrect cache assignments can distort the learned geometry, while EMA may preserve both useful and erroneous evidence. Anchored ranking requires frozen-space consistency and adaptive certainty to remain at least partially informative; if both signals collapse under severe semantic shift, ranking cannot recover reliable views. Finally, our evaluation focuses on CLIP image classification with RN50 and ViT-B/16. Although reduced-view experiments suggest that substantially fewer augmentations can preserve performance, broader VLM architectures and tasks remain outside the present scope.

Conclusion. Reliable test-time adaptation requires answering two different questions: where should class evidence be compared, and which observations should be trusted? GeoRel addresses the first through shared discriminative comparison geometry and the second through anchored, scale-free observation reliability. Across natural-shift and cross-dataset evaluations, the empirical analyses support both mechanisms and show that their prediction paths provide complementary evidence. The resulting principle is to adapt comparison geometry while anchoring observation reliability.

AI use statement. Generative AI tools were used to assist with language editing, manuscript organization, LaTeX conversion, and consistency checks during manuscript preparation. They were

not used to generate experimental results, scientific claims, or conclusions. All AI-assisted content, including citations and technical statements, was reviewed and verified by the authors, who take full responsibility for the final manuscript.

Ethics statement. This work studies test-time adaptation for image classification using existing vision–language models and standard benchmarks and does not introduce new human-subject data. As with other methods built on pretrained models and pseudo-label-based adaptation, the proposed approach may inherit biases and failure modes from the underlying models and data. Performance on standard benchmarks should not be interpreted as evidence of suitability for high-stakes deployment.

Reproducibility statement. Sections 2–3 specify the assumptions, parameterization, objectives, ranking rule, aggregation rule, and prediction procedure. Appendix A provides proofs, and Appendix C records the online sequence.

REFERENCES

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 (ICCV)*, pp. 2281–2291, 2025.

Jungwon Choi and Eunwoo Kim. Dual-modality anchor-guided filtering for test-time prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9368–9377, 2026.

Michael Cogswell, Faruk Ahmed, Ross B. Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. In *International Conference on Learning Representations (ICLR)*, 2016.

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, 2024.

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

Gustavo A. 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 *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 7092–7101, 2025.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012. doi: 10.1561/2200000019.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33999–34019, 2025.

Yiwen Liang, Hui Chen, Yizhe Xiong, Zihan Zhou, Mengyao Lyu, Zijia Lin, Shuaicheng Niu, Sicheng Zhao, Jungong Han, and Guiguang Ding. Advancing reliable test-time adaptation of vision-language models under visual variations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4788–4797, 2025. doi: 10.1145/3746027.3755607.

Sarthak Maharana, Baoming Zhang, Leonid Karlinsky, Rogerio Feris, and Yunhui Guo. BATCLIP: Bimodal online test-time adaptation for CLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1569–1579, 2025.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, and Muhammad Haris Khan. O-TPT: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19942–19951, 2025.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Fei Song, Yi Li, Rui Wang, Jiahuan Zhou, Changwen Zheng, and Jiangmeng Li. Doubly debiased test-time prompt tuning for vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9069–9078, 2026. doi: 10.1609/aaai.v40i11.37863.

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

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pp. 443–450, 2016.

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

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

Min Wang, Miao Jia, Hao Yang, Qing Cheng, and Jincai Huang. CATE: Consensus-aware calibration for test-time prompt tuning via energy anchoring. *Knowledge-Based Systems*, 338:115517, 2026. doi: 10.1016/j.knosys.2026.115517.

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

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3143–3153, 2026.

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

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

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021.

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

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*, volume 37, pp. 67795–67825, 2024b.

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

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

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

A PROOFS

A.1 PROOF OF PROPOSITION 1

Under Assumption 1, the MAP classifier is

$$\widehat{y} = \arg\max_c \left[\boldsymbol{\mu}_c^\top \mathbf{M}_t \sum_{i \in \mathcal{I}} \gamma_i \mathbf{x}_i - \frac{\Gamma}{2} \boldsymbol{\mu}_c^\top \mathbf{M}_t \boldsymbol{\mu}_c \right]. \tag{23}$$

The view index set $\mathcal{I}$ is nonempty, and each $\gamma_i > 0$ is fixed across class hypotheses, so $\Gamma > 0$. For view $i \in \mathcal{I}$, Assumption 1 gives

$$p(\mathbf{x}_i \mid y = c) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_t / \gamma_i|^{1/2}} \exp\Big[-\frac{\gamma_i}{2} (\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_t^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_c) \Big]. \tag{24}$$

Here $|\boldsymbol{\Sigma}_t / \gamma_i|$ denotes the covariance determinant. Conditional independence makes the joint likelihood the product over i. After taking negative logarithms and dropping terms independent of class c, the classification objective is

$$\mathcal{J}(c) = \frac{1}{2} \sum_{i \in \mathcal{I}} \gamma_i (\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \mathbf{M}_t (\mathbf{x}_i - \boldsymbol{\mu}_c). \tag{25}$$

Expanding the quadratic form yields

$$2\mathcal{J}(c) = \sum_{i \in \mathcal{I}} \gamma_i \mathbf{x}_i^\top \mathbf{M}_t \mathbf{x}_i - 2\boldsymbol{\mu}_c^\top \mathbf{M}_t \sum_{i \in \mathcal{I}} \gamma_i \mathbf{x}_i \tag{26}$$

$$+ \Gamma \boldsymbol{\mu}_c^\top \mathbf{M}_t \boldsymbol{\mu}_c. \tag{27}$$

The first term is independent of c. Uniform priors therefore give the MAP rule in Eq. (23). The metric norm is $\|\boldsymbol{\mu}_c\|_{\mathbf{M}_t} = (\boldsymbol{\mu}_c^\top \mathbf{M}_t \boldsymbol{\mu}_c)^{1/2}$. If these class-center norms are equal, the final term is also class-independent; division by the positive scalar Γ yields Eq. (2). □

Interpretation and scope. Proposition 1 is a decomposition of the Bayes rule under Assumption 1. Its conclusion is structural: $\mathbf{M}_t$ controls directional comparison, whereas $\{\gamma_i\}$ control relative observation influence. The proposition does not provide a finite-sample estimator of either quantity and does not imply that empirical covariance inversion is appropriate for VLM-TTA. In particular, no consistency claim is made between the learned metric $\mathbf{M}_\theta$ and the population precision $\mathbf{M}_t$.

The equal-$\mathbf{M}_t$-norm condition is used only to remove the class-dependent quadratic term from the reference MAP rule and expose the bilinear factorization in Eq. (2). Without that condition, Eq. (23) contains the additional class-dependent term $-\frac{\Gamma}{2}\boldsymbol{\mu}_c^\top \mathbf{M}_t \boldsymbol{\mu}_c$. GeoRel's normalized generalized-cosine classifier should therefore be viewed as an operational comparison rule motivated by the factorized role of $\mathbf{M}_t$, rather than as a literal implementation of the full Gaussian discriminant.

A.2 Properties of the learned PSD comparison metric

The shared transform compares image, text, and cache-derived visual representations under one learned geometry. The following properties establish that this operational surrogate defines a valid PSD generalized-cosine geometry and has controlled structured capacity. They do not identify $\mathbf{M}_\theta$ with the population precision $\mathbf{M}_t$ of the reference model. The parameterization requires $O(D + Dr_g)$ parameters and applies $\mathbf{Wq} = \boldsymbol{\Delta}\mathbf{q} + \mathbf{U}(\mathbf{U}^\top \mathbf{q})$ without forming a dense matrix.

PSD generalized cosine. For any $\mathbf{a} \in \mathbb{R}^D$,

$$\mathbf{a}^\top \mathbf{M}_\theta \mathbf{a} = \mathbf{a}^\top \mathbf{W}^\top \mathbf{W}\mathbf{a} = \|\mathbf{W}\mathbf{a}\|_2^2 \geq 0, \tag{28}$$

so $\mathbf{M}_\theta \succeq 0$. Substituting $\mathbf{M}_\theta = \mathbf{W}^\top \mathbf{W}$ into the dot product of two normalized transformed features gives Eq. (8).

Rank of the correlated correction. The structured transform satisfies

$$\operatorname{rank}(\mathbf{M}_\theta - \boldsymbol{\Delta}^2) \leq 2r_g. \tag{29}$$

Because $\boldsymbol{\Delta}$ and $\mathbf{U}\mathbf{U}^\top$ are symmetric,

$$\mathbf{M}_\theta = (\boldsymbol{\Delta} + \mathbf{U}\mathbf{U}^\top)^2 \tag{30}$$

$$= \boldsymbol{\Delta}^2 + \boldsymbol{\Delta}\mathbf{U}\mathbf{U}^\top + \mathbf{U}\mathbf{U}^\top\boldsymbol{\Delta} + \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top. \tag{31}$$

The column space of the correction lies in $\operatorname{span}(\mathbf{U}, \boldsymbol{\Delta}\mathbf{U})$, whose dimension is at most $2r_g$. This proves Eq. (29). □

Margin and Euclidean separation. For normalized transformed text prototypes, $\|\mathbf{z}_c^{\mathrm{txt}} - \mathbf{z}_{c'}^{\mathrm{txt}}\|_2^2 = 2 - 2(\mathbf{z}_c^{\mathrm{txt}})^\top \mathbf{z}_{c'}^{\mathrm{txt}}$. Thus satisfying the margin in Eq. (42) implies $\|\mathbf{z}_c^{\mathrm{txt}} - \mathbf{z}_{c'}^{\mathrm{txt}}\|_2^2 \geq 2m$.

A.3 Ordinal Reliability, Monotone-Transform Invariance, and Geometric Disagreement

For reliability ranking, q_i references the unaugmented query in frozen CLIP space, whereas $-h_i$ reflects the evolving online classifier. GeoRel treats these signals as observable ordinal cues rather than calibrated estimators of latent view precision. Rank normalization removes sensitivity to their raw scales, while geometric fusion favors views supported by both cues.

Proposition 2 (Monotone-transform invariance)**.** *Let f be strictly increasing, and fix score-independent rules for breaking ties in ranking and Top-K selection. Replacing every q_i by $f(q_i)$, or every $-h_i$ by $f(-h_i)$, leaves the corresponding vector of normalized rank scores unchanged. Consequently, the reliability scores and Top-K set remain unchanged.*

Proof. For any i, j, $q_i < q_j$ if and only if $f(q_i) < f(q_j)$; equality is also preserved. Hence all pairwise order and tie relations are unchanged. Applying the same deterministic tie-breaking rule gives the same rank scores. The argument for $-h_i$ is identical, and Eq. (16) then gives the same reliability scores. □

For fixed τ_r, the aggregation weights are therefore unchanged, as is the aggregated prediction whenever the auxiliary probability vectors are held fixed.

Corollary 1 (Ordinal precision preservation under a monotone proxy)**.** *Let $\{\gamma_i\}_{i\in\mathcal{I}}$ denote the latent observation precisions in Assumption 1, and let an observable score satisfy $s_i = f(\gamma_i)$ for a strictly increasing function f. Under the same score-independent tie-breaking rule used by GeoRel,*

$$\mathcal{R}(s_i) = \mathcal{R}(\gamma_i) \qquad \textit{for all } i \in \mathcal{I}.$$

Consequently, any Top-K selection based only on the normalized ranks of $\{s_i\}$ is identical to the corresponding Top-K selection based on the latent precisions.

Proof. Strict monotonicity preserves all pairwise order and equality relations, so the claim follows directly from Proposition 2. □

Scope of the ordinal interpretation. The corollary is conditional and does not assume that either q_i or $-h_i$ is exactly a monotone function of the latent γ_i. Rather, it explains why rank-based reliability is appropriate when an observable signal is informative primarily through ordering rather than calibrated scale. GeoRel uses frozen-space consistency and adaptive certainty as complementary noisy ordinal cues and favors their agreement. Whether this agreement is empirically associated with lower view risk is a separate empirical question examined in Section 4.4 and Appendices D–E.

Proposition 3 (Geometric disagreement penalty)**.** *Write $\widetilde{r}_i^Q = r_i^Q + \epsilon$ and $\widetilde{r}_i^H = r_i^H + \epsilon$. With equal weights, let $A_i = (\widetilde{r}_i^Q + \widetilde{r}_i^H)/2$ and $G_i = \sqrt{\widetilde{r}_i^Q \widetilde{r}_i^H} = u_i$. Then*

$$A_i - G_i = \frac{1}{2}\left(\sqrt{\widetilde{r}_i^Q} - \sqrt{\widetilde{r}_i^H}\right)^2 \geq 0.$$

For equal signal weights, Eq. (17) *holds. More generally, if $\alpha, \beta > 0$ and $\alpha + \beta = 1$, then*

$$(\widetilde{r}_i^Q)^\alpha (\widetilde{r}_i^H)^\beta \leq \alpha \widetilde{r}_i^Q + \beta \widetilde{r}_i^H. \tag{32}$$

Proof. For equal weights, expand

$$\frac{a+b}{2} - \sqrt{ab} = \frac{(\sqrt{a} - \sqrt{b})^2}{2} \tag{33}$$

with $a = \widetilde{r}_i^Q$ and $b = \widetilde{r}_i^H$. The general statement follows from the weighted arithmetic–geometric mean inequality. □

Proposition 3 is an algebraic property of the operational reliability score: at fixed mean rank evidence, stronger disagreement lowers the geometric combination. It does not by itself establish that disagreement is a statistical estimator of latent observation variance. The connection to actual augmented-view error is therefore evaluated empirically rather than assumed by the reference model.

A.4 EMA VARIANCE UNDER STATIONARY OBSERVATIONS

EMA is analyzed separately from the reference-model factorization. Its purpose is to stabilize noisy pseudo-labeled target-visual prototypes used by the primary branch; it is not an estimator of either $\mathbf{M}_t$ or γ_i and is not implied by Proposition 1.

Lemma 1 (EMA covariance with temporal correlation)**.** *Let t index successive memory updates for one class, omitting steps at which its memory is held fixed. Suppose $\mathbf{z}_c^{\mathrm{vis},(t)} = \boldsymbol{\nu}_c + \boldsymbol{\xi}_t$, where $\boldsymbol{\nu}_c$ is fixed and $\boldsymbol{\xi}_t$ is a zero-mean, second-order stationary process with finite covariance. Define $\boldsymbol{\Gamma}_k = \mathrm{Cov}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-k})$. For $m = m_{\mathrm{ema}} \in [0, 1)$, after initialization transients vanish,*

$$\mathrm{Cov}(\mathbf{e}_c) = \frac{1-m}{1+m}\left[\boldsymbol{\Gamma}_0 + \sum_{k=1}^{\infty} m^k (\boldsymbol{\Gamma}_k + \boldsymbol{\Gamma}_k^\top)\right]. \tag{34}$$

For independent noise with covariance $\boldsymbol{\Sigma}_v$, this reduces to $\mathrm{Cov}(\mathbf{e}_c) = \frac{1-m}{1+m}\boldsymbol{\Sigma}_v$.

Proof. The stationary EMA has the mean-square convergent representation $\mathbf{e}_c^{(t)} = \boldsymbol{\nu}_c + (1-m)\sum_{j=0}^{\infty} m^j \boldsymbol{\xi}_{t-j}$. Expanding its covariance and grouping terms by temporal lag gives

$$\mathrm{Cov}(\mathbf{e}_c) = (1-m)^2 \left[\sum_{j=0}^{\infty} m^{2j} \boldsymbol{\Gamma}_0 + \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} m^{2j+k} (\boldsymbol{\Gamma}_k + \boldsymbol{\Gamma}_k^\top) \right] \tag{35}$$

$$= \frac{1-m}{1+m} \left[\boldsymbol{\Gamma}_0 + \sum_{k=1}^{\infty} m^k (\boldsymbol{\Gamma}_k + \boldsymbol{\Gamma}_k^\top) \right]. \tag{36}$$

Stationarity bounds the lag covariances, and $m < 1$ ensures convergence. Independence sets $\boldsymbol{\Gamma}_k = 0$ for $k \geq 1$, yielding the stated special case. □

Thus, under the stated stationary-process assumptions, the lemma quantifies the variance-reduction behavior of the memory mechanism. This result supports the stabilization role of EMA independently of the Gaussian reference model in Assumption 1.

B Inherited ReTA Objective

The inherited residual objective is

$$\mathcal{L}_{\mathrm{ReTA}} = \mathcal{L}_{\mathrm{ent}}^R + \lambda_1^R \mathcal{L}_{\mathrm{surr}}^R + \lambda_2^R \mathcal{L}_{\mathrm{align}}^R, \tag{37}$$

Here $\mathcal{L}_{\mathrm{ent}}^R$, $\mathcal{L}_{\mathrm{surr}}^R$, and $\mathcal{L}_{\mathrm{align}}^R$ denote the inherited entropy, distribution-surrogate, and bidirectional cross-modal alignment losses, respectively, following ReTA (Liang et al., 2025). The coefficients λ_1^R and λ_2^R weight the surrogate and alignment losses. The superscript R distinguishes inherited ReTA quantities, including its alignment loss, from the prototype-level objective $\mathcal{L}_{\text{proto-align}}$ in Eq. (41).

Inherited evaluation settings. We use a positive-cache capacity of three samples per class, a normalized-entropy threshold of 0.1, three adjacent text embeddings, 64 retained SVD components, and residual-loss weights $\lambda_1^R = 0.3$ and $\lambda_2^R = 0.02$. The AugMix candidates use random resized cropping and random horizontal flipping. In Tables 1–2, Avg. is the mean over all five natural-shift datasets and OOD Avg. is the mean over the four shifted ImageNet variants. In Tables 3–4, Avg. is the mean over the ten cross-dataset benchmarks.

Positive-cache logits. Both prediction branches use the scalar correction $\ell^{\mathrm{cache}}(\mathbf{z}_i^{\mathrm{img}}, c)$ in Eqs. (20) and (14). This cache logit correction is distinct from the cache-derived class mean $\widehat{\mathbf{v}}_c$ and its EMA memory $\mathbf{e}_c^{(t)}$.

C Online Procedure and Implementation Details

C.1 Inherited Online State

Following ReTA (Liang et al., 2025), each frozen text feature $\mathbf{t}_{c,p}$ has an online residual $\boldsymbol{\delta}_{c,p}$:

$$\widetilde{\mathbf{t}}_{c,p} = \mathrm{norm}(\mathbf{t}_{c,p} + \boldsymbol{\delta}_{c,p}). \tag{38}$$

ReTA's positive cache is grouped by pseudo-label. For represented class $c \in \mathcal{C}_t$, its cache-derived visual prototype is

$$\widehat{\mathbf{v}}_c^{(t)} = \mathrm{norm}\left(\frac{1}{|\mathcal{K}_c^{(t)}|} \sum_{\mathbf{v} \in \mathcal{K}_c^{(t)}} \mathbf{v} \right). \tag{39}$$

We denote the inherited positive-cache logit correction by $\ell^{\mathrm{cache}}(\mathbf{z}, c)$. The inherited residual objective is summarized in Appendix B; cache admission, eviction, and residual updates otherwise follow ReTA.

C.2 Update Procedure

Notation and parameter domains. Let $\mathcal{V} = \{0, \ldots, V-1\}$, $V = N + 1 \geq 2$, and $\mathcal{I} = \mathcal{V} \setminus \{0\}$. All view features have unit ℓ_2-norm; normalization and transformed cosine expressions use nonzero denominators. We use $C \geq 2$, $P \geq 1$, $1 \leq k \leq C-1$, and $0 < m \leq 2$, with $[a]_+ = \max\{a, 0\}$. The loss weights $\lambda_d, \lambda_a, \lambda_m$ are nonnegative, $0 \leq m_{\text{ema}} < 1$, $\tau, \lambda_{\text{mix}} \in [0, 1]$, $s, \tau_r, \epsilon > 0$, $w_Q, w_H > 0$, and the retained-view count K is an integer with $1 \leq K \leq N$. All classifier softmax operations are over classes, and entropy uses natural logarithms. Online-step superscripts are suppressed when the current state is understood; subscripts in $\mathbf{\Sigma}_t$ and $\mathbf{M}_t$ denote the target domain.

The cache $\mathcal{K}_c^{(t)}$ contains visual features assigned to class c at online step t, and $\mathcal{C}_t = \{c : |\mathcal{K}_c^{(t)}| > 0\}$ denotes represented classes. In Eq. (40), $\text{Cov}(\mathbf{Z}^{\text{txt}})$ is the centered feature covariance across classes. In Eq. (42), $\mathcal{N}_k(c)$ contains the k other classes with the largest transformed-prototype cosine similarity to class c.

Cache and memory boundaries. Before the first cache entry arrives, $\mathcal{L}_{\text{proto-align}} = 0$. ReTA's consistency-aware cache admission, priority-based eviction, and conditional residual-update rule are retained. Cache admission and eviction manage stored entries separately from the differentiable objective. A class initializes $\mathbf{e}_c^{(t)} = \mathbf{z}_c^{\text{vis},(t)}$ on its first observation and retains its memory when absent from the cache. The EMA need not have unit norm; normalization occurs when forming $\mathbf{g}_c^{(t)}$.

Ranking and aggregation details. Ascending ranks range from 1 to N; ties in ranking and Top-K selection use fixed, score-independent deterministic rules. For $N = 1$, Eq. (15) assigns the sole candidate rank score zero. The selected weights satisfy $a_i > 0$ and $\sum_{i \in \mathcal{S}_K} a_i = 1$. Scores and weights encode relative view reliability motivated by Eq. (2).

The scores and weights are operational relative-reliability quantities rather than numerical estimates of γ_i. Eq. (2) motivates differential observation influence; Eqs. (15)–(16) construct an ordinal reliability score, and Eq. (21) realizes the resulting weighting principle at the level of predictive distributions.

For each test image, the online procedure is:

1. Extract V normalized frozen view features $\{\mathbf{x}_i\}_{i \in \mathcal{V}}$: the unaugmented query $\mathbf{x}_0$ for the primary branch and $N = V - 1$ augmented candidates indexed by $\mathcal{I}$.
2. Construct the residual-corrected prompt features in Eq. (38) and read the current positive cache.
3. Apply the shared structured transform through $\mathbf{Wq} = \mathbf{\Delta q} + \mathbf{U}(\mathbf{U}^\top \mathbf{q})$, without materializing a dense matrix, to the image features, prompt features, and available cache-derived visual prototypes; compute $\mathcal{L}_{\text{geom}}$.
4. Take one online update of $\mathcal{L}_{\text{online}}$; the CLIP encoders remain frozen.
5. Refresh the transformed visual prototypes and update the EMA visual-prototype memory for the cache classes.
6. Form the fused class representatives $\mathbf{g}_c^{(t)}$ and compute the geometry-adapted Primary prediction $\mathbf{p}_A$.
7. Compute frozen-space query consistency q_i, per-view auxiliary probabilities $\mathbf{p}_i^{\text{aux}}$, and adaptive predictive certainty $-h_i$ for every augmented view.
8. Convert both evidence channels to within-instance normalized rank scores, compute reliability scores u_i, select the Top-K views, and aggregate their probability vectors into $\mathbf{p}_{\text{Aux}}$.
9. Fuse $\mathbf{p}_A$ and $\mathbf{p}_{\text{Aux}}$ with Eq. (22) and output $\widehat{y}$.

Reference-model correspondence of the online steps. The online procedure separates several mechanisms that should not be conflated. Steps 3 and 6 operationalize the directional-comparison role of $\mathbf{M}_t$ through the learned shared geometry. Steps 7 and 8 operationalize relative observation reliability through anchored ordinal ranking and decision-level weighting. Step 5 is a separate prototype-stabilization mechanism, and Step 9 is a separate branch-fusion rule. Only the first two role correspondences originate from the structural decomposition in Proposition 1; EMA and branch fusion are additional operational components.

Online state and update order. The prediction equations use the current residual parameters and shared transform, whereas the EMA explicitly retains historical transformed prototypes.

Comparison-margin shift diagnostic. For $\Delta m = m_{\text{geo}} - m_{\text{ref}}$, the reference margin is computed from the normalized clean CLIP image feature and the mean residual-corrected text prototype. The adapted-geometry margin is computed from the geometry-adapted clean image feature and text prototype used by the primary branch. Both use cosine similarity to the ground-truth class against the strongest competing class, without cache-logit correction.

Motivation-diagnostic protocol. For each dataset, we draw 16 labeled examples per class to estimate the diagnostic geometry and evaluate exclusively on the remaining held-out samples. All geometry baselines use the same class-balanced split. Frozen CLIP image and text features are ℓ_2-normalized before metric estimation. The inverse-variance baseline uses reciprocal feature-wise variance, while the covariance baseline uses centered features with Ledoit–Wolf shrinkage before inversion. The label-assisted discriminative diagonal learns a positive diagonal transform by minimizing cross-entropy on the estimation split, with transformed image and text representations renormalized before cosine classification. No held-out evaluation sample is used for metric estimation or optimization.

C.3 Exact Cosine Forms and Internal Objective Equivalence

The following identities establish equivalence among GeoRel's internal discriminative objective forms. They do not establish equivalence to the Gaussian reference-model likelihood or imply statistical recovery of $\mathbf{M}_t$. The discriminant energy and redundancy regularizer in Section 3.1 retain the following exact cosine-form losses. For transformed text prototypes $\mathbf{Z}^{\text{txt}} \in \mathbb{R}^{C\times D}$ with one class per row, we use

$$\mathcal{L}_{\text{decorr}} = \left\|\text{offdiag}\left(\text{Cov}(\mathbf{Z}^{\text{txt}})\right)\right\|_F^2, \tag{40}$$

which reduces redundant prototype directions (Cogswell et al., 2016; Zbontar et al., 2021).

For classes represented in the target cache,

$$\mathcal{L}_{\text{proto-align}} = \frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t}\left[1 - (\mathbf{z}_c^{\text{txt}})^\top \mathbf{z}_c^{\text{vis}}\right], \tag{41}$$

aligning textual and target visual evidence under the same geometry (Chen et al., 2025). For an empty cache, this loss and the corresponding attraction term in Eq. (9) are set to zero.

Finally, for the k most similar competing prototypes $\mathcal{N}_k(c)$,

$$\mathcal{L}_{\text{margin}} = \frac{1}{Ck}\sum_{c=1}^{C}\sum_{c'\in\mathcal{N}_k(c)}\left[(\mathbf{z}_c^{\text{txt}})^\top \mathbf{z}_{c'}^{\text{txt}} - (1-m)\right]_+, \tag{42}$$

which separates locally confusable classes (Weinberger & Saul, 2009).

For unit transformed prototypes $\mathbf{a}, \mathbf{b}$,

$$\tfrac{1}{2}d_\theta^2(\mathbf{a},\mathbf{b}) = 1 - \mathbf{a}^\top\mathbf{b},$$
$$[m - \tfrac{1}{2}d_\theta^2(\mathbf{a},\mathbf{b})]_+ = [\mathbf{a}^\top\mathbf{b} - (1-m)]_+.$$

Consequently, $\mathcal{E}_{\text{disc}} = \lambda_a\mathcal{L}_{\text{proto-align}} + \lambda_m\mathcal{L}_{\text{margin}}$ and $\mathcal{R}_{\text{red}} = \mathcal{L}_{\text{decorr}}$. This is an algebraic regrouping with identical coefficients, neighbor sets, and empty-cache convention. In particular, Eq. (7) first transforms and normalizes each prompt, then averages and normalizes the result; the distance formulation preserves that order.

The complete geometry objective and online update are

$$\mathcal{L}_{\text{geom}} = \lambda_d\mathcal{L}_{\text{decorr}} + \lambda_a\mathcal{L}_{\text{proto-align}} + \lambda_m\mathcal{L}_{\text{margin}},$$
$$\mathcal{L}_{\text{online}} = \mathcal{L}_{\text{ReTA}} + \mathcal{L}_{\text{geom}}.$$

We initialize $d = \mathbf{1}$ and the low-rank factor U with zero-mean Gaussian noise of scale 10^{-3}, and optimize the shared geometry parameters with AdamW.

D Extended Reliability Analysis

D.1 Joint Reliability Quadrants

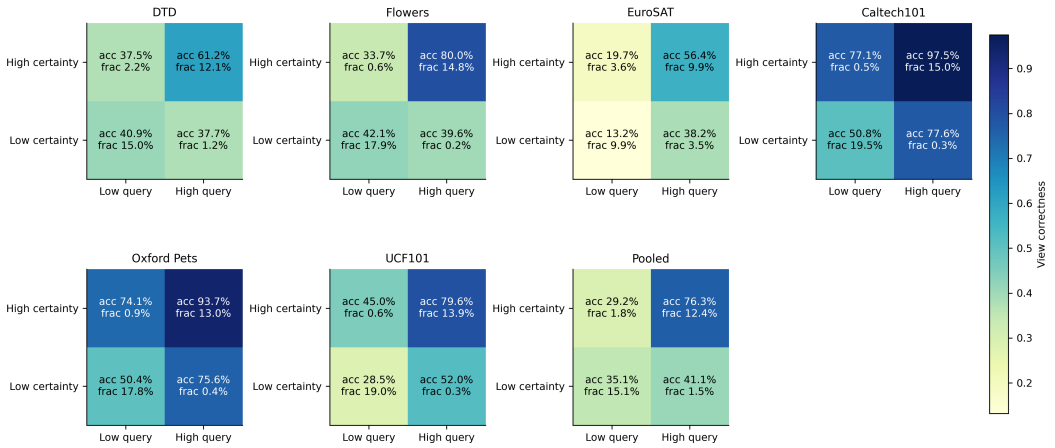

Figure 7: Joint reliability structure of frozen-space query consistency and adaptive predictive certainty with CLIP ViT-B/16. For each test instance, augmented views are partitioned by the top and bottom quartiles of their within-instance consistency and certainty ranks. Each cell reports auxiliary view-level accuracy (*acc*) and the fraction of augmented views in that quadrant (*frac*). The pooled panel computes both metrics on all augmented views concatenated across the six datasets, yielding view-weighted statistics rather than an equal-weight dataset average. The high-consistency/high-certainty quadrant achieves the highest view accuracy across all six datasets and in the pooled analysis.

Joint structure of consistency and certainty. Figure 7 examines the joint reliability structure of consistency and certainty across six datasets using CLIP ViT-B/16. For each test instance, the 63 augmented views are ranked independently by consistency and certainty, and we compare the four extreme rank quadrants defined by the top and bottom quartiles of the two signals. Each cell reports the auxiliary classifier's view-level accuracy together with the fraction of augmented views falling into that quadrant.

Views that are simultaneously high in consistency and certainty (HH) achieve the highest accuracy on all six datasets. Within the high-certainty region, high-consistency views are substantially more accurate than low-consistency views across all six datasets; pooled view accuracy increases from 29.2% to 76.3%. Thus, certainty alone does not identify reliable augmented evidence, while consistency provides substantial complementary information. This supports the dual-signal reliability design.

D.2 Cross-Dataset Disagreement Analysis

Stratified comparison in Figure 6(b). We define mean evidence strength as $A = (r_i^Q + r_i^H)/2$, where r_i^Q is the frozen-space consistency rank, and disagreement as $D = |r_i^Q - r_i^H|$. Within each dataset, augmented views are stratified into four width-0.1 bins over $A \in [0.4, 0.8)$. Within each bin, the bottom and top quintiles of D define low- and high-disagreement groups, respectively; the reported effect is their pooled accuracy difference (high minus low).

Cross-dataset disagreement risk. Figure 8 extends the disagreement analysis to six datasets with CLIP ViT-B/16. For each augmented view, disagreement is measured by the absolute difference between its within-instance consistency and certainty ranks, $|r^Q - r^H|$. Views are grouped into dataset-wise quintiles from low to high disagreement, and risk is defined as one minus auxiliary view-level accuracy.

Figure 8(a) shows that view risk increases with disagreement across all six datasets, despite their different baseline error rates. Figure 8(b) summarizes this pattern through the risk difference between the highest- and lowest-disagreement quintiles. The Q5–Q1 gap is positive for every dataset, ranging

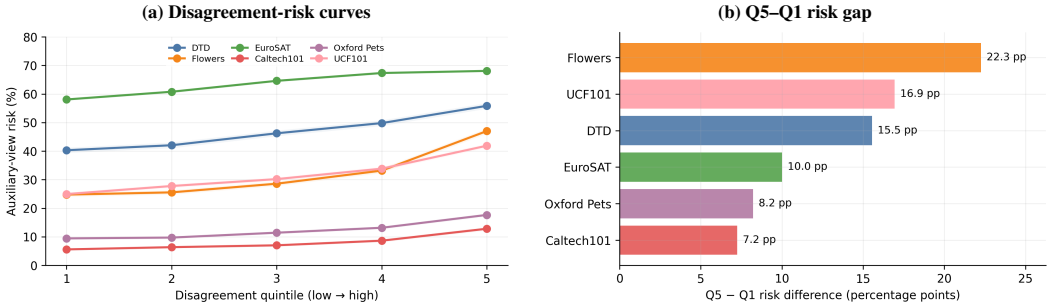

Figure 8: Cross-dataset relationship between reliability-signal disagreement and auxiliary view-level risk with CLIP ViT-B/16. (a) Per-view auxiliary risk across quintiles of $|r^Q - r^H|$, from low to high disagreement. (b) Risk difference between the highest- and lowest-disagreement quintiles. All six datasets exhibit a positive Q5–Q1 risk gap.

from 7.2 to 22.3 percentage points, with the largest separations on Flowers and UCF101. Together, these results show that disagreement between consistency and certainty provides a consistent cross-dataset indicator of unreliable augmented views, supporting agreement-sensitive reliability ranking.

D.3 Continuous Reliability Landscape

Figure 9 provides a continuous view of how consistency and certainty jointly relate to augmented-view reliability. Each hexagonal bin groups augmented views with similar within-instance consistency rank r^Q and certainty rank r^H, and its color denotes the auxiliary view-level accuracy within that bin.

Across the three datasets, higher reliability is generally concentrated in regions with stronger consistency and certainty, while the detailed landscape differs by dataset. Flowers exhibits a pronounced high-accuracy region when both ranks are large, and EuroSAT shows a clear increase in correctness toward higher consistency and certainty. DTD exhibits a smoother landscape, with weaker separation across the two-dimensional rank space. Importantly, the variation is not confined to a single axis: accuracy changes both across consistency at comparable certainty and across certainty at comparable consistency. The continuous landscapes therefore complement the quadrant and disagreement analyses by showing that the two signals provide distinct but interacting information about augmented-view reliability.

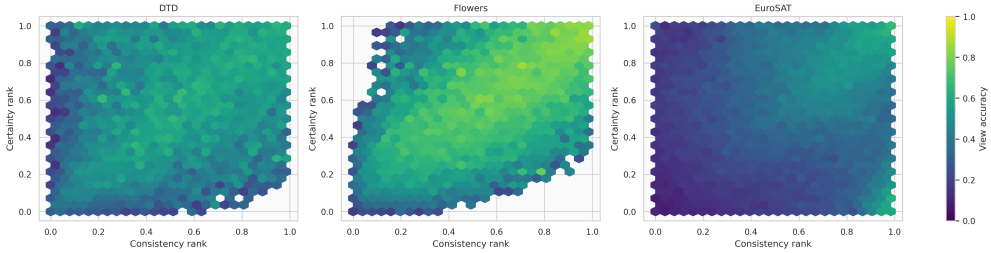

Figure 9: Continuous reliability landscapes for DTD, Flowers, and EuroSAT with CLIP ViT-B/16. The horizontal and vertical axes denote within-instance ranks of frozen-space query consistency and adaptive predictive certainty, respectively. Each hexagonal bin contains augmented views with similar rank pairs (bin width $= 0.05$ in normalized rank space), and its color reports auxiliary view-level accuracy; bins with fewer than 20 views are omitted and shown in white. Higher-accuracy regions depend jointly on both reliability signals, with different landscape structure across datasets.

E From Reliability to Multi-View Decision

E.1 Reliability Ranking Selects More Reliable Views

We next test whether the reliability structure identified in Appendix D translates into better view selection. The production GeoRel pipeline selects the Top-5 views by reliability score from 63 augmented candidates, with the clean/reference query excluded from the candidate set. The selected

views are aggregated using score-softmax weighting with temperature $\tau_r = 0.07$ before fusion with the primary branch. The Auxiliary and Final values for GeoRel are read directly from the saved online production outputs.

For a controlled selector comparison, all alternatives are replayed from the same saved online states and candidate views on the same 512 test instances per dataset–backbone setting. Residuals, geometry parameters, cache and EMA states, Primary predictions, and the remaining fusion settings are held fixed; only view selection and subsequent auxiliary aggregation are changed. Random selection is averaged over three independent selector draws.

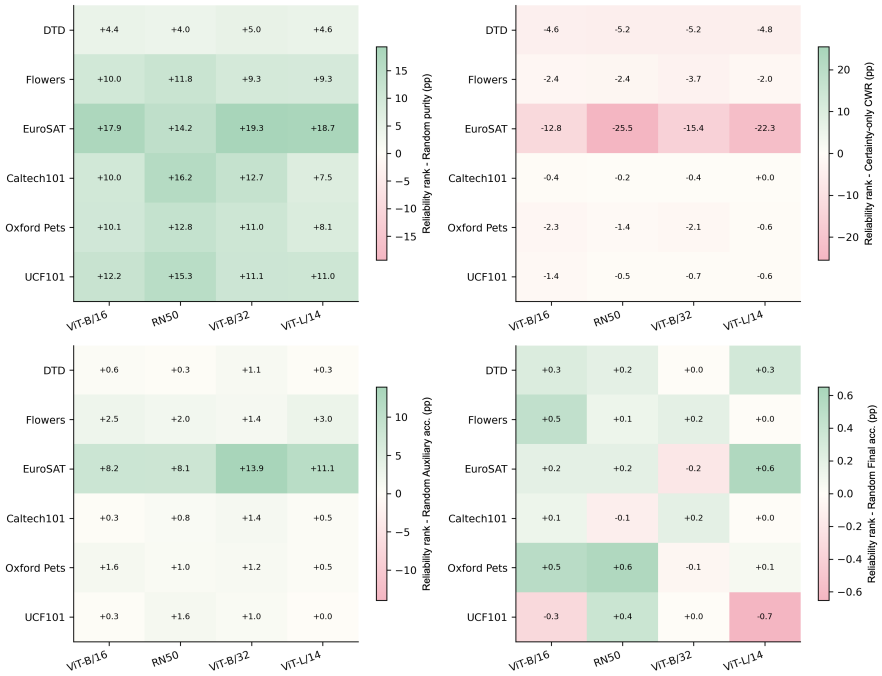

Figure 10: Controlled selector comparison across datasets and CLIP backbones, with samples, candidate views, and adapted states held fixed. Top left: reliability ranking minus Random in Purity@5. Top right: reliability ranking minus Certainty-only in CWR@5. Bottom left and bottom right: reliability ranking minus Random in Auxiliary and Final top-1 accuracy, respectively. Each entry is reported in percentage points.

Figure 10 summarizes the effect of reliability ranking across six datasets and four CLIP backbones. The proposed ranking improves Purity@5 over Random in 24/24 settings by 3.96–19.30 percentage points and reduces the confidently-wrong rate (CWR@5) relative to Certainty-only in 23/24 settings. Under this fixed-state replay protocol, these selection-quality differences are attributable to the selection rule. The results support the role of consistency in filtering high-certainty views that are not well supported by the unaugmented query. Reliability ranking also improves Auxiliary accuracy in 23 settings and matches it in the remaining one, with gains of up to 13.93 percentage points. Changes in Final accuracy are smaller because Final also incorporates the primary branch. These results connect the reliability structure in Appendix D to the role of anchored reliability ranking: combining consistency and certainty produces cleaner multi-view evidence for downstream prediction.

E.2 Robustness to the Retained-View Budget

Figure 11 examines the effect of the retained-view budget K across DTD, Flowers, and EuroSAT with four CLIP backbones. All values are obtained by replaying different retention sizes from matched saved online states while keeping the remaining prediction pipeline fixed. We evaluate $K \in \{1, 3, 5, 8, 16, 32, 63\}$, where $K = 63$ retains all augmented candidates; the clean/reference query is not among these 63 augmented candidates.

Final accuracy remains stable across small-to-moderate retention budgets, while retaining nearly all candidates degrades several dataset–backbone settings. For several dataset–backbone pairs, a small

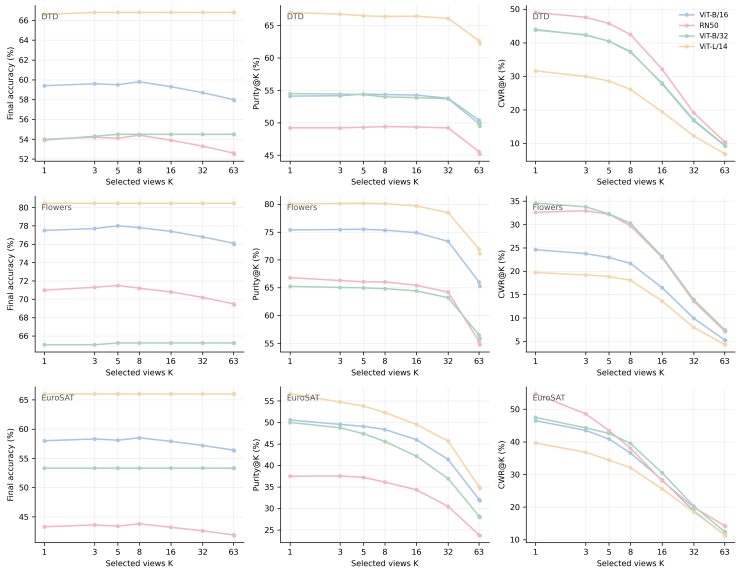

Figure 11: Robustness to the retained-view budget K across DTD, Flowers, and EuroSAT with four CLIP backbones. Left: Final top-1 accuracy. Middle: Purity@K of the retained augmented views. Right: confidently-wrong rate CWR@K. Increasing K progressively includes lower-ranked views. The fused Final prediction remains stable across small-to-moderate retention budgets, while retaining nearly all candidates degrades several settings.

or moderate retained set gives the strongest result without requiring a narrowly tuned value of K. In contrast, Purity@K generally decreases as K grows, with the clearest drop occurring when most or all candidate views are retained. This pattern shows that reliability ranking concentrates more reliable evidence at the top of the candidate list and progressively admits less reliable views as the retention budget increases.

CWR@K also decreases with larger K, but reflects a different aspect of the selected-set composition: adding lower-certainty views reduces the fraction of retained views that are simultaneously high-certainty and incorrect. Together, the three diagnostics show that reliability ranking places cleaner evidence toward the top of the candidate list, while the downstream Primary–Auxiliary fusion remains robust to substantial variation in the retention budget. Thus, the reliability ordering is more important than selecting a narrowly tuned value of K.

E.3 Branch Complementarity across Datasets and Backbones

Across six datasets and four CLIP backbones in Figure 12, Final improves over Auxiliary in all 24 settings. At the reported precision, Final is at or above Primary in 22 of 24; the exceptions are Flowers/ViT-B/32 (-0.4 pp) and Caltech101/RN50 (-0.2 pp). The asymmetry is strongest on EuroSAT, where Final stays close to Primary while substantially improving over Auxiliary.

F Robustness and Efficiency

F.1 Robustness to the Available-View Budget

View-budget protocol. Throughout this section, V denotes the total number of available views per test instance (the unaugmented query for the primary branch plus $V-1$ augmented candidates), using the same notation as Appendix C.2. We evaluate DTD, Oxford Flowers, EuroSAT, and Oxford Pets with CLIP ViT-B/16 and RN50 using 512 test samples per dataset–backbone setting. For each setting, we independently restart the online stream under $V \in \{1, 4, 8, 16, 32, 64\}$ available views per sample. The unaugmented view remains the query for the primary branch and the frozen reliability anchor. For $V > 1$, the reliability mechanism ranks the $N = V - 1$ augmented candidates and retains $K = \min\{5, N\}$ for score-softmax aggregation. For $V = 1$, no augmented candidate is available. The Auxiliary prediction is therefore computed from the sole clean view using the standard ReTA

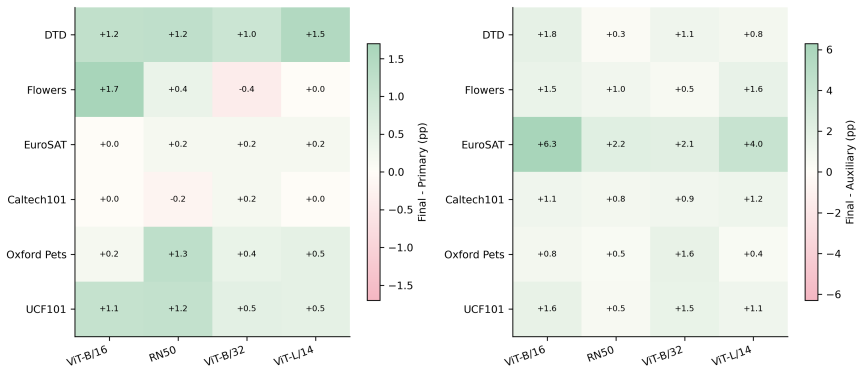

Figure 12: Branch complementarity across datasets and CLIP backbones. Left: Final minus Primary top-1 accuracy. Right: Final minus Auxiliary top-1 accuracy. Each entry is reported in percentage points; positive values (green colors) indicate that the fused Final prediction improves upon the corresponding branch; negative values (pink colors) indicate the reverse. The two panels use different color scales.

classifier with the configured cache correction, so multi-view selection and aggregation reduce to this single-view case. The Final prediction follows Eq. (22): $\mathbf{p}_{\text{final}} = \lambda_{\text{mix}}\mathbf{p}_A + (1 - \lambda_{\text{mix}})\mathbf{p}_{\text{Aux}}$. Different view budgets use nested subsets of the same cached 64-view pool, with the sample order, seed, and all remaining method hyperparameters held fixed.

Figure 13 evaluates GeoRel under different numbers of available views per test instance. Across DTD, Flowers, EuroSAT, and Oxford Pets with both ViT-B/16 and RN50, Final accuracy remains comparatively stable as the view budget is reduced, and the full view pool does not provide a consistent advantage.

The auxiliary branch is more dataset dependent. On EuroSAT, increasing the number of augmented candidates reduces Auxiliary accuracy, indicating that additional augmentations can introduce less useful evidence. The Final prediction nevertheless remains relatively stable because the primary branch provides a complementary prediction path. Across these settings, substantially reduced view budgets often preserve nearly the same Final accuracy as the full 64-view configuration.

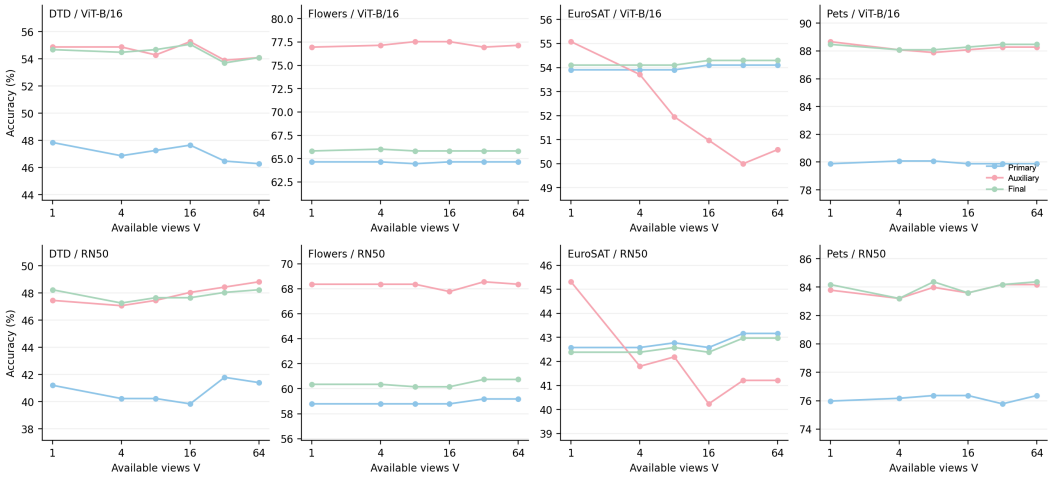

Figure 13: Robustness to the available-view budget across four datasets and two CLIP backbones. Each panel reports Primary, Auxiliary, and Final top-1 accuracy as the number of available views varies. The Final prediction remains comparatively stable under substantially reduced view budgets, while the auxiliary branch exhibits stronger dataset-dependent variation.

F.2 Computational Overhead of GeoRel

Figure 14 separates the additional cost of GeoRel into trainable parameters, reliability ranking and aggregation, and cached online adaptation. For feature dimension D, the structured geometry transform adds $D(r_g + 1)$ trainable parameters, which grows linearly with geometry rank; the default $r_g = 16$ remains small for both backbones. In the standalone synthetic-tensor microbenchmark, reliability ranking and probability aggregation remain below 1 ms across the tested retained-set sizes.

The cached timing measures online processing after visual features have been computed. The identity condition disables geometry adaptation while retaining the inherited ReTA state and reliability ranking, whereas the base condition enables full shared geometry adaptation. The cached online pipeline requires roughly 17–35 ms per sample across the evaluated datasets and backbones. The measurements exclude image encoding and initialization, and therefore isolate the additional online adaptation, branch-prediction, and view-aggregation cost. Thus, once visual features are available, enabling geometry adaptation adds approximately 4.5–14.9 ms per sample in median cached online latency in these measurements.

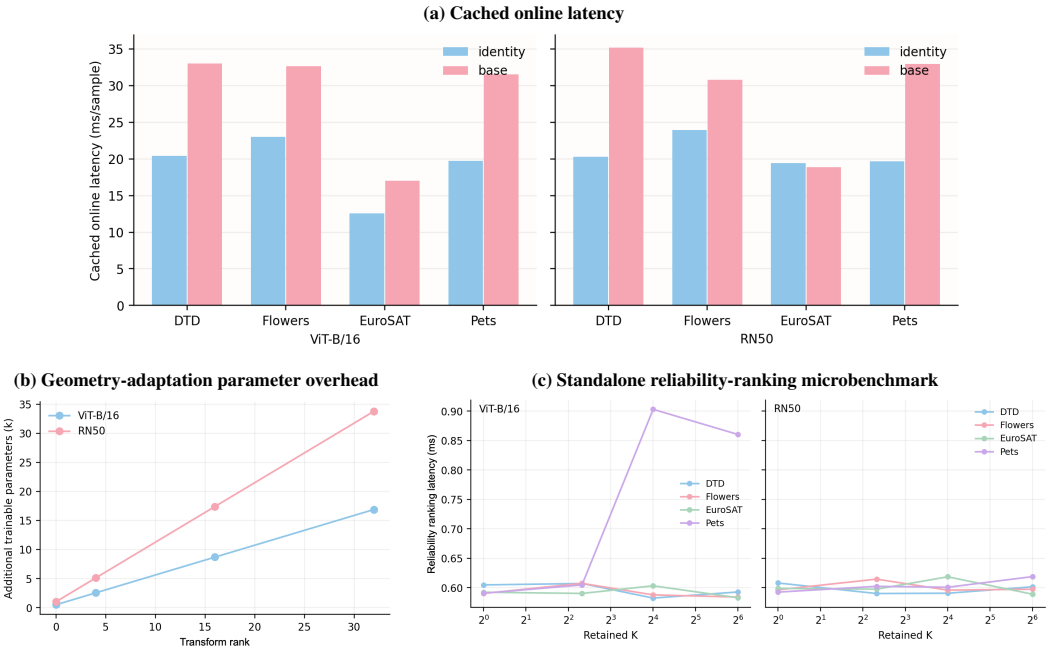

Figure 14: Computational overhead of GeoRel. (a) Cached online latency with and without geometry adaptation after visual feature extraction; timing excludes image encoding and initialization. (b) Additional geometry-adaptation trainable parameters as a function of geometry rank. (c) Standalone synthetic-tensor microbenchmark of reliability ranking and probability aggregation. Panel (c) intentionally sweeps retained sets up to $K = 64$ and is separate from the production protocol, which contains at most 63 augmented candidates.

F.3 View-Encoding Cost under Different Budgets

In these microbenchmarks, image encoding dominates the cost of reliability ranking and aggregation (Figures 14–15). Encoder latency and peak allocated memory both increase with the number of encoded views, with the largest growth for ViT-B/16. Reducing the view budget from 64 to 16 therefore substantially reduces encoding cost, while Figure 13 shows little degradation in Final accuracy on most evaluated settings.

Together, these measurements identify reduced view encoding as an effective route to lower inference cost in the evaluated settings.

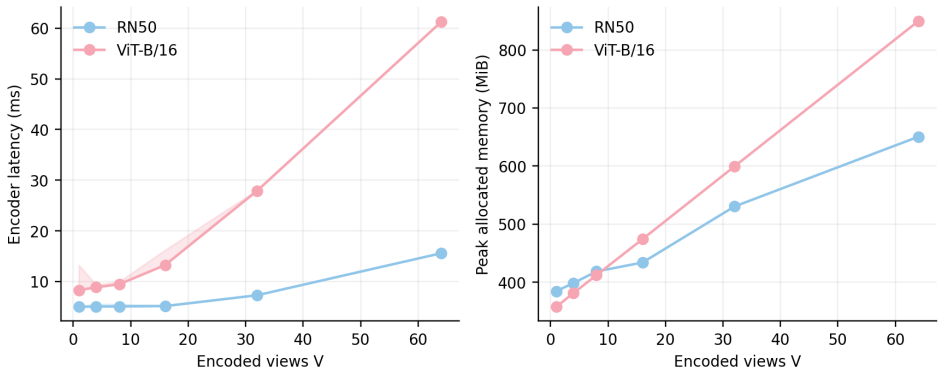

Figure 15: View-encoding cost under different available-view budgets for RN50 and ViT-B/16. Left: encoder latency as the number of encoded views increases. Right: peak allocated memory. Both costs increase with the view budget, linking the reduced-budget robustness in Figure 13 to lower computational and memory cost.

F.4 Configuration-Family Sensitivity Across OOD Shifts and Backbones

Figure 16 evaluates several alternative configuration families relative to the default GeoRel configuration used in the main experiments, across five ImageNet-family datasets and two CLIP backbones. The families jointly vary adapter capacity, regularization, prototype fusion, residual/basis capacity, or auxiliary learning, and therefore represent configuration-level sensitivity rather than single-factor ablations. No alternative family improves uniformly across shifts and backbones; its effect depends on both the target distribution and encoder architecture. For example, Prototype fusion decreases accuracy on all five RN50 benchmarks, whereas Learned auxiliary improves ImageNet, ImageNet-A, and ImageNet-R with ViT-B/16 but decreases ImageNet-Sketch and ImageNet-V2.

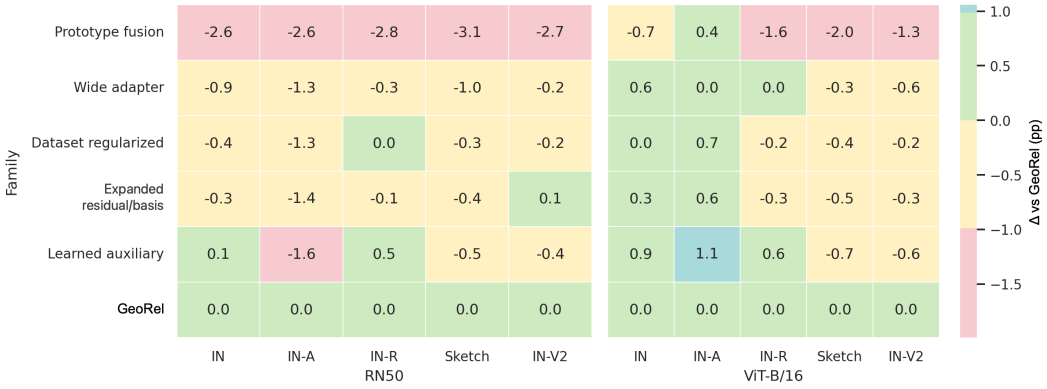

Figure 16: Configuration-family sensitivity across OOD shifts and backbones relative to the default GeoRel configuration. Each cell reports the change in Final accuracy (percentage points) under the same dataset and backbone. Left: RN50. Right: ViT-B/16. Expanded residual/basis uses $n_{\text{res}} = 5$ and basis dimension 128.

F.5 Qualitative Prediction Transitions

Figures 17–19 visualize representative sample-level transitions from the Primary prediction to the fused Final prediction on DTD, ImageNet-R, and Oxford Pets, using both ViT-B/16 and RN50. We show three transition types: rescued cases, where Final corrects a Primary error; harmed cases, where a correct Primary prediction becomes incorrect; and persistent errors, where both predictions remain wrong.

The examples illustrate that the fused prediction can resolve texture, style-shift, and fine-grained category confusions, while also introducing or preserving errors in visually ambiguous cases. These

qualitative transitions complement the aggregate rescue–harm statistics by showing the concrete prediction changes underlying them.

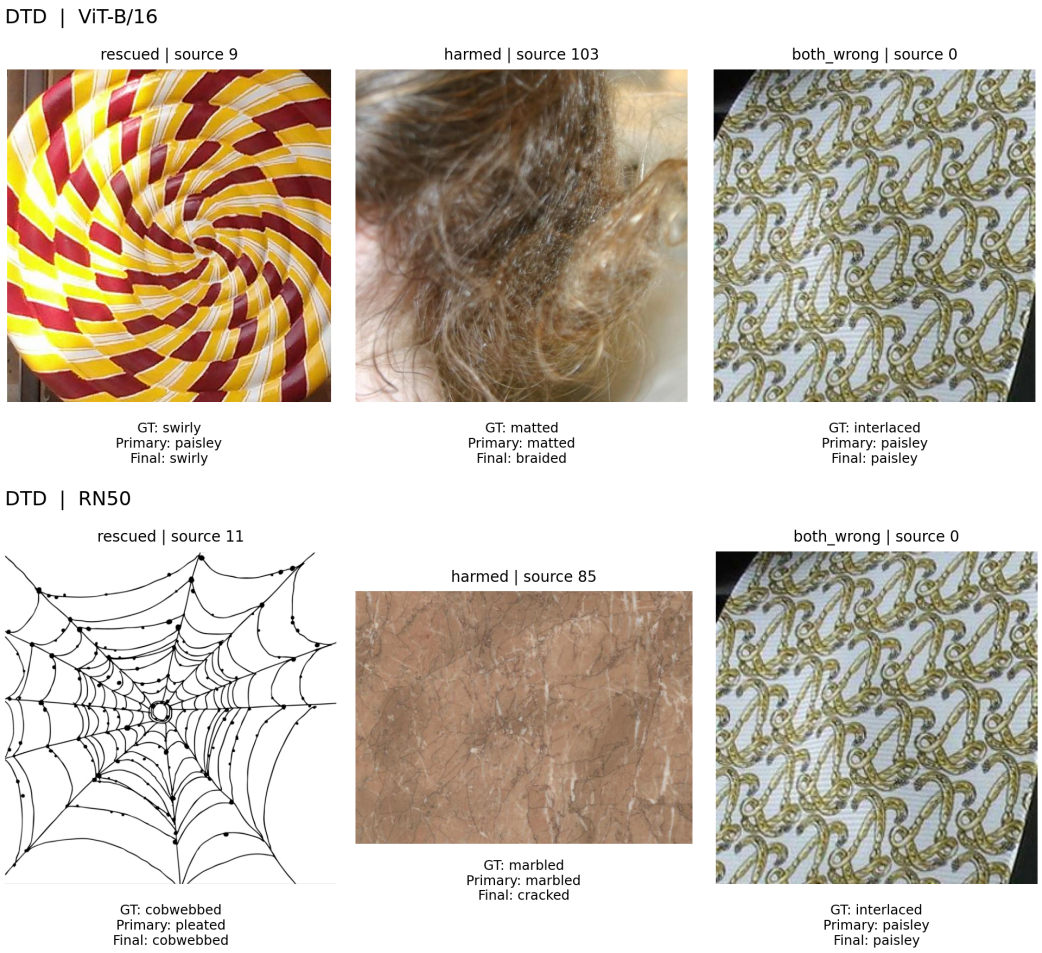

Figure 17: Representative Primary-to-Final prediction transitions on DTD. For each backbone, we show a rescued case (Primary wrong, Final correct), a harmed case (Primary correct, Final wrong), and a persistent error (both wrong). GT denotes the ground-truth class.

F.6 Qualitative Analysis Summary

The qualitative examples in Figures 17–19 illustrate the types of prediction transitions that contribute to the aggregate rescue–harm statistics. Final's ability to correct Primary errors on texture, style-shifted, and fine-grained categories demonstrates the value of reliability-filtered multi-view evidence. Conversely, the harm and persistent-error cases show that fusion is not uniformly beneficial, particularly when visual ambiguity affects both branches similarly.

G Additional Related Work

Test-time adaptation of vision–language models. Test-time adaptation updates a model from unlabeled samples observed during inference. Early fully test-time methods established entropy minimization, continual adaptation, and selective updating as central mechanisms for controlling prediction error and drift (Wang et al., 2021; 2022; Niu et al., 2022). For VLMs, one line adapts predictions from multiple test-time views: TPT optimizes prompts on confident augmentations, DiffTPT diversifies those views with diffusion, and ZERO shows that confidence-based view selection and low-temperature marginalization already form a strong optimization-free baseline (Shu et al., 2022; Feng et al., 2023; Farina et al., 2024). A second line accumulates target-stream evidence

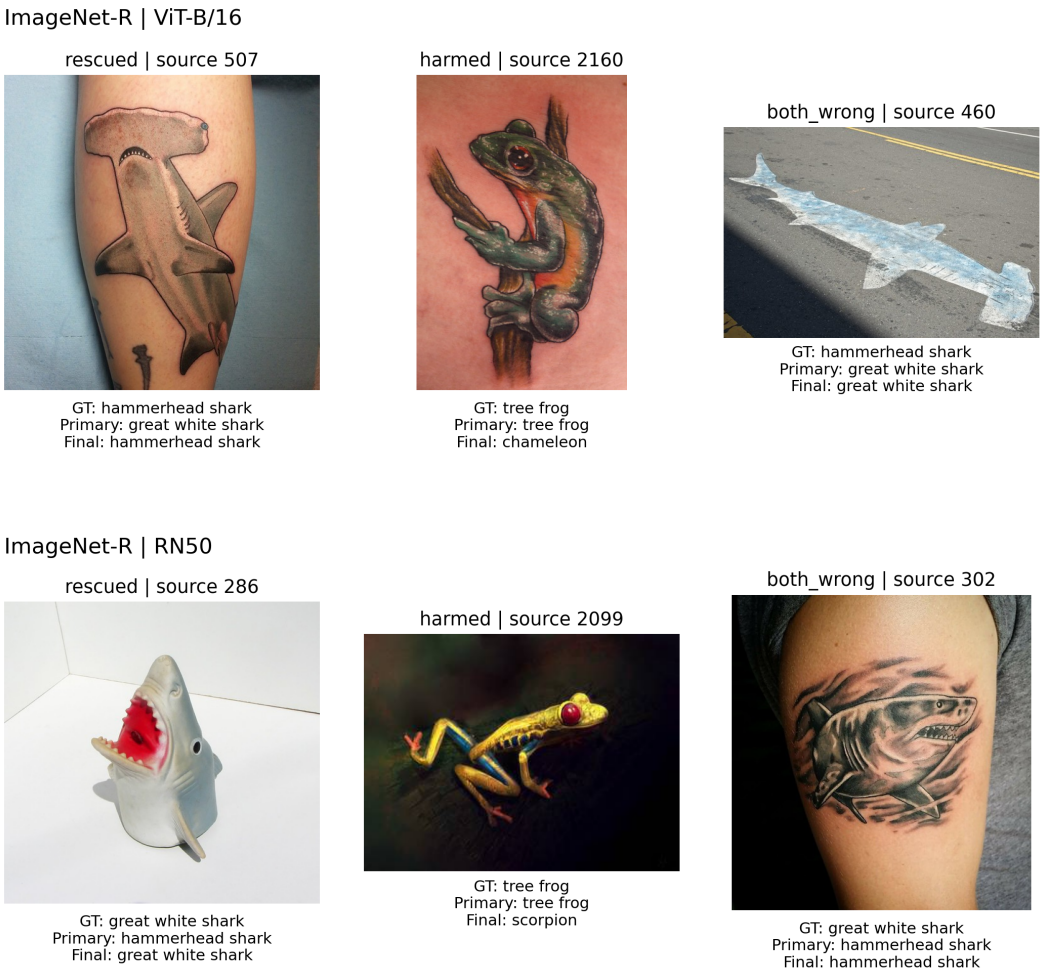

Figure 18: Representative Primary-to-Final prediction transitions on ImageNet-R. For each backbone, we show a rescued case (Primary wrong, Final correct), a harmed case (Primary correct, Final wrong), and a persistent error (both wrong). GT denotes the ground-truth class.

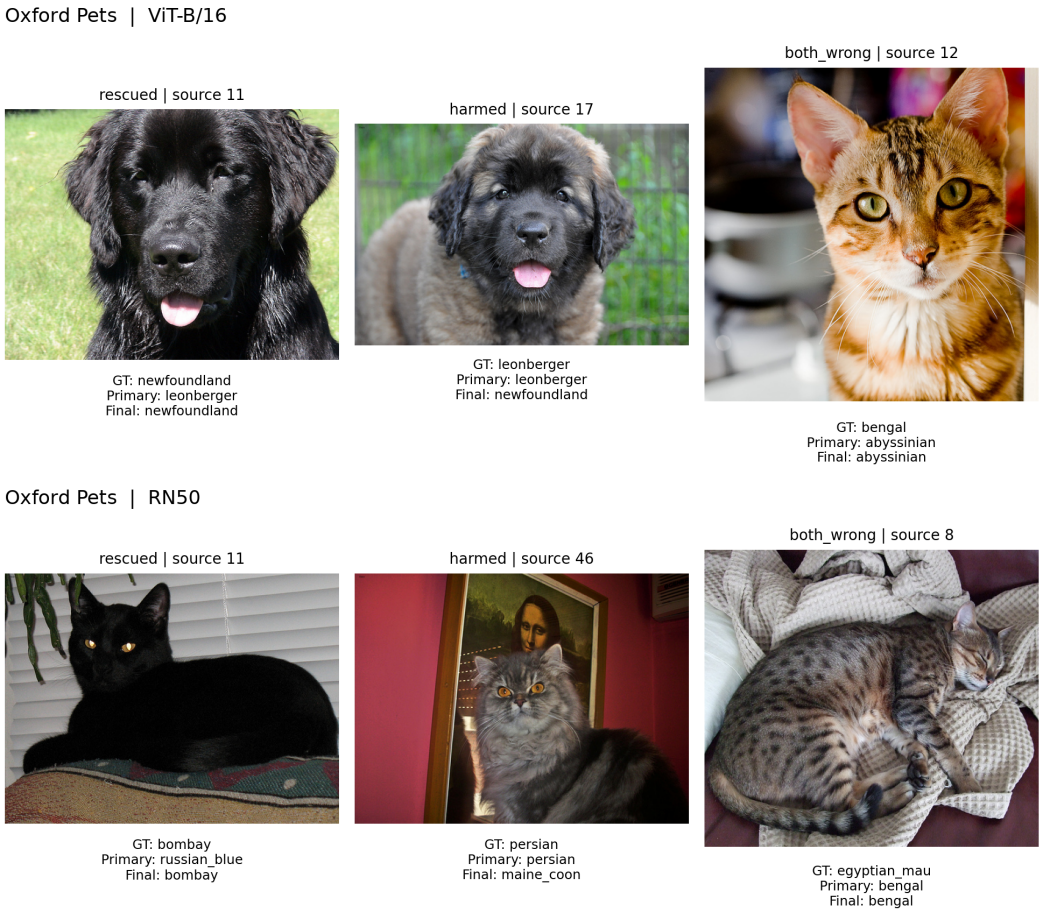

Figure 19: Representative Primary-to-Final prediction transitions on Oxford Pets. For each backbone, we show a rescued case (Primary wrong, Final correct), a harmed case (Primary correct, Final wrong), and a persistent error (both wrong). GT denotes the ground-truth class.

through caches or prototypes, including TDA (Karmanov et al., 2024), DMN (Zhang et al., 2024c), BoostAdapter (Zhang et al., 2024b), DPE (Zhang et al., 2024a), and TPS (Sui et al., 2025). Other methods broaden the adapted quantities through bimodal encoder updates, class embeddings and priors, normalization parameters, multi-cache prototypes, or logit recalibration (Maharana et al., 2025; Zhou et al., 2025; Hakim et al., 2025; Chen et al., 2025; Wu et al., 2026). ReTA is a strong residual- and cache-based baseline that combines residual-corrected prompt features with consistency-aware cache selection and class-distribution calibration (Liang et al., 2025). We retain its prompt and positive-cache machinery, but make the comparison geometry itself an explicit online adaptation target shared across representation types.

Calibration and reliable augmented views. MTA estimates augmented-view inlierness directly (Zanella & Ben Ayed, 2024); C-TPT and O-TPT address prompt-tuning calibration through text-feature dispersion and orthogonality, respectively (Yoon et al., 2024; Sharifdeen et al., 2025). Dual-Modality Anchor-Guided Filtering uses image and text anchors, CATE uses energy-guided consensus, and D2TPT analyzes overconfident errors and cross-modal misalignment (Choi & Kim, 2026; Wang et al., 2026; Song et al., 2026). GeoRel uses a frozen-query consistency signal and adaptive predictive certainty only through their within-instance ordinal structure; their raw magnitudes are not treated as calibrated reliability.

Target comparison geometry under distribution shift. Unlike the supervised setting of large-margin metric learning (Weinberger & Saul, 2009), GeoRel must learn from weak pseudo-labeled evidence online. Unlike covariance alignment methods (Sun et al., 2016; Sun & Saenko, 2016), it optimizes discriminative relations rather than directly matching or inverting covariance. Its geometry is shared across image features, text prototypes, and cache-derived visual prototypes; this shared cross-modal constraint is central to its online setting.