TEXT AS A WITNESS: ASSESSING VISUAL EVIDENCE FOR TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Textual prototypes provide class semantics, while test images reveal appearance in the deployment environment. Test-time adaptation must decide how strongly visual estimates built from inferred assignments should refine a classifier anchored by text. We introduce Witness-Assessed Visual Evidence (WAVE), which organizes this exchange around text as a reference for visual evidence and visual estimates as proposals for adapting class representations. Individual Evidence Assessment (IEA) uses the text class set to assess the response patterns of individual image–class assignments. Collective Evidence Assessment (CEA) assesses whether their aggregate favors its assigned class over its strongest competitor. The two scores jointly determine a proposal's influence, while the original soft assignments determine its visual statistics. Updated class statistics then revise assignments within the batch, closing the feedback loop around fixed text prototypes. WAVE supplies evidence assessment to a statistical inference procedure and operates on frozen embeddings without target labels or encoder backpropagation. Our analysis characterizes the residual geometry underlying IEA and connects proposal influence to regularization and textual decision regions. Across 14 datasets and 11 settings, combining both assessments improves mean accuracy over the stronger individual component by 2.77% on class-sequential streams and 2.15% in small batches sampled from one to four candidate classes. Controls reveal substantial benefits from overall attenuation, with additional gains from class-specific assessment in evaluated ImageNet settings. These findings support regulating the contribution of visual proposals while exposing the limitations of the textual reference.

1 INTRODUCTION

Text and images contribute different information to recognition. Textual class descriptions specify semantic categories and their relationships, while test images reveal how those categories appear in the current environment. A text prototype may not capture the background, style, or appearance of a deployment batch. Visual observations can supply this information, but their class memberships must be inferred. Test-time adaptation therefore depends on an exchange between a semantic reference supplied by text and an estimate of current appearance supplied by images.

The central difficulty is deciding how much influence the visual estimate should receive. Assigning several images to "tiger" creates a candidate estimate for that class. If some images depict cats, using this estimate to adapt the classifier can affect predictions for other images in the batch. The assigned mass indicates how much data supports the estimate numerically; it does not establish the semantic basis for using that estimate to revise a class representation. We ask: when visual data proposes an adaptation update, how should text assess its influence? Figure 1 illustrates this decision: uncertain assignments form a visual proposal, two text-anchored assessments evaluate its evidence, and the resulting influence affects predictions for other images in the same batch.

We organize this decision around two connected levels of evidence: Individual Evidence Assessment (IEA) and Collective Evidence Assessment (CEA). IEA assesses each image–class assignment against the textual class reference; CEA assesses the visual centroid formed from those assignments against the same class set. The two checks ask whether individual assignments support a candidate class and whether their aggregate still favors that class. Here, individual refers to an image–

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.

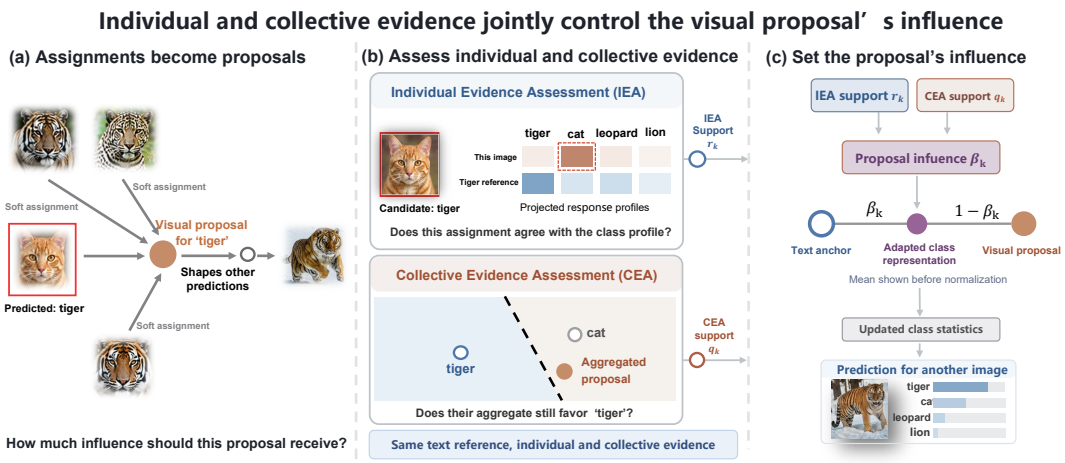


Figure 1: Visual proposals require evidence before influencing class statistics. (a) Soft assignments, including a cat assigned to tiger, form a visual proposal. (b) IEA compares an image–class response profile with its text reference; CEA checks whether the resulting centroid favors its assigned class. (c) These assessments and assigned mass set β_k, which fuses the proposal with the fixed text anchor and affects other images in the batch. Conceptual illustration; encoders and text prototypes remain fixed.

candidate-class pair, while collective refers to the resulting class estimate. Text prototypes remain fixed, and both assessments determine how strongly the visual estimate should influence the class model.

We develop Witness-Assessed Visual Evidence (WAVE) to make this decision. IEA uses projected image–text residual entropy to measure assignment compatibility over the text class set; CEA uses the visual centroid's margin against competing text classes to assess aggregate identity. Their scores set a class-specific fusion weight. Original soft assignments still determine the visual moments, and updated statistics revise assignments within the batch. The underlying solver follows Zanella et al. (2025); WAVE adds the evidence rule controlling proposal influence. We analyze the reference's observable geometry and the resulting fixed-state fusion decision.

Across 14 datasets and 11 settings, WAVE improves over the stronger individual assessment by 2.77 percentage points on class-sequential streams and 2.15 points in B64 batches drawn from one to four candidate classes. Overall attenuation explains a substantial part of the gain, while a class-specific control shows additional gains in evaluated ImageNet small-batch settings. Initial diagnostics do not show residual entropy to be a stronger general-purpose correctness detector than confidence, and some datasets expose limits of the fixed textual reference.

2 Related Work

Test-time adaptation. Tent (Wang et al., 2021) updates batch-normalization parameters by minimizing prediction entropy. Later methods filter uncertain samples, stabilize optimization, or handle changing streams (Niu et al., 2022; 2023; Wang et al., 2022; Gong et al., 2022; Yuan et al., 2023; Press et al., 2023; Su et al., 2024). Parameter-free output adaptation (Boudiaf et al., 2022) and forward-only adaptation (Niu et al., 2024) reduce the need for gradient updates. These studies motivate asking which unlabeled evidence should influence adaptation. WAVE makes that decision for a batch-formed class estimate inside a statistical solver rather than for an encoder update. Its residual score is not claimed to be a generally stronger correctness detector than confidence.

Vision–language adaptation. CLIP (Radford et al., 2021) enables zero-shot recognition but is vulnerable to distribution shift (Hendrycks et al., 2021; Tu et al., 2023). TPT adapts prompts using augmented views of individual images (Shu et al., 2022); calibration and distribution alignment extend this approach (Yoon et al., 2024; Abdul Samadh et al., 2023). MTA instead performs training-free, augmentation-based inference (Zanella & Ben Ayed, 2024). Cache and prototype

methods revise predictions through pseudo-labeled memory or evolving class representatives (Karmanov et al., 2024; Zhang et al., 2024a;c; Chen et al., 2025). Conditional prompting and weight-space anchors also preserve pretrained knowledge during adaptation (Zhou et al., 2022; Zhang et al., 2024b). WAVE keeps encoders and text prototypes fixed and uses the shared text class set to assess the influence of current-batch visual estimates. Its assessments regulate aggregate statistics rather than selecting which test samples enter a cache.

Frozen-feature statistical inference. Graph and transductive approaches jointly infer labels from frozen test features (Stojnić et al., 2024; Zanella et al., 2024; Martin et al., 2024). TransCLIP regularizes Gaussian class statistics toward text embeddings, while Bayesian and Gaussian-alignment variants refine class likelihoods without encoder backpropagation (Zhou et al., 2025; Zhang et al., 2025). StatA (Zanella et al., 2025) addresses class-correlated streams and variable effective class counts through a text-anchored statistical solver; ADAPT (Zhang et al., 2025) updates class means with accumulated high-confidence features. WAVE adopts StatA's solver and adds class-specific evidence control over its visual proposals. Comparing responses over a shared class set relates to accounts of multimodal relational structure (Lu et al., 2026); here the projected text geometry specifies what the comparison can and cannot observe. Modality disagreement raises a related reliability question (Yang et al., 2024), but the present decision concerns the influence of an inferred class estimate.

3 Text References and Visual Proposals

3.1 From Textual Classes to Visual Proposals

Let a test batch contain N unlabeled images and let $K \geq 2$ be the number of candidate classes. Frozen encoders produce unit-normalized image features $\mathbf{f}_i \in \mathbb{R}^d$ and textual class prototypes $\mathbf{t}_k \in \mathbb{R}^d$. Each experimental prototype is the normalized embedding of a class description generated with a single dataset-specific template. The initial zero-shot prediction is

$$\hat{y}_{ik} = \frac{\exp(\kappa \mathbf{f}_i^\top \mathbf{t}_k)}{\sum_{j=1}^{K} \exp(\kappa \mathbf{f}_i^\top \mathbf{t}_j)}, \tag{1}$$

where $\kappa > 0$ is the classification logit scale. Soft assignments $z_{ik} \geq 0$, with $\sum_k z_{ik} = 1$, are initialized by $\hat{y}_{ik}$ and refined during adaptation. The soft class count s_k is the sum of the assignment probabilities for class k. The corresponding weighted average of image features defines a visual class estimate:

$$s_k = \sum_{i=1}^{N} z_{ik}, \qquad \mathbf{v}_k = \frac{1}{s_k} \sum_{i=1}^{N} z_{ik} \mathbf{f}_i, \qquad \hat{\mathbf{v}}_k = \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|_2}. \tag{2}$$

The centroid $\hat{\mathbf{v}}_k$ proposes batch-specific appearance for class k without replacing its fixed text prototype. Its inferred memberships leave its influence to be assessed. Equations omit numerical stabilizers and assume nonzero vectors for normalization.

3.2 The Decision Linking the Two Modalities

The count s_k measures assigned mass, not semantic support. Text-class relations assess the source assignments, while the text classifier assesses their centroid.

Let β_k denote the proposal's visual influence. Low assessed support reduces its contribution, and updated statistics revise assignments within the batch (Sections 4.4–4.5). Agreement with the frozen text reference does not certify label correctness.

4 WAVE: Assessing Proposals in the Cross-Modal Loop

WAVE assesses image-class assignments with Individual Evidence Assessment (IEA) and their centroid with Collective Evidence Assessment (CEA). Both scores control the fusion weight; original soft assignments still form the visual statistics. Updated statistics then revise assignments within the batch.

4.1 INDIVIDUAL EVIDENCE ASSESSMENT (IEA)

A candidate class provides a reference pattern against which to assess an image. For a proposed assignment to class j, IEA compares the image's responses across the class set with those of the textual prototype for j. This comparison examines the semantic support for a source assignment of the visual proposal. We first exclude the mean text direction, defining that direction and its orthogonal projector as

$$\mathbf{u} = \frac{\sum_{k=1}^{K} \mathbf{t}_k}{\left\| \sum_{k=1}^{K} \mathbf{t}_k \right\|_2}, \qquad \mathbf{P} = \mathbf{I} - \mathbf{u}\mathbf{u}^\top, \qquad \mathbf{a}_k = \mathbf{P}\mathbf{t}_k. \tag{3}$$

We assume that the textual prototype sum is nonzero. The projection specifies which component is excluded from assessment; it does not assume that all domain shift or all nonsemantic variation lies in this direction. In particular, $\sum_k \mathbf{a}_k = \mathbf{0}$. Projected vectors are used without further normalization, preserving this centering property and the linear relations below.

The **class similarity profile** of an embedding $\mathbf{x}$ is its vector of inner-product responses to the projected prototypes:

$$\Phi_{\mathbf{P}}(\mathbf{x}) = \left[(\mathbf{P}\mathbf{x})^\top \mathbf{a}_1, \ldots, (\mathbf{P}\mathbf{x})^\top \mathbf{a}_K \right]^\top. \tag{4}$$

For a candidate assignment of image i to class j, we define its profile discrepancy as

$$\boldsymbol{\delta}_{ij} = \Phi_{\mathbf{P}}(\mathbf{f}_i) - \Phi_{\mathbf{P}}(\mathbf{t}_j), \qquad \delta_{ij,k} = [\mathbf{P}(\mathbf{f}_i - \mathbf{t}_j)]^\top \mathbf{a}_k. \tag{5}$$

Equation (5) measures a class-conditioned profile difference, so an image can receive different scores for different candidates. Projection makes this comparison invariant to a residual component along $\mathbf{u}$; profile agreement need not imply embedding agreement. Figure 2 illustrates the scoring path.

4.2 SCORING IEA WITH RESIDUAL ENTROPY

Residual entropy summarizes the profile difference used by IEA. Prediction entropy concerns uncertainty over candidate labels; here the softmax acts on the residual responses themselves. Its entropy describes how those differences are distributed over the reference classes:

$$p_{ij,k} = \frac{\exp(\tau \delta_{ij,k})}{\sum_{r=1}^{K} \exp(\tau \delta_{ij,r})}, \qquad H_{ij} = -\sum_{k=1}^{K} p_{ij,k} \log p_{ij,k}, \tag{6}$$

where $\tau > 0$ controls the residual response scale and is distinct from κ in Eq. (1). The distribution $\mathbf{p}_{ij}$ describes the residual responses over the reference classes. The following proposition establishes the meaning of its entropy.

Proposition 1 (Residual entropy and profile agreement). Under the construction in Eqs. (3)–(6),

$$H_{ij} = \log K \quad \Longleftrightarrow \quad \boldsymbol{\delta}_{ij} = \mathbf{0} \quad \Longleftrightarrow \quad \Phi_{\mathbf{P}}(\mathbf{f}_i) = \Phi_{\mathbf{P}}(\mathbf{t}_j). \tag{7}$$

Moreover, as $\tau \boldsymbol{\delta}_{ij} \to \mathbf{0}$, the entropy deficit satisfies

$$\log K - H_{ij} = \mathrm{KL}(\mathbf{p}_{ij} \,\|\, \mathbf{U}_K) = \frac{\tau^2}{2K} \|\boldsymbol{\delta}_{ij}\|_2^2 + O\big(\|\tau \boldsymbol{\delta}_{ij}\|_2^3\big), \tag{8}$$

where $\mathbf{U}_K$ is the uniform distribution over the K reference classes and logarithms are natural.

Proof. See Appendix D.1.

Proposition 1 motivates mapping higher residual entropy to greater support: its deficit locally measures squared profile discrepancy. The expansion gives no global ordering by Euclidean error, and text-profile compatibility does not certify label correctness.

Candidate evaluation and score mapping. We evaluate residual entropy for a subset of candidate assignments selected by the zero-shot predictions. Let

$$c_{\mathrm{pred}} = \left| \left\{ \arg\max_k \hat{y}_{ik} : 1 \le i \le N \right\} \right|, \qquad M = \min\{K, \max\{c_{\mathrm{pred}}, 5\}\}, \tag{9}$$

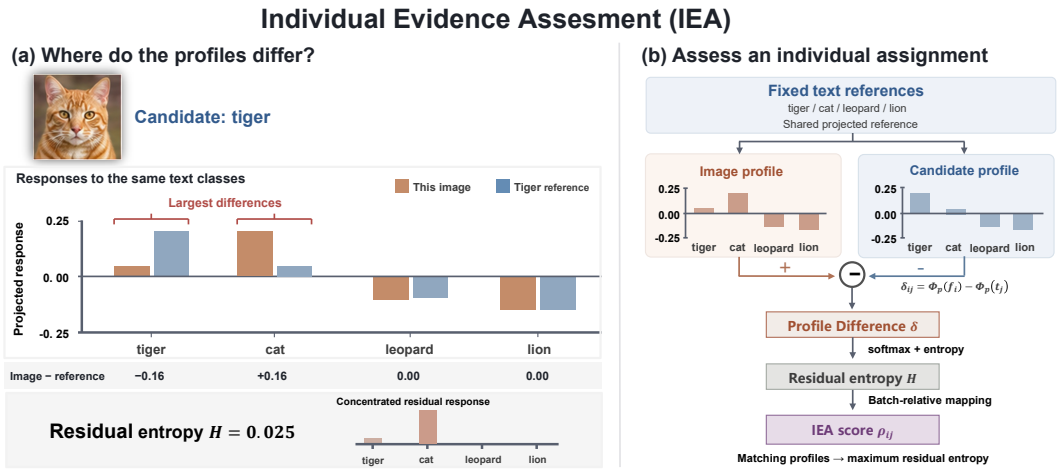


Figure 2: IEA compares an image's projected class-response profile with the candidate tiger profile. (a) Responses differ on tiger (-0.16) and cat ($+0.16$); the concentrated residual gives low entropy ($H = 0.025$ for $\tau = 40$), while matching profiles give maximum entropy. (b) The profile difference $\boldsymbol{\delta}_{ij}$ is mapped to the batch-relative score ρ_{ij}. Values are illustrative; the full score uses all K text classes and measures profile compatibility, not label correctness.

and let $\mathcal{J}_i$ contain the top-M classes for image i. For every evaluated pair $j \in \mathcal{J}_i$, Eq. (6) still uses all K reference classes. Thus, candidate truncation reduces the number of assignments evaluated while preserving the reference used in Proposition 1.

Let H_{ref} be the median of $\{H_{ij} : j \in \mathcal{J}_i,\ 1 \le i \le N\}$. The assignment evidence is

$$\rho_{ij} = \begin{cases} \sigma(\lambda_\rho(H_{ij} - H_{\text{ref}})), & j \in \mathcal{J}_i, \\ \frac{1}{2}, & j \notin \mathcal{J}_i, \end{cases} \tag{10}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ and $\lambda_\rho > 0$. Median centering expresses the assessment relative to the current batch, and unevaluated pairs receive the midpoint weight. These weights are relative evidence scores, without a probability calibration interpretation. WAVE computes $\boldsymbol{\rho}$ once per batch from the frozen embeddings and keeps it fixed during inference. Its fixed pairwise scores contribute to proposal assessment through the current assignments, as specified in Section 4.4.

4.3 Collective Evidence Assessment (CEA)

CEA checks whether the centroid $\hat{\mathbf{v}}_k$ formed by the current assignments favors its proposed class over competing text classes. It assesses the aggregate itself, not an average of IEA scores.

CEA uses the **centroid margin**, defined as the centroid's similarity to its assigned text prototype minus its largest similarity to a competitor:

$$m_k = \hat{\mathbf{v}}_k^\top \mathbf{t}_k - \max_{j \neq k} \hat{\mathbf{v}}_k^\top \mathbf{t}_j, \qquad q_k = \sigma(\gamma m_k), \tag{11}$$

where $\gamma > 0$ controls the score scale. This margin admits a direct decision-region interpretation. For the text-induced region

$$\mathcal{C}_k = \{\mathbf{v} : \mathbf{v}^\top \mathbf{t}_k \ge \mathbf{v}^\top \mathbf{t}_j \text{ for every } j \neq k\}, \tag{12}$$

we have $m_k \ge 0$ if and only if $\hat{\mathbf{v}}_k \in \mathcal{C}_k$. Negative margins indicate that the aggregate favors a competing text class and reduce its evidence weight. The margin assesses centroid identity under the text classifier; it does not establish the purity of the assigned samples.

IEA uses projected response profiles; CEA uses the original text-classification geometry. Neither score subsumes the other, and their combination does not assume independent correctness estimates.

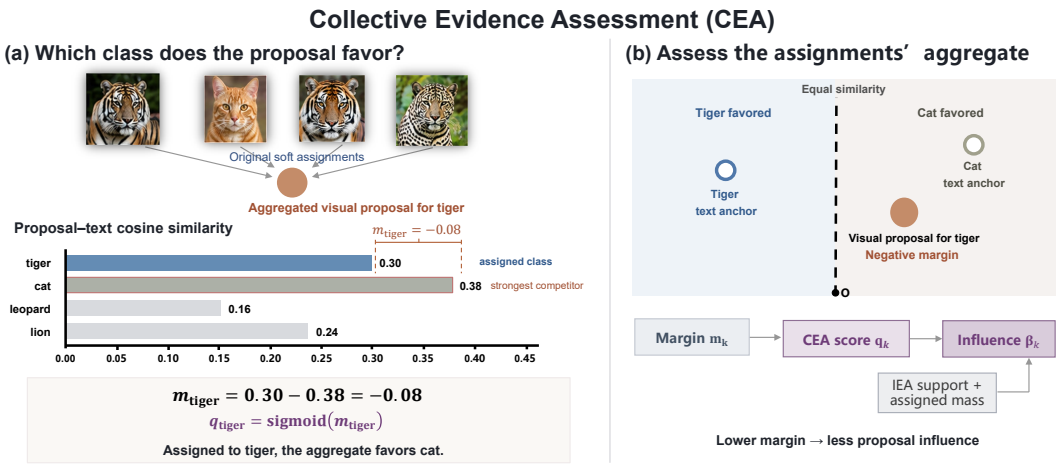


Figure 3: CEA assesses the aggregate visual proposal in the original text-classification geometry. (a) The tiger proposal favors cat (0.30 versus 0.38), giving $m_{\text{tiger}} = -0.08$ and $q_{\text{tiger}} \approx 0.018$ for $\gamma = 50$. (b) It lies beyond the tiger–cat equal-similarity boundary through the origin. At fixed assigned mass and IEA support, a lower margin reduces β_k. Original soft assignments form the centroid; the geometric panel illustrates the margin sign rather than exact coordinates.

4.4 Assigning Influence to the Visual Proposal

Individual and collective evidence meet at proposal influence. IEA supplies evidence about individual assignments, and CEA supplies evidence about their aggregate. We summarize the former by r_k and combine it with q_k to define the assessed support of the proposal:

$$r_k = \frac{\sum_i z_{ik}\rho_{ik}}{s_k}, \qquad n_k^{\text{rel}} = s_k r_k q_k = q_k \sum_i z_{ik}\rho_{ik}. \tag{13}$$

The product implements a soft joint requirement: at fixed assignments, weak support from either the source assessment or the centroid assessment attenuates the proposal. It is a modeling choice evaluated through the component experiments, and the factors are not independent correctness probabilities. With anchor strength $\alpha > 0$, the assessed support determines the visual contribution:

$$\beta_k = \frac{n_k^{\text{rel}}}{n_k^{\text{rel}} + \alpha}, \qquad \alpha > 0. \tag{14}$$

The parameter α sets the overall strength of the statistical anchor, while the evidence determines the class-dependent contribution of visual estimates. For fixed assignments, $0 \le n_k^{\text{rel}} \le s_k$, so $\beta_k \le s_k/(s_k + \alpha)$. Thus, the evidence rule can attenuate the corresponding soft-count update and recover it in the limit of unit evidence. The quantity n_k^{rel} represents weighted support for adaptation, without asserting that the observations are statistically independent.

Returning visual information to the class model. The proposal influences the classifier by contributing to class statistics initialized from text. The mean before output normalization is

$$\bar{\boldsymbol{\mu}}_k = \beta_k \hat{\mathbf{v}}_k + (1 - \beta_k)\mathbf{t}_k, \qquad \boldsymbol{\mu}_k = \frac{\bar{\boldsymbol{\mu}}_k}{\|\bar{\boldsymbol{\mu}}_k\|_2}. \tag{15}$$

Here $\bar{\boldsymbol{\mu}}_k$ denotes the mean before output normalization, and $\bar{\boldsymbol{\mu}}_k - \mathbf{t}_k = \beta_k(\hat{\mathbf{v}}_k - \mathbf{t}_k)$. This relation makes the action of the assessment explicit: it changes the magnitude of the proposed displacement. Visual means and covariance estimates remain weighted by the original z_{ik}; neither assessment directly reweights their contributing samples. Updated class statistics can subsequently affect the assignments and hence later estimates.

The statistical realization follows the text-anchored model of Zanella et al. (2025). We retain its covariance anchor and assignment solver, specified in Appendix C. The sigmoid mappings and the interpolation in Eq. (15) implement influence control; the geometric assessment is defined by the profile discrepancy and the aggregate's position relative to the text decision regions.

4.5 Closing the Loop within the Test Batch

The solver alternates assignment and statistics updates. Text fixes $\boldsymbol{\rho}$ for the batch; current assignments form centroids and the class-level r_k, while CEA updates q_k. Their fusion weights update the class statistics used in the next assignment phase. Algorithm 1 in Appendix C specifies the order.

Text prototypes and encoders remain fixed. Each batch starts fresh and discards its adapted state afterward, yielding episodic, batch-transductive evaluation. WAVE adds evidence-controlled influence to the solver of Zanella et al. (2025).

Assessment cost and chunked implementation are detailed in Appendix E.

5 Geometry and Limits of the Text Reference

The projected text set determines which residual directions IEA can assess. For $\mathbf{T}_P = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$, Eq. (5) gives $\boldsymbol{\delta}_{ij} = \mathbf{T}_P^\top \mathbf{P}(\mathbf{f}_i - \mathbf{t}_j)$: residuals orthogonal to its span are invisible. The span has rank at most $\min(d-1, K-1)$, and bounded residual logits yield quadratic bounds on the entropy deficit in this visible component. These are geometric bounds, not rankings by label correctness. At a fixed update state, $Q_k = r_k q_k$ makes Eq. (14) equivalent to a class-specific anchor strength α/Q_k when $Q_k > 0$. Thus assessment attenuates the soft-count update; it cannot strengthen it. The centroid margin checks a different, original text decision geometry. A nonnegative margin keeps the pre-normalization fusion segment in the assigned text region, but the implemented soft gate does not guarantee this identity when the margin is negative. Appendix F gives the divergence and spectral bounds, fixed-state optimization interpretation, sufficient margin condition, and proofs. None establishes label correctness or convergence of the alternating solver.

6 Experiments

We evaluate whether individual and collective evidence provide useful control of visual influence.

6.1 Experimental Setup

We use frozen CLIP ViT-B/16 features and one dataset-specific text prompt per class on 14 datasets. The 11 settings comprise four class-correlated streams (128 images per batch), three B64 candidate-class ranges (1–4, 2–10, 5–25), three B1000 ranges (5–25, 25–50, 50–100), and a full-test-set batch. Low, Medium, and High streams use Dirichlet sampling parameters $\eta = 0.1, 0.01, 0.001$, respectively; Separate presents classes sequentially. For each of five seeds, each stream setting reconstructs 100 streams and retains complete 128-image batches. Each B64 or B1000 setting samples 1,000 tasks; All is one full-test-set task. A stream score averages its batch accuracies before averaging streams, while batch tasks receive equal weight even when insufficient samples reduce their size.

Each incoming stream batch starts with fresh adaptation state and is adapted before its own predictions; no state carries across batches. Ground-truth labels construct the controlled test distributions and assess accuracy but never update the solver. Each setting averages dataset accuracies over five seeds. Appendix E gives exact task sampling, splits, prompts, timing, hyperparameters, and implementation details.

StatA-hard uses hard predicted-class counts, whereas IEA-only, CEA-only, and WAVE use the same soft-count rule; the StatA-soft control isolates this difference. The two single-assessment variants disable the other score. ADAPT is rerun under the same backbone, prompts, and evaluation harness. Accuracy differences are absolute percentage-point changes computed before rounding.

6.2 Does Assessing Visual Proposals Improve Adaptation?

WAVE obtains mean accuracies of 71.82% and 72.10% on highly correlated and class-sequential streams, and 73.26% in B64 batches sampled from one to four candidate classes (Table 1). Its largest advantages over StatA-hard occur in these concentrated distributions. WAVE exceeds StatA-hard in all eleven settings, including gains of 4.40, 5.13, and 5.12% in the three settings above.

Table 1: Mean accuracy across 14 datasets in each of 11 settings. StatA-hard disables both assessments; IEA-only and CEA-only retain one each. ADAPT is rerun under the same protocols. Δ_{hard} compares WAVE with StatA-hard; Δ_{single} compares it with the stronger single assessment. Macro averages the setting-level values.

Setting	Zero-shot	StatA-hard	IEA-only	CEA-only	ADAPT	WAVE	Δ_{hard}	Δ_{single}
Online Low	63.26	64.86	64.94	65.04	66.01	65.10	$+0.24$	$+0.06$
Online Medium	63.27	66.81	68.75	68.95	64.99	70.22	$+3.41$	$+1.27$
Online High	63.27	67.42	69.78	69.84	64.28	71.82	$+4.40$	$+1.98$
Online Separate	63.27	66.98	69.31	69.34	63.41	72.10	$+5.13$	$+2.77$
B64, 1–4	63.17	68.14	71.11	70.81	56.72	73.26	$+5.12$	$+2.15$
B64, 2–10	63.02	67.32	69.88	70.00	60.44	71.08	$+3.76$	$+1.08$
B64, 5–25	63.16	65.48	66.83	67.32	63.07	67.43	$+1.95$	$+0.11$
B1000, 5–25	63.24	67.15	68.51	68.86	57.54	70.26	$+3.11$	$+1.40$
B1000, 25–50	63.26	67.40	68.37	68.58	61.18	69.10	$+1.70$	$+0.51$
B1000, 50–100	63.21	66.74	67.43	67.32	63.90	67.95	$+1.21$	$+0.52$
All	63.27	67.32	67.38	67.21	67.20	67.88	$+0.56$	$+0.50$
Macro	—	**66.87**	**68.39**	**68.48**	**62.61**	**69.65**	$\mathbf{+2.78}$	$\mathbf{+1.12}$

These comparisons evaluate complete configurations; Appendix G separates the counting-rule and component effects.

At a fixed requested batch size, the gain decreases as the candidate-class range expands. For B64, it decreases from 5.12% in the 1–4 setting to 1.95% in the 5–25 setting. For B1000, it decreases from 3.11% in the 5–25 setting to 1.21% in the 50–100 setting. The smallest gains occur on Online Low ($+0.24\%$) and on the All setting ($+0.56\%$), which pools all classes in a single batch. The method is therefore most promising in the evaluated conditions with concentrated class distributions. Batch size, class composition, and sampling protocol should be considered jointly when interpreting these trends.

6.3 Comparison and Component Effects

WAVE exceeds the rerun ADAPT in ten of eleven settings (Table 1); ADAPT leads only on Online Low. Holding the soft-count rule fixed, WAVE exceeds both single-assessment variants in all settings, by 2.77 points on Online Separate and 2.15 points in B64, 1–4. Its gains over StatA-soft are 5.36, 2.93, and 0.90 points across the three B64 ranges. The full dataset–setting matrix and component breakdown are in Appendix G.

6.4 Overall Attenuation and Class-Specific Allocation

The evidence rule also reduces visual influence overall. In a StatA-hard sweep over $\alpha \in \{0.5, 1, 2, 2.5, 3, 3.5\}$, selecting the best value for each setting by its reported evaluation accuracy raises macro accuracy from 66.88% to 69.18%, versus 69.65% for WAVE. This is an oracle analysis within the tested grid, not a label-free tuning protocol; seven settings select its upper endpoint. On ImageNet B64, a Constant Gate that matches the active-class mean influence at its own state trails WAVE by 0.44, 0.76, and 1.06 points across the three class ranges. Thus global attenuation explains much of the gain, with additional evidence for class-specific allocation in those ImageNet settings. Table 2 reports the setting-level sweep. It shows that WAVE is higher in seven of eleven settings, including all three B64 settings, but lower on Online Low, Online Separate, and two B1000 settings. The $\alpha = 1$ StatA-hard values come from a separate sweep and differ slightly from Table 1. Constant Gate uses one quality multiplier g across classes while preserving soft-count dependence; at each update it matches the active-class mean WAVE-formula influence on its own assignments. This matching is local to the Constant branch's evolving assignments, not a trajectory-level match to the independently evolving WAVE run. The reported aggregates do not establish statistical significance. The gate equation and full diagnostic results are in Appendix G.

Table 2: Anchor-strength control across 14 datasets per setting. StatA-hard uses the best α^* selected within the tested grid by evaluation accuracy; WAVE uses $\alpha = 1$. The Δ column is WAVE minus grid-selected StatA-hard.

Setting	StatA-hard, $\alpha = 1$	α^*	StatA-hard, grid-selected	Δ
Online Low	64.85	2.5	65.19	-0.09
Online Medium	66.81	3.5	69.49	$+0.73$
Online High	67.44	3.5	71.42	$+0.40$
Online Separate	66.98	3.5	72.57	-0.46
B64, 1–4	68.16	3.5	72.25	$+1.01$
B64, 2–10	67.32	3.5	69.31	$+1.77$
B64, 5–25	65.48	2.5	65.74	$+1.69$
B1000, 5–25	67.16	3.5	71.29	-1.03
B1000, 25–50	67.40	3.5	69.35	-0.25
B1000, 50–100	66.74	2.0	66.83	$+1.12$
All	67.32	0.5	67.50	$+0.38$
Macro	**66.88**	Per setting	**69.18**	$\mathbf{+0.47}$

6.5 Score Diagnostics and Failure Boundaries

We evaluate the scores before adaptation on ImageNet and DTD to distinguish diagnostic behavior from the later feedback loop. In the two displayed B64 settings, residual entropy yields correctness-detection AUROCs of 0.530/0.523 on ImageNet and 0.627/0.631 on DTD, below maximum softmax probability (0.815/0.811 and 0.825/0.800). Thus it is not a generally stronger correctness detector; these comparisons do not test value conditional on matched confidence. At the aggregate level, the association with soft class purity is more favorable in some count strata: for ImageNet B64, 1–4 with $4 \leq s_k < 16$, Spearman correlations for IEA, CEA, and their product are 0.07, 0.44, and 0.31. Sparse and saturated strata limit this reading.

On EuroSAT B64, 5–25, WAVE falls from 45.10% (StatA-hard) to 35.07%; both single-assessment variants also decline. This is consistent with over-attenuating useful visual updates, without identifying its cause. Appendix G gives diagnostic definitions, plots, and dataset-level results.

7 Discussion

WAVE controls a visual proposal within StatA's alternating solver (Zanella et al., 2025); it does not establish that iteration itself adds benefit. The text reference constrains what IEA can observe, and its scores are neither correctness probabilities nor independent verification of the encoders. Stronger anchoring recovers much of WAVE's gain, while class-specific allocation adds setting-dependent improvements under the tested Constant Gate. The theoretical results concern score geometry and fixed-state fusion, not convergence, calibration, or final accuracy. These limits and the EuroSAT failure motivate testing influence rules against imperfect text references.

8 Conclusion

WAVE organizes test-time adaptation around text as a reference for visual evidence and visual estimates as proposals for refining class representations. IEA assesses the image–class assignments supporting a proposal, and CEA assesses its aggregate direction. Their scores jointly regulate the proposal's contribution to class statistics, which feed back into assignments within the batch. Text prototypes remain fixed throughout this exchange. The analysis characterizes the text-visible residual geometry and the role of assessed support in mean fusion. Experiments show benefits under concentrated class distributions, substantial contributions from overall attenuation, and additional setting-dependent gains from class-specific assessment. Failures emphasize that a fixed textual reference can also restrict useful visual updates. These findings motivate further evaluation of how to allocate visual influence when both the reference and the proposal are imperfect.

REFERENCES

Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Align Your Prompts: Test-Time Prompting with Distribution Alignment for Zero-Shot Generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-Free Online Test-Time Adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL `https://web.stanford.edu/~boyd/cvxbook/`.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust Continual Test-time Adaptation Against Temporal Correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric P. Xing. Efficient Test-Time Adaptation of Vision-Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jianglin Lu, Hailing Wang, Kuo Yang, Yitian Zhang, Simon Jenni, and Yun Fu. The indra representation hypothesis for multimodal alignment. *Advances in Neural Information Processing Systems*, 38:156980–157000, 2026.

Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive Zero-Shot and Few-Shot CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient Test-Time Model Adaptation without Forgetting. In *International Conference on Machine Learning (ICML)*, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards Stable Test-time Adaptation in Dynamic Wild World. In *International Conference on Learning Representations (ICLR)*, 2023.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-Time Model Adaptation with Only Forward Passes. In *International Conference on Machine Learning (ICML)*, 2024.

Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. RDumb: A simple approach that questions our progress in continual test-time adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Vladan Stojnić, Yannis Kalantidis, and Giorgos Tolias. Label Propagation for Zero-shot Classification with Vision-Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yongyi Su, Xun Xu, and Kui Jia. Towards Real-World Test-Time Adaptation: Tri-net Self-Training with Balanced Normalization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Weijie Tu, Weijian Deng, and Tom Gedeon. A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (CLIP). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual Test-Time Domain Adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time Adaption against Multi-modal Reliability Bias. In *International Conference on Learning Representations (ICLR)*, 2024.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, volume 2024, pp. 8636–8660, 2024.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust Test-Time Adaptation in Dynamic Scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Maxime Zanella and Ismail Ben Ayed. On the Test-Time Zero-Shot Generalization of Vision-Language Models: Do We Really Need Prompt Learning? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Maxime Zanella, Benoît Gérin, and Ismail Ben Ayed. Boosting Vision-Language Models with Transduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025.

Ce Zhang, Simon Stepputtis, Katia P. Sycara, and Yaqi Xie. Dual Prototype Evolving for Test-Time Generalization of Vision-Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. DePT: Decoupled Prompt Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. BoostAdapter: Improving Vision-Language Test-Time Adaptation via Regional Bootstrapping. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024c.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-Free Test-Time Adaptation via Probabilistic Gaussian Alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional Prompt Learning for Vision-Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian Test-Time Adaptation for Vision-Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

A Dataset Splits and Text Prompts

Dataset releases and splits. We use the official dataset layouts: Caltech101 (`101_ObjectCategories`), DTD r1.0.1, EuroSAT (`2750`), FGVC Aircraft 2013b, Oxford Flowers (`jpg` and official labels), Food101 (`images` and `meta`), Oxford-IIIT Pets (`images` and `annotations`), Stanford Cars (official train/test images and devkit annotations), and SUN397. UCF101 uses the CoOp midframe release (`UCF-101-midframes`). ImageNet uses ILSVRC2012 images with the official `ILSVRC2012_devkit_t12`; ImageNet-A, ImageNet-R, and ImageNet-Sketch use their official benchmark releases. The nine JSON-based datasets in Table 3 use the fixed CoOp (Zhou et al.) train/validation/test split files. FGVC Aircraft uses the official variant split, and ImageNet uses its official validation set.

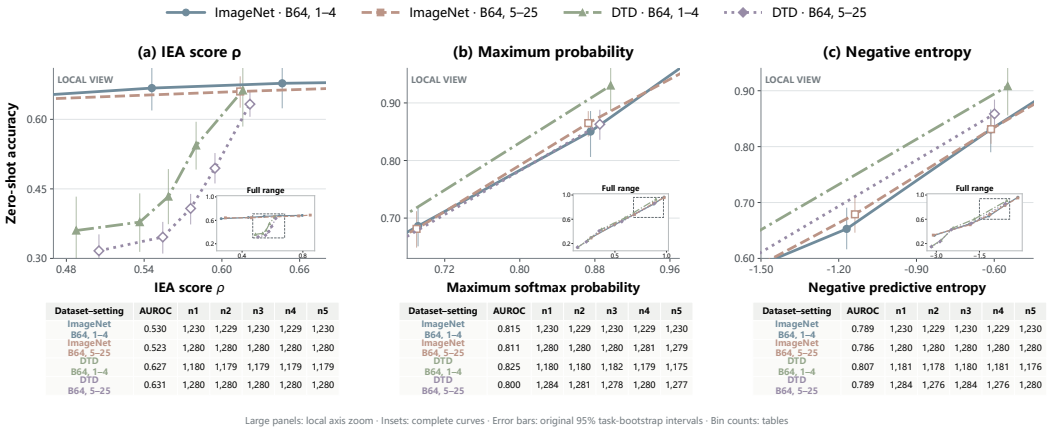
Table 3: Evaluation data sources and prompt templates. JSON-based datasets use the `test` entry of the fixed CoOp (Zhou et al.) split files. ImageNet uses the official validation set as its evaluation set. ImageNet-A, ImageNet-R, and ImageNet-Sketch use all images in the corresponding evaluation directories. The placeholder {} in each template is replaced by the class name after replacing underscores with spaces. Each resulting text embedding is normalized and used directly as the class prototype.

Dataset	Evaluation source	Template
ImageNet	`imagenet/val`, loaded with ImageFolder	`a photo of a {}.`
ImageNet-A / -R / -Sketch	Corresponding `images` directory; Sketch: `imagenet-s/images`	`a photo of a {}.`
FGVC Aircraft	`images_variant_test.txt`	`a photo of a {}, a type of aircraft.`
Caltech101	`split_zhou_Caltech101.json`	`a photo of a {}.`
DTD	`split_zhou_DescribableTextures.json`	`{} texture`
EuroSAT	`split_zhou_EuroSAT.json`	`a centered satellite photo of {}.`
Food101	`split_zhou_Food101.json`	`a photo of {}, a type of food.`
Oxford-IIIT Pets	`split_zhou_OxfordPets.json`	`a photo of a {}, a type of pet.`
Oxford Flowers	`split_zhou_OxfordFlowers.json`	`a photo of a {}, a type of flower.`
Stanford Cars	`split_zhou_StanfordCars.json`	`a photo of a {}.`
SUN397	`split_zhou_SUN397.json`	`a photo of a {}.`
UCF101	`split_zhou_UCF101.json`	`a photo of a person doing {}.`

The evaluation call uses `dataset.template` through `clip_classifier(..., reduce=None)`. Although the ImageNet dataset definitions also contain a seven-template `custom_templates` list, that list is not used by this evaluation path.

B VISUAL DIAGNOSTICS AND DATASET-LEVEL BOUNDARIES

The following figures present our diagnostic plots, including sparse-group counts and negative results. The individual and collective assessments are labeled IEA and CEA, and WAVE denotes the complete model. The class-level scores are r_k, q_k, and their product $Q_k = r_k q_k$. The class-count ranges in plot titles denote requested sampling pools, not the total number of classifier prototypes. The analyses are described in Sections G.2 and 6.5.



Dataset–setting	AUROC	n1	n2	n3	n4	n5
ImageNet B64, 1–4	0.530	1,230	1,229	1,230	1,229	1,230
ImageNet B64, 5–25	0.523	1,280	1,280	1,280	1,280	1,280
DTD B64, 1–4	0.627	1,180	1,179	1,179	1,179	1,179
DTD B64, 5–25	0.631	1,280	1,280	1,280	1,280	1,280

Dataset–setting	AUROC	n1	n2	n3	n4	n5
ImageNet B64, 1–4	0.815	1,230	1,229	1,230	1,229	1,230
ImageNet B64, 5–25	0.811	1,280	1,280	1,280	1,281	1,279
DTD B64, 1–4	0.825	1,180	1,180	1,182	1,179	1,175
DTD B64, 5–25	0.800	1,284	1,281	1,278	1,280	1,277

Dataset–setting	AUROC	n1	n2	n3	n4	n5
ImageNet B64, 1–4	0.789	1,230	1,229	1,230	1,229	1,230
ImageNet B64, 5–25	0.786	1,280	1,280	1,280	1,280	1,280
DTD B64, 1–4	0.807	1,181	1,178	1,180	1,181	1,176
DTD B64, 5–25	0.789	1,284	1,276	1,284	1,276	1,280

Figure 4: Initial-state sample reliability on ImageNet and DTD. Samples are grouped by residual score for their zero-shot predicted class, maximum softmax probability, or negative prediction entropy, covering both B64 settings of each dataset. Each panel lists the correctness-detection AUROC and per-bin sample counts in the table below; AUROC evaluates correctness detection. Large panels show local axis zooms and insets show the full curves; intervals use task-level bootstrap resampling. These curves compare score rankings without matching confidence between groups.

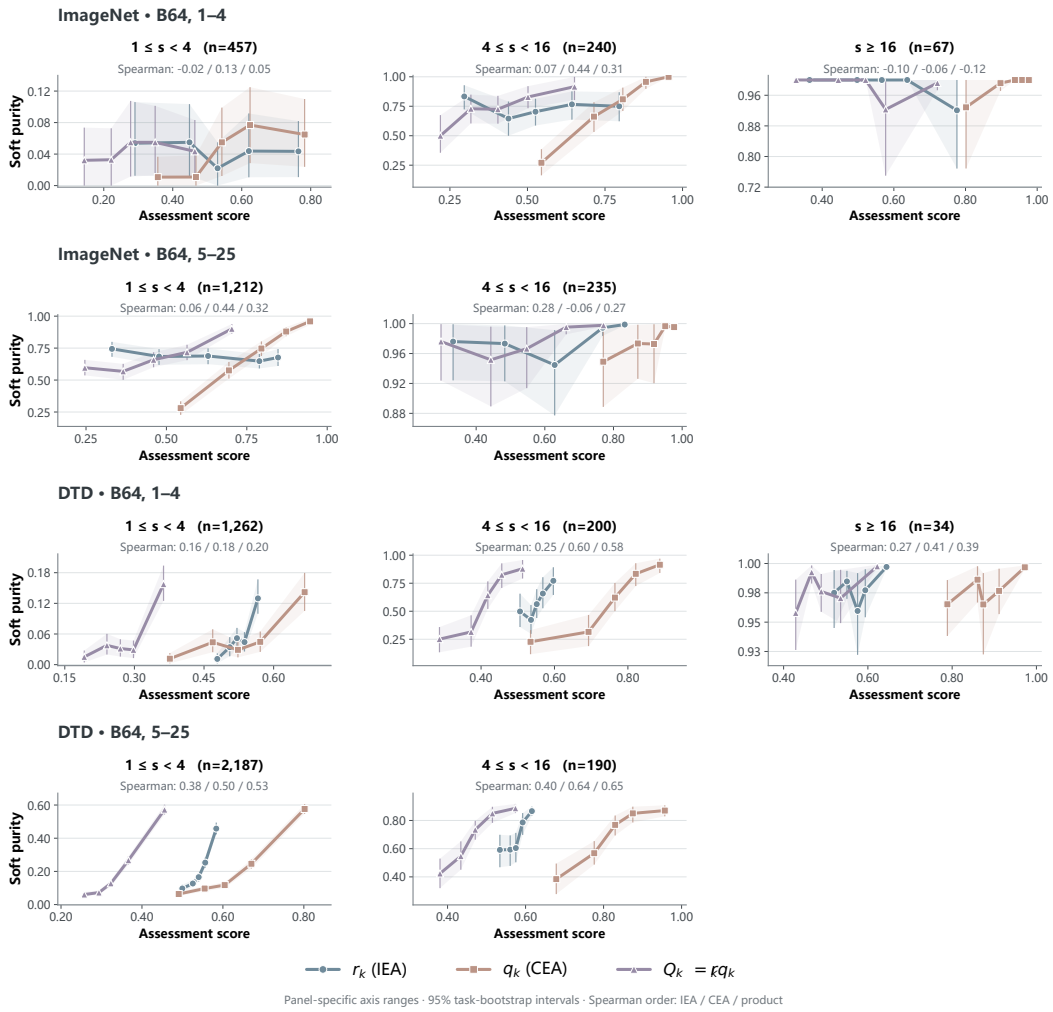


Figure 5: Initial class quality versus soft purity, stratified by soft count. Blue, orange, and purple curves denote r_k (IEA), q_k (CEA), and $Q_k = r_k q_k$, respectively; Spearman correlations follow that order. Each curve bins its own score, so the bins need not contain the same observations across scores. Blank positions denote unobserved $s_k \geq 16$ strata. Sparse and saturated strata limit interpretation.

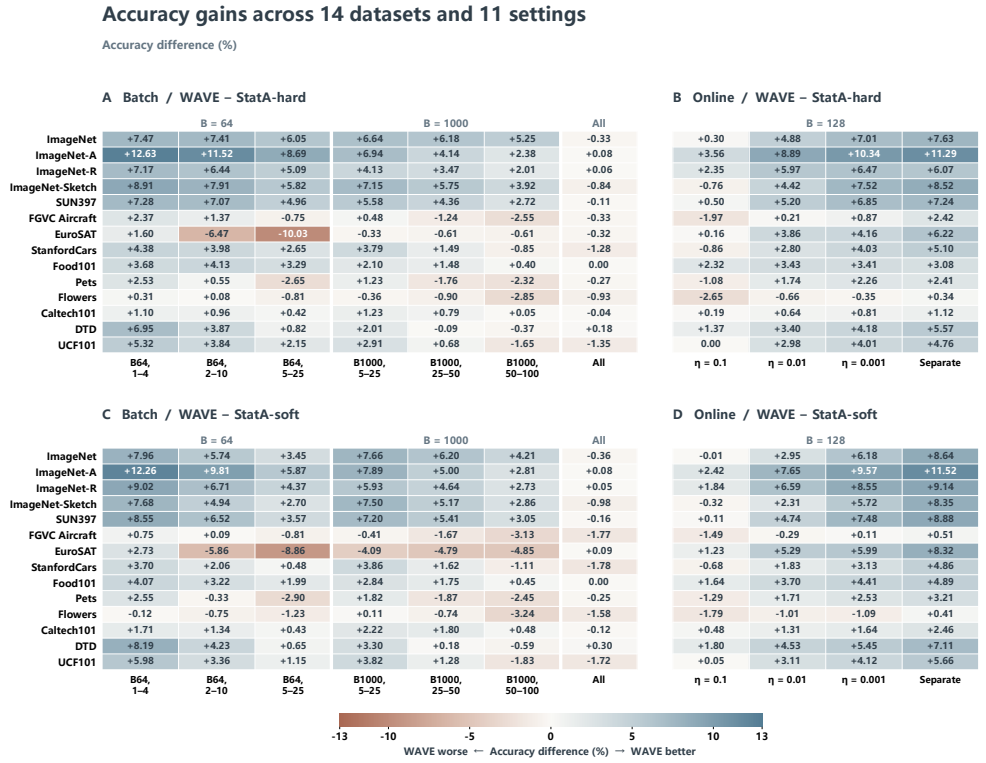


Figure 6: Accuracy changes across 14 datasets and 11 settings. The hard baseline is StatA-hard, and the soft baseline is StatA-soft. Colors are centered at zero with a symmetric scale. Different batch sizes are displayed separately, and η denotes the stream sampling parameter.

C Statistical Implementation and Reproducibility

This appendix specifies the statistical implementation used to evaluate WAVE. The Gaussian class model, covariance anchor, graph-regularized assignment solver, and mean interpolation follow StatA (Zanella et al., 2025). WAVE supplies the class-specific influence coefficient through Eqs. (13) and (14). The equations below are retained to make the procedure reproducible without requiring the reader to consult another paper.

C.1 Class Statistics and Initialization

Estimating moments and applying the weights. We instantiate the influence rule with the statistical anchor and moment structure of Zanella et al. (2025). The anchor mean is $\mathbf{t}_k$, and the shared initial diagonal covariance is

$$\mathbf{\Sigma}_0 = \mathrm{Diag}\left(\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{y}_{ik} (\mathbf{f}_i - \mathbf{t}_k)^{\odot 2} \right), \tag{16}$$

where $\odot 2$ denotes elementwise squaring and Diag maps a vector to a diagonal matrix. A small positive floor keeps covariance entries positive. We initialize $\boldsymbol{\mu}_k = \mathbf{t}_k$ and $\mathbf{\Sigma}_k = \mathbf{\Sigma}_0$.

Given the current assignments, we update the mean using Eq. (15) with the evidence-controlled coefficient in Eq. (14). The diagonal covariance update is

$$\begin{aligned} \mathbf{V}_k &= \mathrm{Diag}\left(\frac{1}{s_k} \sum_i z_{ik} (\mathbf{f}_i - \boldsymbol{\mu}_k)^{\odot 2} \right), \\ \mathbf{\Sigma}_k &= \beta_k \mathbf{V}_k + (1 - \beta_k) \left[\mathbf{\Sigma}_0 + \mathrm{Diag}\big((\mathbf{t}_k - \boldsymbol{\mu}_k)^{\odot 2} \big) \right]. \end{aligned} \tag{17}$$

This construction separates estimation and influence control. Equations (2) and (17) use the original assignment weights z_{ik} to compute visual moments; neither ρ_{ik} nor q_k directly reweights those moments. Their direct role is to control β_k. Updated statistics subsequently affect assignments, so later estimates can still change indirectly. As the evidence-weighted count approaches zero, the statistics approach the textual anchor. Text therefore provides both the reference for assessing visual evidence and the prior whose influence increases when that evidence receives low weight. Both roles depend on the quality of the pretrained model and class descriptions.

C.2 Assignment Solver and Complete Algorithm

Assignment updates. To make the adaptation procedure self-contained, we specify how the estimated class statistics produce new assignments. We use the text and neighborhood regularization of Zanella et al. (2025); this solver supplies the statistical inference component of WAVE. Define the Gaussian log score, omitting a constant shared by all classes, as

$$L_{ik} = -\frac{1}{2} \left[(\mathbf{f}_i - \boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k^{-1} (\mathbf{f}_i - \boldsymbol{\mu}_k) + \log |\mathbf{\Sigma}_k| \right]. \tag{18}$$

Let $\mathbf{W}$ be the sparse directed nearest-neighbor affinity matrix, with $W_{ij} = \mathbf{f}_i^\top \mathbf{f}_j$ on retained edges and zero elsewhere. With n_{nbr} neighbors per sample, write the scaled symmetric affinity as $\mathbf{A} = (50/n_{\mathrm{nbr}})(\mathbf{W} + \mathbf{W}^\top)$. Each inner assignment step is

$$z_{ik}^{(\ell+1)} = \frac{\hat{y}_{ik} \exp\big(L_{ik} + (\mathbf{A}\mathbf{Z}^{(\ell)})_{ik} \big)}{\sum_{j=1}^{K} \hat{y}_{ij} \exp\big(L_{ij} + (\mathbf{A}\mathbf{Z}^{(\ell)})_{ij} \big)}, \tag{19}$$

where $\mathbf{Z} = [z_{ik}]$. Equation (19) is implemented as a row-wise softmax of $\log \hat{y}_{ik} + L_{ik} + (\mathbf{A}\mathbf{Z}^{(\ell)})_{ik}$. The log-determinant contribution is already included in L_{ik}. Between assignment phases, WAVE reassesses the current centroids and recomputes visual influence before updating the class statistics. The complete procedure is specified by these updates; we do not require a convergence claim for a single fixed objective with adaptive weights.

Algorithm 1: WAVE (Witness-Assessed Visual Evidence).

Input: frozen features $\{\mathbf{f}_i\}_{i=1}^N$, text prototypes $\{\mathbf{t}_k\}_{k=1}^K$, and hyperparameters.

1. Compute zero-shot predictions using Eq. (1), set $\mathbf{Z} = \hat{\mathbf{Y}}$, initialize the class statistics using Eq. (16), and construct the neighbor graph.
2. Construct the textual witness's projected reference and candidate sets using Eqs. (3) and (9). Compute the fixed IEA scores $\boldsymbol{\rho}$ using Eqs. (5), (6), and (10).
3. For $b = 0, \ldots, B$:
   - Compute the Gaussian log scores in Eq. (18) and perform L inner assignment steps using Eq. (19).
   - If $b = B$, finish inference.
   - Form the current visual proposals and compute their CEA scores using Eqs. (2) and (11).
   - Combine both assessments and compute adaptation weights $\boldsymbol{\beta}$ using Eqs. (13) and (14).
   - Update the means and covariances using Eqs. (15) and (17).
4. Return the adapted assignments and predictions $\arg\max_k z_{ik}$ for the current batch.

Algorithm 1 implements the exchange within each batch. The fixed text prototypes supply initial class semantics, the IEA reference, and the reference for CEA. Visual proposals contribute to updated means and covariances, which revise assignments and thereby change later proposals. IEA is fixed within the batch; its assignment-weighted contribution, CEA, and the fusion weights evolve. Assignments, class statistics, the graph, and assessment scores are reinitialized at every call. Frozen feature caches and textual prototypes are reused, while adapted state is discarded between batches.

The default configuration uses $\kappa = 100$, $\tau = 40$, $\lambda_\rho = 2$, $\gamma = 50$, $\alpha = 1$, $n_{\mathrm{nbr}} = 3$, $B = 10$, and $L = 5$. Thus, inference performs ten statistics updates and eleven assignment phases. The default lower bound on β_k is zero. Numerical denominators use a small stabilizer, including $r_k = \sum_i z_{ik}\rho_{ik}/(s_k + \varepsilon)$ and $\beta_k = n_k^{\mathrm{rel}}/(n_k^{\mathrm{rel}} + \alpha + \varepsilon)$.

Computational cost. Assignment evidence is computed once per batch. Direct evaluation of the NM candidate residuals against K prototypes costs $O(NMKd)$ operations and can be processed in chunks, retaining the $N \times K$ evidence matrix without storing all candidate response distributions simultaneously. Each subsequent class-evidence update requires $O(NKd)$ operations for centroids and $O(K^2d)$ for their similarities to the textual prototypes. Moment estimation and Gaussian scoring are also processed in chunks to avoid materializing a full $N \times K \times d$ tensor. All operations act on frozen embeddings, so adaptation requires no encoder backpropagation.

D PROOFS AND SCOPE OF THE THEORETICAL RESULTS

We use the projected, unrenormalized text vectors of Eq. (3), all K reference classes for each evaluated candidate pair, and natural logarithms. The text-prototype sum is nonzero. Statements about normalized centroids and means assume that the corresponding pre-normalization vectors are nonzero. Fixed-state claims treat assignments and assessment scores as given; they do not assert that these quantities are statistically independent.

D.1 PROOF OF PROPOSITION 1

Since $\sum_k \mathbf{a}_k = \mathbf{0}$, the profile discrepancy satisfies $\sum_k \delta_{ij,k} = 0$. Entropy reaches $\log K$ exactly when $\mathbf{p}_{ij} = \mathbf{U}_K$. For a finite logit vector, softmax is uniform if and only if all logits are equal. With $\tau > 0$ and the zero-sum constraint, this common value must be zero. Thus $H_{ij} = \log K$ if and only if $\boldsymbol{\delta}_{ij} = \mathbf{0}$, which is equivalent to profile agreement by definition.

For the local expansion, write $\mathbf{x} = \tau\boldsymbol{\delta}_{ij}$ and $\psi(\mathbf{x}) = \log \sum_k e^{x_k}$. At the origin,

$$\nabla\psi(\mathbf{0}) = \frac{1}{K}\mathbf{1}, \qquad \nabla^2\psi(\mathbf{0}) = \frac{1}{K}\mathbf{I} - \frac{1}{K^2}\mathbf{1}\mathbf{1}^\top. \tag{20}$$

For $\mathbf{1}^\top\mathbf{x} = 0$, smoothness gives

$$\begin{aligned} \psi(\mathbf{x}) &= \log K + \frac{1}{2K}\|\mathbf{x}\|_2^2 + O(\|\mathbf{x}\|_2^3), \\ \mathbf{x}^\top\nabla\psi(\mathbf{x}) &= \frac{1}{K}\|\mathbf{x}\|_2^2 + O(\|\mathbf{x}\|_2^3). \end{aligned} \tag{21}$$

The softmax identity $\log p_k = x_k - \psi(\mathbf{x})$ yields $\log K - H = \mathbf{x}^\top\nabla\psi(\mathbf{x}) - \psi(\mathbf{x}) + \log K$. Substitution proves Eq. (8). The expansion is for fixed K as $\|\mathbf{x}\|_2 \to 0$. □

D.2 PROOF OF PROPOSITION 2

The identity $\mathrm{KL}(\mathbf{p}\|\mathbf{U}_K) = \log K - H(\mathbf{p})$ follows directly from the definition of KL divergence. Since $\nabla\psi(\mathbf{x}) = \mathbf{p}$ and $\psi(\mathbf{0}) = \log K$, Eq. (25) gives the Bregman equality in Eq. (26). To obtain the integral, define

$$F(t) = t\mathbf{x}^\top\nabla\psi(t\mathbf{x}) - \psi(t\mathbf{x}) + \psi(\mathbf{0}).$$

Then $F(0) = 0$, $F(1) = D$, and differentiation cancels the first-order terms:

$$F'(t) = t\mathbf{x}^\top\nabla^2\psi(t\mathbf{x})\mathbf{x}.$$

Integration over $[0, 1]$ proves the identity.

For the bounds, let $\mathbf{p}(t) = \mathrm{softmax}(t\mathbf{x})$. Its Hessian is

$$\nabla^2\psi(t\mathbf{x}) = \mathrm{Diag}(\mathbf{p}(t)) - \mathbf{p}(t)\mathbf{p}(t)^\top. \tag{22}$$

If $\|\mathbf{x}\|_\infty \le R$, then for every k and $t \in [0, 1]$,

$$\frac{e^{-2R}}{K} \le p_k(t) \le \frac{e^{2R}}{K}.$$

The quadratic form in Eq. (22) is the weighted variance of the coordinates of $\mathbf{x}$. Writing $p_{\min}(t)$ and $p_{\max}(t)$ for the smallest and largest weights,

$$\begin{aligned} \mathbf{x}^\top\nabla^2\psi(t\mathbf{x})\mathbf{x} &= \min_{c\in\mathbb{R}} \sum_k p_k(t)(x_k - c)^2 \\ &\ge p_{\min}(t) \min_{c\in\mathbb{R}} \sum_k (x_k - c)^2 = p_{\min}(t)\|\mathbf{x}\|_2^2, \\ \mathbf{x}^\top\nabla^2\psi(t\mathbf{x})\mathbf{x} &\le \sum_k p_k(t)x_k^2 \le p_{\max}(t)\|\mathbf{x}\|_2^2. \end{aligned} \tag{23}$$

The lower-bound equality uses $\sum_k x_k = 0$, so the unweighted minimizing constant is zero. Inserting the weight bounds into the integral and using $\int_0^1 t\,dt = 1/2$ proves Eq. (27). □

The bounded-logit result does not require an additional algorithmic threshold: R describes the domain over which the bound is evaluated. Constants may be very loose at the deployed residual scale. It is a comparison of mathematical quantities, not a confidence interval, statistical calibration, or claim that the score measures semantic error outside the text-visible geometry.

D.3 Proof of Proposition 3

Let $\mathbf{G} = \mathbf{T}_P\mathbf{T}_P^\top$. This matrix is symmetric positive semidefinite, with range $\mathcal{S}_T$ and kernel $\mathcal{S}_T^\perp$. Write $\mathbf{r} = \mathbf{\Pi}_T\mathbf{r} + (\mathbf{I} - \mathbf{\Pi}_T)\mathbf{r}$. The second component is annihilated by $\mathbf{T}_P^\top$, hence

$$\|\boldsymbol{\delta}\|_2^2 = \mathbf{r}^\top\mathbf{G}\mathbf{r} = (\mathbf{\Pi}_T\mathbf{r})^\top\mathbf{G}(\mathbf{\Pi}_T\mathbf{r}). \tag{24}$$

The eigenvalues of $\mathbf{G}$ restricted to its nonzero range lie between λ_- and λ_+. Applying the Rayleigh-quotient bounds proves Eq. (28). Positivity of λ_- proves the zero-discrepancy equivalence. Substituting $\mathbf{x} = \tau\boldsymbol{\delta}$ in Eq. (27) and then applying Eq. (28) gives Eq. (29).

Finally, $\mathbf{u}^\top\mathbf{T}_P = \mathbf{0}$ and $\mathbf{T}_P\mathbf{1} = \mathbf{0}$ imply the rank bound. In the rank-zero case, $\mathbf{T}_P = \mathbf{0}$ and every profile discrepancy is zero, so positive spectral constants are not defined and are not used. □

This proof only concerns the linear profile comparison at fixed coordinates. It does not establish invariance after arbitrary renormalization or relabeling of the inputs. Components invisible to the projected assignment check may still affect the original text classifier or the aggregate check.

D.4 Proof of Proposition 4

For fixed e_k, $\hat{\mathbf{v}}_k$, and $\mathbf{t}_k$, the gradient and Hessian of the objective in Eq. (30) are

$$2(e_k + \alpha)\mathbf{m} - 2e_k\hat{\mathbf{v}}_k - 2\alpha\mathbf{t}_k, \qquad 2(e_k + \alpha)\mathbf{I}.$$

Since $e_k + \alpha > 0$, the objective is strictly convex and its unique stationary point is

$$\mathbf{m} = \frac{e_k}{e_k + \alpha}\hat{\mathbf{v}}_k + \frac{\alpha}{e_k + \alpha}\mathbf{t}_k.$$

This is exactly $\bar{\boldsymbol{\mu}}_k$. For $e_k = s_kQ_k$ and $Q_k > 0$, division of numerator and denominator by Q_k gives the equivalent anchor strength α/Q_k. Since $Q_k \le 1$, this strength is at least α, establishing the soft-count upper bound on β_k. Subtracting $\mathbf{t}_k$ from the fusion equality and taking norms proves the displacement identity. The zero-quality case follows by setting $e_k = 0$. □

No sampling-distribution assumption is used. In particular, e_k need not be an unbiased count or a posterior precision. The minimizer statement holds before the implementation normalizes the mean. It is not a claim that the covariance update or assignment solver minimizes this quadratic objective.

D.5 Proof of Proposition 5

For each competitor $j \neq k$, linearity gives

$$\mathbf{w}_\beta^\top(\mathbf{t}_k - \mathbf{t}_j) = (1 - \beta)\mathbf{t}_k^\top(\mathbf{t}_k - \mathbf{t}_j) + \beta\hat{\mathbf{v}}_k^\top(\mathbf{t}_k - \mathbf{t}_j).$$

The first inner product is at least a_k, and the second is at least m_k. Their coefficients are nonnegative, so taking the minimum over j proves Eq. (33). For unit prototypes,

$$a_k = \min_{j \neq k}(1 - \mathbf{t}_k^\top\mathbf{t}_j) \ge 0.$$

It is strictly positive when no competing unit prototype equals $\mathbf{t}_k$. If $m_k \ge 0$, the lower bound in Eq. (33) is nonnegative. If $a_k > 0$ and $m_k < 0$, solving $a_k - \beta(a_k - m_k) \ge 0$ gives Eq. (34). Finally, for any $c > 0$, $g_k(c\mathbf{x}) = c\,g_k(\mathbf{x})$, so division by a nonzero vector's norm preserves the margin sign and region membership. □

The sufficient bound uses the two separate worst-case margins and need not be tight. It is not enforced by the implementation. In particular, a small sigmoid score alone need not ensure a small β_k when assigned mass is large. The zero vector belongs to the closed cone but cannot be normalized, which is why normalized conclusions explicitly exclude it. Identity throughout this proposition means identity under the fixed text prototypes, not correct labels or unchanged predictions under the adapted Gaussian model.

Table 4: Evaluation protocols. Repetitions are sampled tasks or reconstructed streams within each of the five random seeds. Candidate-class ranges control the sampling pool; the number of classes actually observed in a batch can be smaller.

Protocol	Setting	Batch size	Class sampling	Repetitions
Online	Low / Medium / High	128	$\eta = 0.1/0.01/0.001$	100 streams each
Online	Separate	128	Classes appear sequentially	100 streams
Batch	B64, 1–4 / 2–10 / 5–25	64	Candidate classes sampled from the indicated range	1,000 tasks each
Batch	B1000, 5–25 / 25–50 / 50–100	1,000	Candidate classes sampled from the indicated range	1,000 tasks each
Batch	All	Entire test split	All available classes	1 task

E Detailed Evaluation Protocol

Datasets and backbone. We evaluate on 14 classification datasets: ImageNet, ImageNet-A, ImageNet-R, ImageNet-Sketch, SUN397, FGVC Aircraft, EuroSAT, Stanford Cars, Food101, Oxford-IIIT Pets, Oxford Flowers, Caltech101, DTD, and UCF101. All reported runs use CLIP ViT-B/16 in evaluation mode, with the preprocessing returned by `clip.load`. ImageNet uses its official validation directory. ImageNet-A, ImageNet-R, and ImageNet-Sketch use their complete image directories. FGVC Aircraft uses the variant test list, and the remaining datasets use the test entries of their Zhou split files. Appendix A lists the exact split files and templates. Each class uses a single dataset-specific prompt, with underscores in class names replaced by spaces before text encoding. Its normalized text embedding initializes the class prototype. The experiments use neither seven-prompt nor 80-prompt ensembling. The image and text encoders remain frozen.

Test distributions. Following the evaluation setting of statistical adaptation (Zanella et al., 2025), we consider class-correlated streams and independently sampled batches. We retain the label Online for the stream settings, which use episodic adaptation on each incoming batch. Table 4 summarizes the 11 settings. We denote the stream sampling parameter by η to distinguish it from WAVE's margin scale γ in Eq. (11). Smaller positive η produces more concentrated class sequences. The Separate setting presents classes sequentially and is implemented separately from the positive-parameter sampler.

For batch tasks, the candidate-class count is sampled from the specified inclusive range, classes are selected, and images are drawn without replacement from their pooled samples. If the selected classes provide fewer images than the requested batch size, all available images are used. When the candidate count reaches the dataset's class count, all classes are eligible. For online evaluation, each task reconstructs a stream and partitions it into batches of 128, dropping an incomplete final batch. Ground-truth labels are used to construct the controlled class distributions and evaluate predictions; the adaptation solver uses only unlabeled features and textual prototypes. For positive stream parameter η, let N_{test} be the test-set size and define $S = \min\{\lfloor N_{\text{test}}/128 \rfloor, K\}$. Independently for each class k, we sample $\mathbf{p}_k \sim \text{Dirichlet}(\eta \mathbf{1}_S)$ and divide its samples among the S time slots using integer cut% from the cumulative proportions. We shuffle each slot internally, concatenate the slots in order, and retain only complete batches of 128.

Online state and prediction timing. StatA-hard, StatA-soft, and WAVE reset assignments, means, covariances, graph structure, and assessment scores for every incoming batch. They adapt jointly on the current batch and then compute its top-1 predictions, so each sample can influence other predictions within that batch. No adapted parameters or memory are carried forward. All test images have their frozen CLIP features extracted in advance, and these features and textual prototypes are reused. Each adaptation call accesses the current batch's features without using later batches to update its solver. This is batch-transductive evaluation along a stream, rather than a predict-before-update protocol.

Implementation. We use the settings in Appendix C: $\alpha = 1$, residual scale $\tau = 40$, assignment-score slope $\lambda_\rho = 2$, margin scale $\gamma = 50$, three graph neighbors, ten statistics updates, and five inner assignment steps per phase. Assignment scores are computed once per batch, whereas centroid

assessments and visual influence weights are updated during inference. Experiments use NVIDIA GeForce RTX 4090/4080 GPUs, Python 3.10.0, NumPy 1.26.4, PyTorch 2.11.0+cu128, CUDA 12.8, and torchvision 0.26.0, with OpenAI CLIP ViT-B/16. Each experiment is run with seeds 2, 3, 5, 7, and 11, and reported accuracies are averaged over these five seeds.

Comparisons. We use the zero-shot CLIP scores recorded with the complete WAVE runs and the StatA reproduction with both assessments disabled, denoted StatA-hard. The IEA-only variant retains ρ_{ik} and sets $q_k = 1$; the CEA-only variant retains q_k and sets $r_k = 1$. The complete WAVE model uses both. StatA-hard counts samples assigned to each predicted class rather than summing soft assignment probabilities, while the three assessment-enabled variants use soft counts. We therefore distinguish comparisons among assessment variants from comparisons against StatA-hard, which also change the counting rule. The completed StatA-soft control uses $\beta_k = s_k/(s_k + \alpha)$ and is analyzed in Sections G.5 and G.2. In addition, we compare against ADAPT (Zhang et al., 2025), rerun with the same evaluation harness under the same backbone, prompts, and test-distribution protocols (Section 6.3).

Metrics and reporting. We report top-1 classification accuracy in percent. For batch evaluation, we average task accuracies equally over 1,000 tasks, or use the single full-split accuracy for All. Tasks remain equally weighted when sample availability reduces their batch size. For stream evaluation, we first average the complete 128-sample batch accuracies within each stream and then average equally over 100 streams. Since retained batch sizes are equal, a stream's score equals accuracy on its retained samples; dropped tails are excluded. Each setting's aggregate is the arithmetic mean of the 14 per-dataset scores. A cross-setting summary gives equal weight to each of the 11 settings. We average each experiment over the five seeds. Accuracy differences, also written with %, are absolute differences on the percentage accuracy scale, not relative changes, and are calculated before rounding.

All setting-level comparisons use 14 datasets. The zero-shot column in Table 1 uses the WAVE evaluation records; the ADAPT records differ by at most 0.01 on these averages. The separate fixed-state diagnostics use task-bootstrap intervals, as specified in Section 6.5.

Assessment cost. Evaluating M candidate assignments per image against K text prototypes costs $O(NMKd)$ by direct computation and is performed once per batch. At each statistics update, centroid estimation costs $O(NKd)$ and centroid–text comparisons cost $O(K^2d)$. Candidate responses can be processed in chunks without retaining an $N \times M \times K$ tensor. All assessment operations act on frozen embeddings.

F Detailed Geometry of the Reference and Proposal Influence

The reference–proposal interpretation raises two geometric questions: what information can the text reference reveal about a visual proposal's source, and how do the resulting scores affect its influence? We connect residual entropy to an observable profile discrepancy, characterize directions hidden from that comparison, and analyze mean fusion and CEA. These are deterministic properties of the stated geometry at a fixed update state; they require neither correct pseudo-labels nor independent observations. All results use the idealized equations without numerical stabilizers, with $\beta_{\min} = 0$, and assume that normalized vectors are nonzero. Complete proofs are in Appendix D.

F.1 Residual Entropy as an Information-Geometric Discrepancy

For an evaluated pair (i, j), write $\mathbf{x} = \tau\boldsymbol{\delta}_{ij}$ and $D_{ij} = \log K - H_{ij}$. By construction, $\mathbf{1}^\top\mathbf{x} = 0$. Define the log-partition function $\psi(\mathbf{x}) = \log \sum_{k=1}^{K} e^{x_k}$ and its Bregman divergence

$$B_\psi(\mathbf{y}, \mathbf{x}) = \psi(\mathbf{y}) - \psi(\mathbf{x}) - \nabla\psi(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}). \tag{25}$$

The relation between entropy, log-partition functions, and convex duality is standard in exponential-family analysis (Wainwright & Jordan, 2008). Here it gives an interpretation of the particular residual distribution used by WAVE.

Proposition 2 (Information-geometric form and bounded-logit control). For the residual distribution in Eq. (6),

$$D_{ij} = \mathrm{KL}(\mathbf{p}_{ij} \| \mathbf{U}_K) = B_\psi(\mathbf{0}, \mathbf{x}) = \int_0^1 t\, \mathbf{x}^\top \nabla^2 \psi(t\mathbf{x}) \mathbf{x}\, dt. \tag{26}$$

If $\|\mathbf{x}\|_\infty \le R$ for $R \ge 0$, then

$$\frac{e^{-2R}}{2K} \|\mathbf{x}\|_2^2 \le D_{ij} \le \frac{e^{2R}}{2K} \|\mathbf{x}\|_2^2. \tag{27}$$

Thus, residual entropy measures a divergence of residual responses from the uniform reference. Proposition 1 gives its second-order behavior near agreement; Proposition 2 bounds it over any specified bounded-logit region. The constants can be loose when R is large, so Eq. (27) does not assert a useful global ranking by Euclidean discrepancy. Nor does it rank assignments by label correctness. The batch-relative transformation in Eq. (10) preserves the entropy ordering among evaluated pairs within a batch, but is not a calibrated divergence or correctness probability. These results do not apply to the imputed midpoint scores of unevaluated candidates.

F.2 The Subspace Visible to the Textual Reference

The text set determines which residual directions the assignment check can observe. Let $\mathbf{T}_P = [\mathbf{a}_1, \ldots, \mathbf{a}_K] \in \mathbb{R}^{d \times K}$, let $\mathcal{S}_T = \mathrm{range}(\mathbf{T}_P)$, and let $\mathbf{\Pi}_T$ be the orthogonal projector onto $\mathcal{S}_T$. For the projected residual $\mathbf{r}_{ij} = \mathbf{P}(\mathbf{f}_i - \mathbf{t}_j)$, Eq. (5) becomes $\boldsymbol{\delta}_{ij} = \mathbf{T}_P^\top \mathbf{r}_{ij}$.

Proposition 3 (Observable residual and spectral bounds). Assume $\mathrm{rank}(\mathbf{T}_P) > 0$. Let $\lambda_- > 0$ and λ_+ be the smallest positive and largest eigenvalues of $\mathbf{T}_P \mathbf{T}_P^\top$. Then

$$\lambda_- \|\mathbf{\Pi}_T \mathbf{r}_{ij}\|_2^2 \le \|\boldsymbol{\delta}_{ij}\|_2^2 \le \lambda_+ \|\mathbf{\Pi}_T \mathbf{r}_{ij}\|_2^2. \tag{28}$$

In particular, $\boldsymbol{\delta}_{ij} = \mathbf{0}$ if and only if $\mathbf{\Pi}_T \mathbf{r}_{ij} = \mathbf{0}$. Combining this result with Proposition 2, if $\|\tau \boldsymbol{\delta}_{ij}\|_\infty \le R$, then

$$\frac{\tau^2 e^{-2R} \lambda_-}{2K} \|\mathbf{\Pi}_T \mathbf{r}_{ij}\|_2^2 \le D_{ij} \le \frac{\tau^2 e^{2R} \lambda_+}{2K} \|\mathbf{\Pi}_T \mathbf{r}_{ij}\|_2^2. \tag{29}$$

The assignment check therefore assesses the component visible through the projected text set. A residual in $\mathcal{S}_T^\perp$ is invisible to this check, and weakly represented directions can produce small response discrepancies. Since $\sum_k \mathbf{a}_k = \mathbf{0}$, the rank is at most $\min(d-1, K-1)$; the reference need not identify the entire projected residual. If $\mathbf{T}_P = \mathbf{0}$, all residual profiles vanish and the check is uninformative. This characterization explains why profile agreement is weaker than embedding agreement and why the reference geometry matters. It does not establish that a dataset's empirical failures are caused by its spectrum. The aggregate check uses the original text decision geometry and is not reduced to this projected subspace.

F.3 Evidence-Controlled Fusion as a Regularized Decision

The assessments determine influence after the assignment-weighted estimate has been formed. Write $e_k = n_k^{\mathrm{rel}}$ and hold the current assignments, visual direction, and scores fixed. The following statement is a quadratic regularization interpretation of the mean-fusion step, using standard convex optimization principles (Boyd & Vandenberghe, 2004).

Proposition 4 (Fixed-state shrinkage and displacement). For $e_k \ge 0$ and $\alpha > 0$, the mean before output normalization in Eq. (15) is the unique minimizer

$$\bar{\boldsymbol{\mu}}_k = \operatorname*{argmin}_{\mathbf{m} \in \mathbb{R}^d} \left\{ e_k \|\mathbf{m} - \hat{\mathbf{v}}_k\|_2^2 + \alpha \|\mathbf{m} - \mathbf{t}_k\|_2^2 \right\}. \tag{30}$$

Let $Q_k = r_k q_k \in [0, 1]$, with $s_k > 0$. For $Q_k > 0$,

$$\beta_k = \frac{s_k}{s_k + \alpha / Q_k} \le \frac{s_k}{s_k + \alpha}, \qquad \|\bar{\boldsymbol{\mu}}_k - \mathbf{t}_k\|_2 = \beta_k \|\hat{\mathbf{v}}_k - \mathbf{t}_k\|_2. \tag{31}$$

For $Q_k = 0$, $\beta_k = 0$ and $\bar{\boldsymbol{\mu}}_k = \mathbf{t}_k$.

Quality discounting is thus equivalent to increasing the anchor strength separately for each class at the current state. It can attenuate the corresponding soft-count displacement but cannot increase it. This equivalence makes the global-α and Constant Gate comparisons in Sections 6.4 and G.5 necessary for interpreting the empirical gain. The displacement identity concerns the mean before normalization, not Euclidean displacement after normalization or the complete Gaussian decision rule.

Equation (30) is a subproblem interpretation, not a derivation of the entire algorithm from a fixed objective or a Bayesian posterior. The assessed mass is not a calibrated effective sample size. Retaining the original assignment weights avoids direct score-based reweighting of moments, but does not make their estimates unbiased. Changing β_k changes both fused statistics and later assignments; the covariance update also depends on the updated mean. The factorization $Q_k = r_k q_k$ remains a modeling choice evaluated experimentally, rather than a consequence of independent probabilities.

F.4 CEA: Margin along the Proposal Update

Extend the centroid margin to any vector by defining

$$g_k(\mathbf{x}) = \min_{j \neq k} \mathbf{x}^\top (\mathbf{t}_k - \mathbf{t}_j). \tag{32}$$

Then $g_k(\hat{\mathbf{v}}_k) = m_k$ and $\mathcal{C}_k = \{\mathbf{x} : g_k(\mathbf{x}) \geq 0\}$. The region is a convex cone because it is an intersection of homogeneous halfspaces.

Proposition 5 (Margin along the fusion segment). For any $\beta \in [0, 1]$, let $\mathbf{w}_\beta = (1 - \beta)\mathbf{t}_k + \beta\hat{\mathbf{v}}_k$ and $a_k = g_k(\mathbf{t}_k)$. Then

$$g_k(\mathbf{w}_\beta) \geq (1 - \beta)a_k + \beta m_k. \tag{33}$$

Unit-normalized text prototypes give $a_k \geq 0$. Consequently, if $m_k \geq 0$, the whole fusion segment belongs to $\mathcal{C}_k$. If $a_k > 0$ and $m_k < 0$, the sufficient condition

$$\beta \leq \frac{a_k}{a_k - m_k} \tag{34}$$

ensures $\mathbf{w}_\beta \in \mathcal{C}_k$. Whenever $\mathbf{w}_\beta \neq \mathbf{0}$, normalization preserves this region membership.

The proposition connects the aggregate check to the fusion decision: a centroid that favors a competitor requires a sufficiently small visual coefficient to retain its assigned text identity. At fixed s_k and r_k, WAVE's coefficient increases with $q_k = \sigma(\gamma m_k)$, so a smaller margin produces stronger attenuation. However, the implemented soft rule does not enforce Eq. (34); we do not claim unconditional identity preservation. Even satisfying this condition only concerns the fused prototype under the fixed text classifier. It does not certify the predictions of the adapted Gaussian classifier or agreement with true labels.

The propositions describe what the reference observes and how assessed evidence controls a visual proposal. They also delimit the cross-modal interpretation: a text reference can be internally consistent while semantically inaccurate, and a conservative update can suppress useful visual information. The results concern the geometry of scores and fixed-state fusion. Convergence of the feedback loop, calibration of correctness, and accuracy gains require separate evidence.

G Extended Experimental Analyses

G.1 Comparison with ADAPT

Among published methods, ADAPT (Zhang et al., 2025) is the closest competitor to our setting: it is backpropagation-free, operates on frozen CLIP features without target labels, and models class-conditional Gaussians, refining class means from high-confidence features accumulated in per-class knowledge banks. We therefore rerun ADAPT under the same backbone, prompts, datasets, and the 11 test-distribution settings of Table 4, using the same evaluation harness and logging (154 runs covering all dataset–setting pairs). Table 1 reports the results.

WAVE exceeds ADAPT in ten of the eleven settings. In B64, 1–4, ADAPT falls 6.45% below zero-shot CLIP, while WAVE exceeds ADAPT by 16.54%. ADAPT retains an advantage only on Online Low (66.01% vs. 65.10%). These results highlight the benefit of assessing visual proposals in concentrated batches.

G.2 Holding the Counting Rule at Soft Assignments

The StatA-soft comparison uses the same soft-count rule as the assessment-enabled configurations and covers all 14 datasets and 11 settings. Figure 6 in Appendix B shows the full dataset–setting matrix against both StatA-hard and StatA-soft. In the three B64 settings, WAVE's mean gains over StatA-soft are 5.36, 2.93, and 0.90%, respectively. Gains on the Low, Medium, High, and Separate stream settings are 0.28, 3.17, 4.56, and 6.00%.

G.3 Do Source and Centroid Assessments Work Together?

We next compare the two assessments within the soft-count configurations. The IEA-only and CEA-only columns of Table 1 show that combining IEA with CEA outperforms either assessment alone in all eleven settings. The incremental gains are particularly clear on Online High, Separate, and B64 1–4, where the complete method exceeds the stronger single-assessment variant by 1.98, 2.77, and 2.15%, respectively. These results support combining the checks for these adaptation settings, without implying that their product is a uniformly better estimator of label correctness.

Neither assessment dominates across all settings. CEA-only has higher mean accuracy than IEA-only in eight settings, while IEA-only is stronger in B64 1–4, B1000 50–100, and All. The complete model's margin over the stronger single assessment is smallest on Online Low ($+0.06$), B64 5–25 ($+0.11$), and All ($+0.50$), so the benefit of combining the checks is positive throughout but small in these settings.

The two assessments also differ at the dataset level. On Oxford Flowers, IEA-only decreases accuracy relative to StatA-hard in all 11 settings, with a mean change of -0.86%. CEA-only improves in nine settings and has a mean change of $+0.44\%$. This difference supports their distinct empirical behavior. It does not by itself establish that either score estimates label correctness, or that the joint gain is independent of the overall visual influence; we examine the latter factor in Sections 6.4 and G.5.

G.4 How Much Does Overall Shrinkage Explain?

The semantic checks attenuate visual influence. An important alternative explanation is that the method benefits mainly from being more conservative, irrespective of which class receives which weight. Increasing the anchor strength α in StatA-hard also reduces visual influence. We first examine this explanation with a StatA-hard sweep over $\alpha \in \{0.5, 1, 2, 2.5, 3, 3.5\}$. Within each setting, we select the value with the highest mean accuracy across the 14 evaluated datasets. Because this selection uses the reported evaluation accuracies, it is an **oracle selection within the tested grid**, used to examine the sensitivity of the comparison. It is not a label-free hyperparameter-selection protocol. The $\alpha = 1$ rows in Tables 1 and 2 come from separate component and anchor-sweep experiments, respectively, accounting for their small numerical differences.

Stronger anchor regularization substantially narrows the average gap. A single shared value $\alpha = 3.5$ gives a cross-setting macro accuracy of 68.95%, compared with 69.65% for WAVE. Selecting α separately for each setting raises StatA-hard to 69.18%. WAVE therefore exceeds these two summaries by 0.70 and 0.47%, respectively, compared with its 2.78% advantage over StatA-hard. These results show that the regularization strength accounts for a substantial portion of the default-baseline gap.

WAVE remains higher in seven of the 11 settings under the grid-selected comparison, including all three B64 settings. Its margins of 1.01–1.77% in these small-batch settings identify a useful focus for further controlled evaluation. However, seven settings select the largest tested value, $\alpha = 3.5$, so the sweep does not establish an optimum over a wider range. No results outside the reported grid are currently available.

G.5 Does Class-Specific Quality Allocation Add Value?

To examine class-specific allocation, the Constant Gate control preserves count dependence while removing class-specific variation in the quality multiplier. At each update, on the Constant branch's own current assignments $\mathbf{Z}^C$, we compute the WAVE-formula target weights $\tilde{\beta}_k(\mathbf{Z}^C)$. We choose a

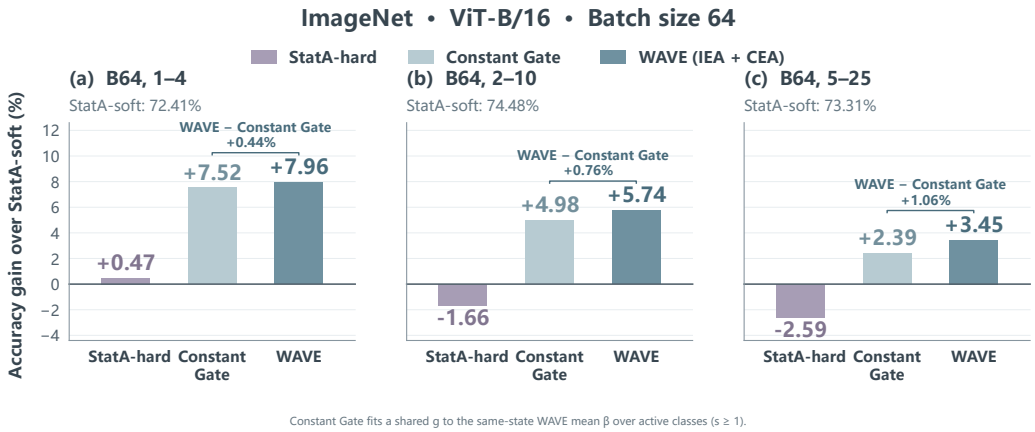


Figure 7: ImageNet B64 performance relative to StatA-soft. Each Constant setting has 1,000 tasks. Both Constant Gate and WAVE improve over StatA-soft, and WAVE exceeds Constant Gate by 0.44%, 0.76%, and 1.06% for the three requested class ranges, respectively.

shared $g \in [0, 1]$ by bisection such that

$$\beta_k^C = \frac{g s_k^C}{g s_k^C + \alpha}, \qquad \frac{1}{|\mathcal{A}|} \sum_{k \in \mathcal{A}} \beta_k^C = \frac{1}{|\mathcal{A}|} \sum_{k \in \mathcal{A}} \tilde{\beta}_k(\mathbf{Z}^C), \quad \mathcal{A} = \{k : s_k^C \geq 1\}. \tag{35}$$

The implementation uses 48 bisection steps, includes numerical stabilizers, and falls back to all classes if $\mathcal{A}$ is empty. Matching uses no labels. It matches the active-class mean to the WAVE formula at the Constant branch's state, not to the independently evolving WAVE run. Class weights remain count-dependent even though g is shared.

For requested class ranges 1–4, 2–10, and 5–25, StatA-soft obtains 72.41%, 74.48%, and 73.31% accuracy. Constant Gate obtains 79.92%, 79.46%, and 75.70%, while WAVE obtains 80.37%, 80.22%, and 76.76%. Figure 7 shows WAVE exceeding Constant Gate by 0.44, 0.76, and 1.06%. The substantial improvement of both gated methods over StatA-soft indicates that the overall attenuation level contributes to performance. WAVE's additional gains are consistent with value from class-specific quality allocation, subject to the branch-local matching rule and unverified pairing across runs. We do not infer statistical significance from these aggregate scores.

G.6 What Do the Scores Measure Empirically?

The reference–proposal loop depends on agreement with a shared text reference. We therefore examine whether IEA and CEA reflect label-based quality before interpreting their adaptation gains. We evaluate the assessments before adaptation, fixing $z_{ik} = \hat{y}_{ik}$ for every method. This separates their initial diagnostic behavior from changes induced by iterative inference. We sample 100 tasks for each feasible dataset–setting combination on ImageNet and DTD, using the six batch configurations in Table 4. DTD has 47 classes, so its requested 50–100-class setting is skipped, giving 11 feasible combinations. Actual batches can be smaller than requested when the selected classes contain insufficient samples. Labels are used for task construction and retrospective evaluation, but not to compute assessment scores.

For class-level analysis, we measure soft purity: the sum of assignment probabilities contributed by samples whose true label is class k, divided by the soft class count s_k:

$$\text{purity}_k = \frac{\sum_i z_{ik} \mathbb{1}[y_i = k]}{s_k}, \qquad s_k > 0. \tag{36}$$

This soft purity evaluates the class estimate's supporting assignments. The class-level plots use r_k for IEA, q_k for CEA, and $Q_k = r_k q_k$ for their product. Figures 4 and 5 in Appendix B report the two displayed B64 settings (1–4 and 5–25 candidate classes) for both datasets. Intervals in the reliability curves use 1,000 task-level bootstrap resamples.

Assignment scores have limited correctness-detection ability. Figure 4 bins samples by their assessment score for the zero-shot predicted class and reports accuracy and correctness-detection AUROC. The residual score obtains AUROCs of 0.530 and 0.523 on ImageNet and 0.627 and 0.631 on DTD. Maximum softmax probability obtains 0.815, 0.811, 0.825, and 0.800, respectively. Thus, residual entropy is not a stronger general-purpose correctness ranking in these settings. Its geometric interpretation establishes compatibility with a text profile, while its association with true labels depends on the dataset. These comparisons do not test conditional value after matching prediction confidence.

Aggregate scores provide partial support for class reliability. Figure 5 relates scores to soft purity within the count strata $1 \leq s_k < 4$, $4 \leq s_k < 16$, and $s_k \geq 16$. In several nonsaturated strata, the aggregate score has a clearer positive association with purity than the assignment score. For example, in ImageNet B64 1–4 with $4 \leq s_k < 16$, Spearman correlations for assignment, aggregate, and product scores are 0.07, 0.44, and 0.31. In DTD B64 5–25 with $1 \leq s_k < 4$, they are 0.38, 0.50, and 0.53. Saturated high-purity strata and sparse groups exhibit weaker or unstable relationships. The diagnostics provide partial support for aggregate assessment and limited, setting-dependent evidence of complementarity. Final-accuracy gains from combining the scores should therefore be distinguished from uniformly superior reliability estimation. In Figure 5, the blank positions in the third column of the second and fourth rows correspond to ImageNet and DTD in the B64, 5–25 setting. No classes in the collected tasks had soft count $s_k \geq 16$, so the statistics for that stratum are undefined and the positions are left blank. This absence reflects an unobserved stratum, not missing experiments or a plotting error.

G.7 When Does the Textual Reference Restrict Useful Adaptation?

The textual witness provides a reference tied to the pretrained class geometry, so its assessments need not favor useful adaptation in every dataset. Figure 6 in Appendix B reports the complete dataset–setting matrix of accuracy changes against both baselines. The complete model improves on ImageNet-A and ImageNet-R in every setting, with mean gains of 7.31 and 4.48% over StatA-hard. Oxford Flowers and EuroSAT exhibit more limited or negative changes.

A pronounced failure occurs on EuroSAT in B64 5–25: accuracy decreases from 45.10% for StatA-hard to 35.07% for WAVE, a loss of 10.03%. Both single-assessment variants also decrease accuracy in this setting. Together with the All results, this case is consistent with the possibility that stronger reliance on the textual anchor suppresses useful visual adaptation. Determining whether the cause is an inaccurate text reference, a score that penalizes valid visual variation, or excessive attenuation requires diagnostics beyond final accuracy. The current results establish this failure mode without identifying a unique cause.