ORDINAL COMPATIBILITY FOR PREDICTION AND ADAPTATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Online test-time adaptation gives each unlabeled sample two roles: it must be classified immediately, and it can contribute to the model state used for subsequent predictions. These roles call for different uses of the same evidence. A sample may require prediction correction while warranting a conservative update to the evolving class distributions. We introduce **P**rediction and **A**daptation via **C**ompatibility **E**vidence (PACE), a framework that uses ordinal compatibility between fixed semantic knowledge and adaptive distributional estimates to guide both decisions. For each sample, we compare the class rankings induced by text prototypes and online Gaussian models, and convert their rank displacements into a shared compatibility vector. At the class level, compatibility provides a log prior that adjusts dynamic predictions. At the sample level, its posterior-weighted expectation produces a continuous gate that controls the total responsibility assigned to Gaussian updates. This design connects prediction correction and selective adaptation through one relational signal while keeping their operations distinct. PACE requires no encoder updates or backpropagation. Under dataset-specific oracle tuning on full target streams, the full method improves top-1 accuracy over its rank-free counterpart in all 21 evaluations, with an average gain of 1.09 percentage points. The ordinal-only variant achieves a 1.02-point mean gain, identifying compatibility as the main source of improvement. The gate reduces the fraction of update mass assigned to incorrect classes in five of six diagnostic domains. Matched-mass controls favor compatibility selection over uniform attenuation in three of four domains, while a same-state diagnostic shows that corrected predictions can still carry substantial wrong-class allocation mass.

1 INTRODUCTION

Online test-time adaptation enables vision-language models to learn from unlabeled inputs encountered during deployment. For models such as CLIP (Radford et al., 2021), incoming examples provide information about a target environment that fixed text prototypes alone may not capture. Distributional adaptation offers an efficient way to accumulate this information: class statistics are updated along the test stream and used to infer the labels of subsequent samples, without optimizing the encoders through backpropagation (Han et al., 2026; Zhang et al., 2026; Guan et al., 2026). In this setting, every test sample has two roles. It is an instance to classify now, and a potential source of evidence for the state that will classify future inputs.

These roles require different decisions. Prediction correction affects one output, whereas updating a class distribution can change many later outputs. A normalized class posterior allocates a sample across classes but does not determine its total update weight. Reliable adaptation therefore needs to control both current prediction and evidence admitted to the evolving state (Lee et al., 2024; Lee, 2026; Zanella et al., 2025); correlated streams can compound poor updates (Gong et al., 2022; Wang et al., 2022). Prior methods often use auxiliary evidence for only one role (Figure 1) (Yu et al., 2024).

Frozen text prototypes and online Gaussian models offer complementary views of each test sample. Both rank the candidate classes, even when their score scales differ. We turn each class's rank displacement into an ordinal compatibility value. Agreement supplies a persistent semantic reference to the evolving distributional state without assuming that either source is always correct (Martin et al., 2024).

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.

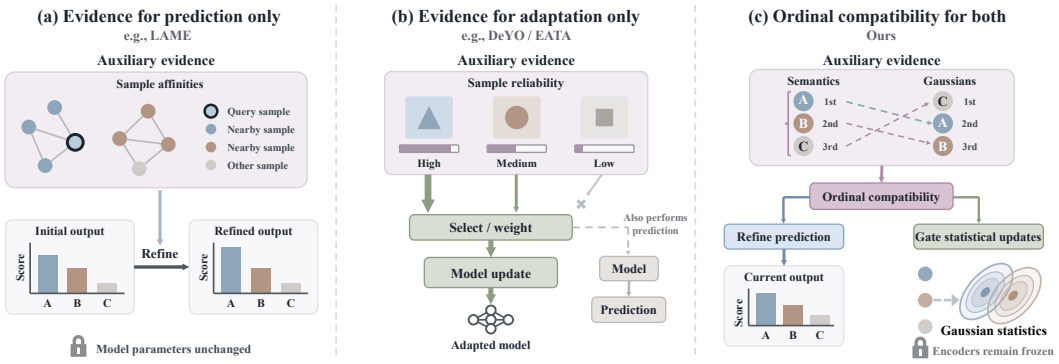


Figure 1: Auxiliary evidence in online adaptation. (a) Methods such as LAME use affinities to refine only the current prediction. (b) DeYO and EATA select or weight only model updates. (c) PACE compares fixed semantic and adaptive Gaussian class rankings once: classwise agreement corrects the current prediction, while its posterior-weighted sample aggregate controls how much evidence enters subsequent Gaussian updates.

The same compatibility signal serves two granularities. Its classwise values penalize dynamic predictions that disagree with the semantic ranking; their posterior-weighted aggregate gates the sample's total contribution to future Gaussian statistics. Prediction and update can therefore respond differently to the same input. PACE implements these operations without changing either encoder: it computes compatibility from the pre-update state, emits the current prediction, and then commits gated responsibilities to the state. The gate scales a separately normalized class allocation, so a corrected label is never automatically fed back as an update target.

Under dataset-specific oracle tuning on the full target streams, PACE improves over its rank-free counterpart on all 21 evaluations by 1.09 percentage points on average. The ordinal-only variant gains 1.02 points, identifying compatibility as the main contributor. Paired prediction changes, update-responsibility diagnostics, and matched-mass controls test the two mechanisms; the latter also reveal a domain where the gate hurts. These results characterize mechanism behavior under the stated oracle protocol, not independently tuned deployment performance.

Our contributions are:

- A shared ordinal compatibility measure between frozen semantic prototypes and adaptive Gaussian class models.
- A classwise prediction prior and a samplewise responsibility gate that independently control prediction and online adaptation.
- Component and diagnostic evidence across 21 evaluations, including matched-mass controls and the failure of update gating on SUN397.

2 METHOD

PACE uses the relationship between fixed semantic knowledge and evolving class distributions to determine how test evidence enters prediction and adaptation. Its central object is a classwise compatibility vector computed from their relative class rankings. Figure 2 summarizes the framework: the two rankings are compared once, and the resulting compatibility enters a classwise correction in the prediction lane and a samplewise gate in the adaptation lane, with the timing constraint that the prediction always precedes the state update. We first define this shared evidence, then describe its use as a prediction prior and as a gate on online updates.

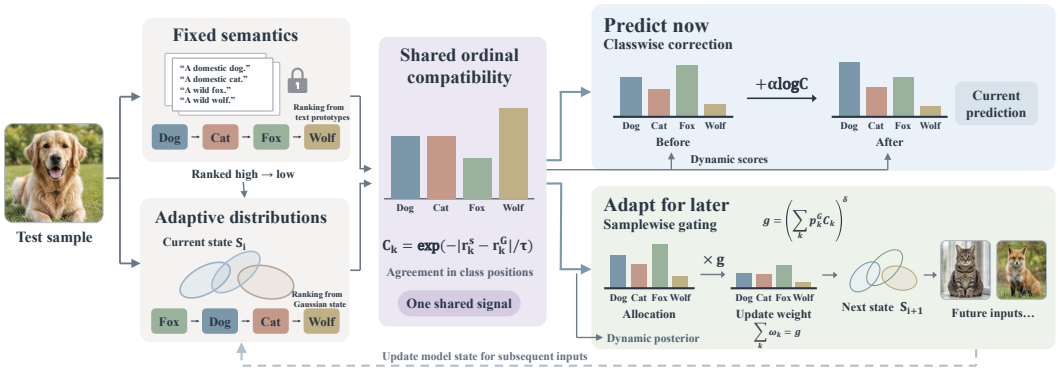


Figure 2: PACE computes one ordinal compatibility vector from fixed semantic and adaptive Gaussian rankings before updating the state. In the prediction branch, a classwise log prior corrects dynamic scores before fusion with zero-shot logits. In the adaptation branch, posterior-weighted compatibility yields a continuous gate on normalized class allocation. Prediction precedes the Gaussian update, and both encoders remain frozen.

2.1 Online Setting and Distributional State

For an unlabeled stream $\{x_i\}_{i=1}^T$ over K classes, frozen encoders yield normalized image features $z_i \in \mathbb{R}^D$ and class prototypes $\{t_k\}_{k=1}^K$. We compute semantic similarity and the zero-shot posterior:

$$s_{ik}^{\mathrm{S}} = z_i^\top t_k, \qquad p_{ik}^{\mathrm{CLIP}} = \mathrm{softmax}_k(\kappa s_i^{\mathrm{S}}), \tag{1}$$

where $\kappa > 0$ is the pretrained logit scale. Before processing x_i, the adaptive state is $\mathcal{S}_i = \{\mu_{ik}, \Sigma_{ik}, n_{ik}\}_{k=1}^K$, comprising class means, covariance estimates, and effective counts accumulated from earlier inputs. Following distributional adaptation (Han et al., 2026), the dynamic score is a Gaussian log likelihood,

$$s_{ik}^{\mathrm{G}} = -\frac{1}{2}(z_i - \mu_{ik})^\top \widetilde{\Sigma}_{ik}^{-1}(z_i - \mu_{ik}) - \frac{1}{2}\log\left|\widetilde{\Sigma}_{ik}\right|, \tag{2}$$

up to a class-independent constant. Here $\widetilde{\Sigma}_{ik}$ denotes the positive-definite covariance used for scoring, including the base estimator's regularization or covariance sharing; the construction follows the Gaussian discriminant tradition adapted to frozen features (Wang et al., 2024a; Haselhoff et al., 2024), and the challenge of adapting class covariances from scarce online samples is also studied in exemplar-free incremental learning (Goswami et al., 2023; Rypeść et al., 2024). PACE acts on the resulting class scores and update weights; it does not require optimizing either encoder. All quantities used to predict x_i are evaluated with $\mathcal{S}_i$, and its update produces $\mathcal{S}_{i+1}$ only after the prediction is emitted.

2.2 Shared Ordinal Compatibility

The semantic and Gaussian scores reflect different sources of knowledge and need not have comparable scales. We characterize their relationship through the position assigned to each class. Let R_{ik}^{S} and R_{ik}^{G} be descending ranks, with rank 1 assigned to the highest score. For $K \geq 2$, we normalize the ranks and compute their displacement:

$$r_{ik}^{\mathrm{S}} = \frac{R_{ik}^{\mathrm{S}} - 1}{K - 1}, \qquad r_{ik}^{\mathrm{G}} = \frac{R_{ik}^{\mathrm{G}} - 1}{K - 1}, \qquad d_{ik} = \left|r_{ik}^{\mathrm{S}} - r_{ik}^{\mathrm{G}}\right|. \tag{3}$$

A consistent deterministic rule resolves ties. We map the displacement to compatibility,

$$C_{ik} = \exp\left(-\frac{d_{ik}}{\tau}\right), \qquad \tau > 0. \tag{4}$$

The vector C_i retains the classwise structure of agreement. A class receives high compatibility when the two sources place it at similar relative positions, and lower compatibility as these positions

diverge. Consequently, agreement among low-ranked classes does not by itself identify a likely label; compatibility must be interpreted together with the class scores.

Both branches of PACE consume this same vector, computed once from the unmodified semantic and Gaussian rankings. In particular, the Gaussian ranking in Eq. equation 3 precedes the geometric and ordinal corrections below. This gives prediction and adaptation a common reference for evaluating the current distributional interpretation.

2.3 Classwise Evidence for Prediction

Geometric refinement. We first refine the dynamic scores using the geometry of the feature relative to a class anchor. The anchor combines fixed text semantics with the current class mean:

$$b_{ik} = \beta t_k + (1-\beta)\mu_{ik}, \qquad e_{ik} = 1 - \cos^2(z_i, b_{ik}), \qquad h_{ik} = s^{\mathrm{G}}_{ik} - \gamma e_{ik}, \tag{5}$$

where $\beta \in [0,1]$ and $\gamma \geq 0$. For a nonzero anchor, e_{ik} is the squared component of the normalized feature orthogonal to the anchor direction. It provides a geometric adjustment to the Gaussian score, echoing observation-side anchor corrections studied in representation learning (Boudiaf et al., 2022; Ge et al., 2024). We denote the resulting posterior by $p^{\mathrm{G}}_i = \mathrm{softmax}(h_i)$. Setting $\gamma = 0$ removes this auxiliary refinement while retaining the ordinal mechanism.

Compatibility prior. We incorporate classwise compatibility as a log prior on the dynamic scores:

$$\widetilde{h}_{ik} = h_{ik} + \alpha \log C_{ik} = h_{ik} - \frac{\alpha}{\tau} d_{ik}, \qquad \alpha \geq 0. \tag{6}$$

This correction reduces the dynamic contribution of classes with larger rank disagreement. Equivalently, the corrected dynamic posterior is

$$q_{ik} = \frac{p^{\mathrm{G}}_{ik} C^{\alpha}_{ik}}{\sum_{j=1}^{K} p^{\mathrm{G}}_{ij} C^{\alpha}_{ij}}. \tag{7}$$

Thus, the prior changes the relative support for candidate classes according to their compatibility with the fixed semantic reference, in the same multiplicative spirit as prior-correction and label-shift rules that reweight posteriors toward a reference distribution (Park et al., 2023). The term *prior* refers to this multiplicative reweighting; C_{ik} is an agreement measure and is not assumed to be a calibrated probability of correctness.

The final prediction combines the zero-shot logits with the corrected dynamic scores:

$$\ell_{ik} = \kappa s^{\mathrm{S}}_{ik} + \lambda_i \widetilde{h}_{ik}, \qquad \lambda_i = \min\left(\rho \frac{\bar{n}_i}{V}, \eta\right), \qquad \widehat{y}_i = \arg\max_k \ell_{ik}, \tag{8}$$

where $\bar{n}_i = K^{-1} \sum_k n_{ik}$, V is the number of views used to normalize the estimator's counts, ρ controls the growth of the dynamic contribution, and η caps it. The fusion therefore responds to the effective evidence accumulated in the online state. This state-count dependence parallels non-parametric test-time classifiers whose decisions likewise track the support accumulated on the stream (Iwasawa & Matsuo, 2021; Zhang et al., 2023). Setting $\alpha = 0$ disables the ordinal prediction correction.

2.4 Samplewise Evidence for Adaptation

Prediction uses compatibility to adjust relative class support. Adaptation additionally requires a decision about the total weight of the sample in the evolving statistics. We aggregate the same compatibility vector under the geometry-adjusted posterior:

$$\overline{C}_i = \mathbb{E}_{k \sim p^{\mathrm{G}}_i}[C_{ik}] = \sum_{k=1}^{K} p^{\mathrm{G}}_{ik} C_{ik}, \qquad g_i = \overline{C}_i^{\,\delta}, \qquad \delta \geq 0. \tag{9}$$

The expectation emphasizes agreement on classes that the current distributional model considers plausible. Disagreement among these classes reduces $\overline{C}_i$ and, through g_i, the sample's contribution

to adaptation. Since $C_{ik} \in [\exp(-1/\tau), 1]$, the gate satisfies $g_i \in [\exp(-\delta/\tau), 1]$. It provides continuous attenuation, with $\delta = 0$ recovering an ungated update. Posterior aggregation of reliability in the same spirit appears in confidence-based adaptation and failure prediction (Zhang et al., 2026; Zhu et al., 2022).

We separately specify a normalized class allocation a_i. The supported choices can be expressed as

$$a_{ik} = \zeta p_{ik}^{\mathrm{CLIP}} + (1 - \zeta) p_{ik}^{\mathrm{G}}, \qquad \zeta \in \{0, 0.5, 0.8, 1\}. \tag{10}$$

These choices allocate evidence using the Gaussian posterior, an equal mixture, a CLIP-dominant mixture, or the CLIP posterior, respectively. For any such allocation, PACE assigns the update responsibility

$$\omega_{ik} = g_i a_{ik}, \qquad \sum_{k=1}^{K} \omega_{ik} = g_i. \tag{11}$$

Eq. equation 11 separates two decisions explicitly: a_i determines where evidence is allocated, while g_i determines how much evidence is admitted. Multiplying by the gate preserves the relative allocation within a sample. Renormalizing ω_i across classes would cancel the gate, so the Gaussian estimator consumes the unnormalized responsibilities.

The posterior in both Eqs. equation 9 and equation 10 is p_i^{G}, before the rank prior. Neither the corrected posterior q_i nor the final prediction $\widehat{y}_i$ is fed back as the allocation. This construction allows the prediction prior to adjust the current decision while the update gate independently controls its influence on subsequent states, with both operations grounded in C_i.

2.5 Responsibility-Weighted Online Updates

The Gaussian estimator consumes the unnormalized responsibilities in Eq. equation 11. For each class, its effective count increases by ω_{ik}, and the same weight enters the recursive mean and covariance updates. Thus, g_i changes the total admitted evidence while a_i distributes it across classes. Predictions use $\mathcal{S}_i$; only afterward do these updates produce $\mathcal{S}_{i+1}$. The recursions, multiple-view extension, and complete processing order are given in Appendix A.

At a fixed state, changing the prediction-prior strength α leaves update responsibilities unchanged; changing the gate power δ alters update mass and possibly later states and fusion weights. With geometry disabled, $(\alpha, \delta) = (0, 0)$ is Rank-free, and enabling either parameter isolates one ordinal operation. Compatibility is invariant to strictly increasing transformations of either score vector, although the full predictor and gate also depend on score magnitudes. The fixed-state condition under which gating reduces incorrect-class responsibility is derived in Appendix A.

3 Related Works

Test-time adaptation of image classifiers. Parameter-updating methods adapt to unlabeled data through entropy minimization or selective gradients: Tent (Wang et al., 2021), MEMO (Zhang et al., 2022), SAR (Niu et al., 2023), and DeYO (Lee et al., 2024) exemplify this line, whose reliability limits have been analyzed by Press et al. (2024). Parameter-free distributional adaptation instead fits online class Gaussians (Boudiaf et al., 2022). In classifier TTA, reliability is generally inferred from the adapting model's entropy, gradients, or features. Agreement with a frozen semantic reference offers a different source of evidence: a corrected current label need not imply a safe contribution to future class statistics. PACE follows the Gaussian approach but uses this cross-source agreement to assess both predictions and update weights.

Cache-based vision-language adaptation. Frozen CLIP models can use auxiliary test features without encoder optimization. CALIP refines outputs with parameter-free attention (Guo et al., 2023), while ZERO uses classwise Gaussian models of cached features to reweight predictions (Wang et al., 2024a). TDA, Dual Memory Networks, and BoostAdapter maintain feature caches with different storage and retrieval rules (Karmanov et al., 2024; Zhang et al., 2024b;a). These methods use test evidence through the examples retained and the votes they cast. PACE instead compares the semantic and evolving distributional class rankings before deciding how to correct the current output and how much responsibility to admit into future statistics.

Distributional and prompt adaptation. DOTA updates Gaussian class statistics online and fuses their posterior with zero-shot logits (Han et al., 2026); ADAPT uses probabilistic Gaussian alignment without backpropagation (Zhang et al., 2026), and multimodal Gaussian calibration extends this family (Xu et al., 2026). These methods motivate PACE's distributional state, but do not derive a shared classwise and samplewise control signal from its agreement with fixed semantics. StatA addresses non-i.i.d. streams by adjusting adaptation magnitude at batch level (Zanella et al., 2025); PACE controls each sample's update mass. Prompt-based methods such as TPT and DiffTPT tune text prompts at test time (Shu et al., 2022; Feng et al., 2023), building on prompt learning before deployment (Zhou et al., 2022a). PACE holds the text branch fixed as the reference against which distributional evidence is assessed.

4 Experiments

We examine how ordinal compatibility supports current prediction and the accumulation of evidence for future inputs. Our evaluation combines six component configurations across 21 datasets and domains with paired prediction outcomes, reliability statistics, update-gate distributions, matched-mass gate controls, a same-state prediction–update diagnostic, and a matched efficiency comparison on VisDA-2017. A windowed temporal diagnostic tracks how gains evolve along the stream. We also place the results alongside published test-time adaptation baselines on the overlapping classification benchmarks. The main question is whether the two uses of compatibility account for the observed gains, and how much the auxiliary geometry term adds.

4.1 Experimental Setup

We evaluate CLIP ViT-B/16 (Radford et al., 2021) on 21 streams: ten classification datasets, four Office-Home domains (Venkateswara et al., 2017), VisDA-2017 (Peng et al., 2017), and six DomainNet domains (Peng et al., 2019). The core experiments use one image feature, dataset-specific text prompts, batch size one, and a fixed canonical order. Six domains receive detailed diagnostics. Dataset splits, prompts, preprocessing, hardware, and the separate multi-view ImageNet-A/R/V2 protocol are specified in Appendix B and Appendix D.

Evaluation scope. We tune parameters separately for each dataset using labels from its complete target test stream after replay. No independent validation split or held-out retest is used. These are oracle-tuned mechanism results, not estimates of independently selected deployment hyperparameters. Candidates start from fresh Gaussian states, share cached features and stream order within comparisons, and never use labels in per-sample prediction or updates. Appendix E records selected settings.

Rank-free disables the ordinal prior, gate, and geometry. Prediction and Update enable the corresponding ordinal operation alone; Pure Rank enables both, Geometry Only enables geometry alone, and PACE enables all three. We report top-1 accuracy and unweighted means across evaluations; differences are percentage points (pp). Full metric definitions and diagnostic protocols appear in Appendix B and Appendix C.

4.2 Contributions of Ordinal Compatibility and Geometry

Table 1 separates the two ordinal operations from geometry. Prediction and Update improve mean accuracy over Rank-free by 0.435 and 0.703 pp, respectively. Pure Rank gains 1.019 pp, and full PACE gains 1.088 pp; both improve all 21 evaluations. The ordinal gain therefore accounts for most of the improvement.

Adding geometry to Pure Rank contributes 0.070 pp on average, with mixed effects by dataset. EuroSAT has the largest PACE gain (+6.44 pp), but the mean gain remains 0.821 pp without it. The Prediction and Update effects overlap: their joint gain is 0.119 pp smaller than their summed individual gains on average. These comparisons remain descriptive under target-stream oracle tuning.

Table 1: Core component comparison under dataset-specific oracle tuning. Accuracy is in percent. P, U, and G denote Prediction, Update, and Geometry Only. Pure Rank enables P and U; PACE additionally enables geometry. Bold indicates the highest value at the displayed precision, including ties. The final row averages the 21 evaluations. All components are disabled in Rank-free.

Dataset / domain	Rank-free	P	U	Pure Rank	G	PACE
Caltech101	94.24	94.28	94.30	94.30	94.26	**94.32**
DTD	49.00	50.24	50.50	50.96	49.16	**51.06**
EuroSAT	66.78	67.90	71.91	73.04	67.04	**73.22**
FGVC Aircraft	26.94	26.97	27.06	27.15	26.98	**27.18**
Food101	86.04	87.01	87.06	**87.08**	86.07	87.06
Oxford Flowers	76.65	76.53	76.93	76.87	76.70	**76.94**
Oxford Pets	92.42	92.83	92.47	92.87	92.44	**92.91**
Stanford Cars	69.53	69.73	69.56	**69.74**	69.51	69.71
SUN397	68.92	68.93	68.78	69.03	68.88	**69.56**
UCF101	73.04	74.04	73.74	74.12	73.13	**74.20**
Office-Home Art	85.13	85.62	85.73	86.32	85.27	**86.44**
Office-Home Clipart	70.29	70.56	70.80	70.95	70.36	**71.00**
Office-Home Product	91.46	91.26	91.77	91.99	91.57	**92.07**
Office-Home Real World	91.60	91.81	91.74	92.38	91.70	**92.47**
VisDA-2017	86.94	87.50	87.82	88.22	87.09	**88.25**
DomainNet Clipart	74.01	74.59	74.53	**74.71**	74.07	74.68
DomainNet Infograph	52.49	52.61	53.11	53.18	52.57	**53.24**
DomainNet Painting	70.85	71.39	71.14	71.56	70.94	**71.63**
DomainNet Quickdraw	18.03	19.12	18.95	19.44	18.21	**19.50**
DomainNet Real	84.83	85.39	85.71	85.96	84.97	**85.97**
DomainNet Sketch	66.28	66.30	66.62	**67.00**	66.25	66.92
Mean	71.213	71.648	71.916	72.232	71.294	**72.301**

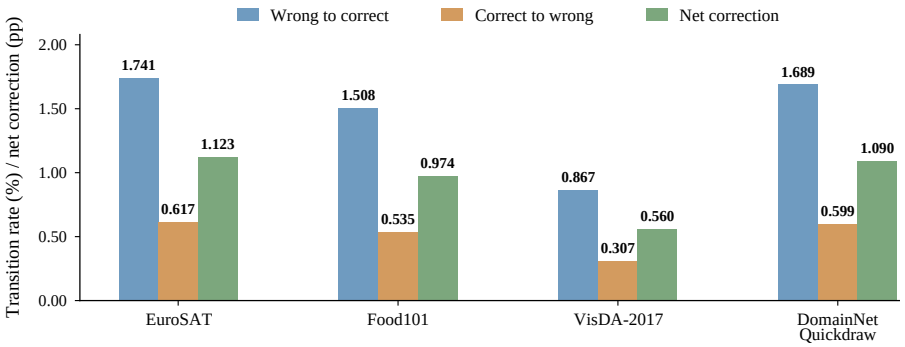


Figure 3: Paired prediction changes relative to Rank-free: beneficial, harmful, and net correction rates.

4.3 Prediction and Update Mechanisms

Against Rank-free at the same sample indices, the Prediction prior changes 3,012 wrong outputs to correct and 1,119 correct outputs to wrong across 331,415 samples, a pooled net correction of 0.571 pp. Nineteen of 21 evaluations have positive net correction. Compatibility is also positively associated with Rank-free prediction correctness in all six diagnostic domains (mean AUROC 0.589), but it is not a calibrated correctness probability. Figure 3 shows the paired changes; full counts, decile plots, and reliability bins are in Appendix C.

Table 2 shows that the gate reduces the fraction of admitted responsibility assigned to incorrect classes (WRM) in five of six diagnostic domains; mean WRM falls from 0.223 to 0.203. SUN397 is the exception: WRM rises from 0.245 to 0.247 and Update accuracy falls by 0.14 pp. At matched total update mass (Table 3), the compatibility gate exceeds a constant gate in accuracy on EuroSAT, VisDA-2017, and DomainNet Quickdraw, while SUN397 remains unfavorable. Confidence-based controls and a common fusion-schedule control qualify this selection effect (Appendix C).

Table 2: Update-gate diagnostics in six domains. WRM is the admitted mass assigned to wrong classes; Gate(c/w) are mean gates by allocation correctness, and M_{eff} is mean admitted mass. Base uses no gate; both configurations disable geometry.

Dataset / domain	Base WRM	Update WRM	ΔWRM (pp)	Gate(c)	Gate(w)	M_{eff}
EuroSAT	0.286	0.241	−4.510	0.774	0.615	0.728
Food101	0.118	0.108	−1.000	0.858	0.779	0.849
VisDA-2017	0.132	0.119	−1.300	0.841	0.746	0.828
DomainNet Quickdraw	0.415	0.376	−3.900	0.728	0.618	0.683
SUN397	0.245	0.247	+0.200	0.802	0.812	0.804
DomainNet Real	0.142	0.129	−1.300	0.839	0.752	0.826
Mean	0.223	0.203	−1.968	0.807	0.720	0.786

Table 3: Matched-mass update control: accuracy gains over Rank-free (pp) and wrong responsibility mass (WRM). Full controls and late-stream results appear in Table 11.

Domain	Constant	MaxProb	PACE gate	WRM(no)	WRM(PACE)
EuroSAT	+1.35	+3.54	+5.13	0.286	0.241
VisDA-2017	+0.21	+0.91	+0.88	0.132	0.119
DomainNet Quickdraw	+0.19	+0.67	+0.92	0.415	0.376
SUN397	+0.03	+0.07	−0.14	0.245	0.247

As a separate reliability check, we compare the posterior-weighted compatibility $\overline{C}_i$ with Rank-free correctness using labels only after replay. Correct samples have higher mean compatibility than incorrect samples in all six diagnostic domains; AUROC ranges from 0.567 to 0.619, and Spearman correlation from 0.093 to 0.187 (Table 10). These modest associations justify testing compatibility as a gate but do not make it a calibrated probability of correctness. The bin-level plots in Appendix C expose the low accuracy that can persist even at high compatibility.

At the same pre-update state, a joint diagnostic (Table 4) separates the two roles: samples corrected by the prediction prior retain mean wrong-class allocation mass of 0.296, versus 0.108 for samples already correct under Rank-free. Correcting a label therefore does not certify its update allocation. Full counts, gate distributions, and temporal-window results appear in Appendix C.

The pooled joint diagnostic covers 135,088 samples from four domains. Among the 1,501 $W \to C$ samples corrected by the prior, the mean gate is 0.629, below the 0.825 mean for the 76,041 $C \to C$ samples. The $W \to W$ group has the largest wrong allocation mass (0.437). These groups use labels only after replay for analysis; their means are descriptive and do not isolate compatibility's information beyond baseline confidence. They nevertheless show why a corrected prediction cannot be used as a sufficient condition for assigning full update mass.

The matched-mass control tests whether selecting samples matters beyond reducing update weight. Constant gates admit the same total mass as PACE but assign it uniformly; shuffled gates preserve PACE's gate values while breaking their sample correspondence. Entropy, maximum-probability, and probability-margin gates provide confidence-based alternatives, each calibrated to the same total mass. Table 3 shows that PACE exceeds constant weighting in three domains. It has the highest all-stream gain and lowest WRM among these gates on EuroSAT and Quickdraw, whereas MaxProb is slightly better on VisDA-2017. SUN397 remains a counterexample: its PACE gate lowers accuracy and increases WRM. The complete gate, late-half, and fusion-weight records appear in Table 11.

WRM divides incorrect-class responsibility by total admitted responsibility. Uniformly scaling all sample weights cannot improve this fraction when states and allocations are held fixed; a lower WRM therefore reflects selective weighting in that comparison. In an online replay, gates can also change later class statistics and the effective-count-dependent fusion weight. A control with a common fusion schedule still favors compatibility over constant weighting in the first three domains, but not SUN397 (Table 17). Thus, the evidence supports sample selection in those tested streams without asserting that semantic–Gaussian agreement always identifies safe updates.

Table 4: Same-state joint diagnostic across four domains. Transitions compare Rank-free with Prediction Only; wrong mass averages $1 - a_{i,y_i}$. Gates and allocations precede state updates. Labels are used only for analysis.

Transition	N	Compatibility	Gate	Wrong mass
$W \to C$	1,501	0.535	0.629	0.296
$C \to W$	534	0.556	0.664	0.316
$C \to C$	76,041	0.663	0.825	0.108
$W \to W$	57,012	0.568	0.685	0.437
Total	135,088	–	–	–

Table 5: Compatibility-source comparison: top-1 accuracy (%). Means are recomputed from six dataset rows; bold marks the row maximum. OH and DN abbreviate Office-Home and DomainNet.

Dataset	Rank	Score	Top-1	Random
FGVC Aircraft	27.18	**27.20**	27.07	26.98
EuroSAT	**73.22**	71.91	70.86	66.84
OH Clipart	71.00	70.91	**71.04**	70.36
VisDA-2017	**88.25**	88.03	87.79	87.01
DN Infograph	53.24	**53.29**	53.15	52.53
DN Quickdraw	**19.50**	19.18	19.02	18.10
Mean	55.398	55.087	54.822	53.637

4.4 Additional Controls and Evaluation Scope

Table 5 tests whether the ordinal structure matters. It compares normalized class-rank displacement with per-sample standardized score distance, top-1 agreement, and a randomized ranking. Rank compatibility has higher mean accuracy across six datasets by 0.312, 0.577, and 1.762 pp, respectively. Random rank is worse on all six datasets, and top-1 agreement is worse on five, supporting the use of relationships beyond the leading class. The rank–score advantage is less uniform: score distance is higher by 0.02 pp on Aircraft and 0.05 pp on Infograph, and the mean rank advantage shrinks to 0.112 pp after excluding EuroSAT. These are single-run descriptive differences, not significance estimates.

The mixed semantic–visual anchor adds only a small mean gain. A three-domain grid favors a neighborhood around $\alpha/\tau = 2$ and $\delta = 1$, although this is not an independently selected default. With fixed hyperparameters, Full and Pure Rank exceed Rank-free in all 18 pairs of six domains and three tested stream orders; SUN397's Update-only failure persists. Appendix C gives the detailed controls.

Published DOTA results provide descriptive context: PACE's oracle-tuned 21-stream mean is 72.301% versus DOTA's reported 70.498% under different protocols. On the matched 12-class VisDA-2017 implementation, PACE processes 103.72 versus 86.27 images/s, without establishing a general speed advantage. Full comparisons and timing details are in Appendix C. All accuracy claims remain conditional on target-stream oracle tuning.

5 Conclusion

PACE connects current prediction and online Gaussian adaptation through ordinal compatibility between fixed semantic and evolving class rankings. Its classwise prior corrects dynamic scores, while a samplewise gate controls admitted update mass. Under target-stream oracle tuning, Pure Rank gains 1.02 pp and full PACE gains 1.09 pp over Rank-free on average across 21 evaluations. The joint and matched-mass diagnostics show why prediction correction and update selection need separate assessment; SUN397 exposes a gate failure. Future work should test independently selected parameters and identify when rank agreement ceases to indicate reliable updates.

REFERENCES

Motasem Alfarra, Hani Itani, Alejandro Pardo, Shyma Yaser Alhuwaider, Merey Ramazanova, Juan Camilo Perez, Zhipeng Cai, Matthias Müller, and Bernard Ghanem. Evaluation of test-time adaptation under computational time constraints. In *International Conference on Machine Learning (ICML)*, 2024.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8344–8353, 2022.

Hyeongheon Cha, Dong Min Kim, Hye Won Chung, Taesik Gong, and Sung-Ju Lee. Snap: Low-latency test-time adaptation with sparse updates. volume 38, pp. 90170–90212, 2026.

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

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

Yanqi Ge, Qiang Nie, Ye Huang, Yong Liu, Chengjie Wang, Feng Zheng, Wen Li, and Lixin Duan. Beyond prototypes: Semantic anchor regularization for better representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1887–1895, 2024.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. volume 35, pp. 27253–27266, 2022.

Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost Van De Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. volume 36, pp. 6582–6595, 2023.

Zenghao Guan, Yucan Zhou, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. volume 38, pp. 6616–6639, 2026.

Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 746–754, 2023.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. volume 38, pp. 142772–142798, 2026.

Anselm Haselhoff, Kevin Trelenberg, Fabian Küppers, and Jonas Schneider. The gaussian discriminant variational autoencoder (gdvae): A self-explainable model with counterfactual explanations. In *European Conference on Computer Vision*, pp. 305–322. Springer, 2024.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple method to improve robustness and uncertainty under data shift. In *International conference on learning representations*, volume 2, pp. 6, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. volume 34, pp. 2427–2440, 2021.

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

Jae-Hong Lee. Parameter dynamics of online machine learning and test-time adaptation. volume 38, pp. 68273–68302, 2026.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, volume 2024, pp. 46977–47002, 2024.

Zhihui Lin, Tianyu Yang, Maomao Li, Ziyu Wang, Chun Yuan, Wenhao Jiang, and Wei Liu. Swem: Towards real-time video object segmentation with sequential weighted expectation-maximization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1352–1362. IEEE, 2022.

Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28816–28826, 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*, pp. 16888–16905. PMLR, 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.

Sunghyun Park, Seunghan Yang, Jaegul Choo, and Sungrack Yun. Label shift adapter for test-time adaptation under covariate and label shifts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16375–16385. IEEE, 2023.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. VisDA: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415. IEEE, 2019.

Ori Press, Ravid Shwartz-Ziv, Yann LeCun, and Matthias Bethge. The entropy enigma: Success and failure of entropy minimization. In *International Conference on Machine Learning (ICML)*, 2024.

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

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

Grzegorz Rypeść, Sebastian Cygert, Tomasz Trzciński, and Bartłomiej Twardowski. Task-recency bias strikes back: Adapting covariances in exemplar-free class incremental learning. volume 37, pp. 63268–63289, 2024.

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. volume 35, pp. 14274–14289, 2022.

Sudarshan Sreeram, Young D. Kwon, and Cecilia Mascolo. Tempora: Characterising the time-contingent utility of online test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2026.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5385–5394, 2017.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201. IEEE, 2022.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free clip-based adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 28841–28859, 2024a.

Ziqiang Wang, Zhixiang Chi, Yanan Wu, Li Gu, Zhi Liu, Konstantinos Plataniotis, and Yang Wang. Distribution alignment for fully test-time adaptation with dynamic online data streams. In *European Conference on Computer Vision*, pp. 332–349. Springer, 2024b.

Jinglin Xu, Yi Li, Chuxiong Sun, Xiao Xu, Jiangmeng Li, and Fanjiang Xu. Multi-modal test-time adaptation via adaptive probabilistic Gaussian calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yongcan Yu, Lijun Sheng, Ran He, and Jian Liang. Stamp: Outlier-aware test-time adaptation with stable memory replay. In *European Conference on Computer Vision*, pp. 375–392. Springer, 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.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. volume 35, pp. 38629–38642, 2022.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. volume 37, pp. 67795–67825, 2024a.

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

Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International conference on machine learning*, pp. 41647–41676. PMLR, 2023.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. volume 38, pp. 76280–76310, 2026.

Bowen Zhao, Chen Chen, and Shu-Tao Xia. DELTA: Degradation-free fully test-time adaptation. In *International Conference on Learning Representations (ICLR)*, 2023a.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2023b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16795–16804. IEEE, 2022a.

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

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *European conference on computer vision*, pp. 518–536. Springer, 2022.

USE OF AI ASSISTANCE

We use GPT-6 for paper polishing. All technical content, experiments, results, and conclusions are the authors' own; the language model was used only for editing and proofreading.

A ESTIMATOR DETAILS AND INTERPRETATION

A.1 RESPONSIBILITY-WEIGHTED ONLINE UPDATES

PACE passes ω_i to the weighted Gaussian estimator. To make its effect explicit, consider the undiscounted recursive update for a single feature. Writing $\Delta_{ik} = z_i - \mu_{ik}$, the count, mean, and recursive covariance updates take the form used in distributional adaptation (Han et al., 2026), analogous to exponential-moving-average recursions employed by online adaptation and EM-style update schemes (Niu et al., 2022; Lin et al., 2022; Dai & Yang, 2025):

$$\begin{aligned} n_{i+1,k} &= n_{ik} + \omega_{ik}, \\ \mu_{i+1,k} &= \mu_{ik} + \frac{\omega_{ik}}{n_{ik} + \omega_{ik}} \Delta_{ik}, \\ \Sigma_{i+1,k} &= \frac{n_{ik}\Sigma_{ik} + \omega_{ik}\Delta_{ik}\Delta_{ik}^{\top}}{n_{ik} + \omega_{ik}}. \end{aligned} \tag{12}$$

Here the covariance recursion uses residuals around the pre-update mean; covariance regularization is applied when constructing the scoring covariance in Eq. equation 2. Initialization, covariance sharing, and any statistical decay are settings of the underlying Gaussian estimator. Their configuration must be specified for each component comparison. Eq. equation 12 displays the undiscounted case to isolate the role of the responsibilities.

For $n_{ik} > 0$, the effective step toward z_i is $g_i a_{ik}/(n_{ik} + g_i a_{ik})$. A smaller gate therefore reduces the movement of an established class mean as well as the increase in its count. The covariance receives the same reduced evidence weight. When a class has zero prior count, its first mean estimate is determined by the incoming feature; the gate still controls the effective mass attached to that estimate.

With multiple update views, the estimator accumulates their weighted contributions: ω_{ik}, $\omega_{ik} z_i$, and $\omega_{ik}\Delta_{ik}\Delta_{ik}^{\top}$ in the corresponding numerators are replaced by sums over the selected views. Each view's compatibility and responsibility are evaluated against the same pre-update state, and the resulting statistics are committed after prediction. The view selection and aggregation protocol belong to the base estimator; compatibility gates the responsibility supplied by that protocol.

Algorithm 1: PACE on an unlabeled test stream.

1. Initialize the Gaussian state $\mathcal{S}_1$ and fix the image encoder, text prototypes, and adaptation settings.
2. For each incoming sample x_i, extract its prediction feature and the selected update views.
3. Using $\mathcal{S}_i$, compute semantic and Gaussian scores, their normalized ranks, and the shared compatibility vector for each processed feature (Eqs. equation 1–equation 4).
4. Compute the geometry-adjusted scores and posterior (Eq. equation 5).
5. Apply the compatibility prior, fuse with the zero-shot logits, and emit $\widehat{y}_i$ (Eqs. equation 6–equation 8).
6. Reuse compatibility to compute the gate, normalized allocation, and unnormalized update responsibilities (Eqs. equation 9–equation 11).
7. Accumulate the selected views' weighted statistics, update the Gaussian state, and proceed to x_{i+1}.

The procedure uses no test labels, backward passes, or encoder parameter updates. In addition to the base Gaussian computation, constructing compatibility requires two class rankings, costing $O(K \log K)$ per feature, and $O(K)$ operations for the prior and gate. The optional anchor-based

geometry requires $O(KD)$ operations. Compatibility introduces no persistent state that grows with the stream length.

A.2 Interpretation and Component Isolation

Ordinal evidence. Any strictly increasing transformation applied to either score vector preserves its ranking and hence C_i. Compatibility can therefore compare the two sources without calibrating their raw score scales. This invariance applies to the compatibility vector itself; the complete predictor and gate also depend on score magnitudes through the fused logits and p_i^{G}.

Distinct interventions from shared evidence. At a fixed state, α changes the classwise prediction correction and leaves the update responsibilities unchanged. With the allocation rule and other settings fixed, changing α alone therefore preserves the state trajectory. By contrast, δ changes the admitted update mass, which can affect later class statistics and the count-dependent fusion weight. With geometry disabled ($\gamma = 0$), setting $(\alpha, \delta) = (0, 0)$ gives the rank-free configuration; enabling either parameter isolates one use of compatibility. Setting $\gamma = 0$ with both uses enabled gives the pure ordinal configuration. Eq. equation 6 also shows that prediction depends directly on the ratio α/τ, whereas τ additionally shapes the compatibility aggregated by the update gate.

When gating improves update quality. A smaller update mass alone does not establish better evidence selection. For a diagnostic collection with one normalized allocation per sample, let $\varepsilon_i = 1 - a_{i,y_i}$ be the allocation assigned to incorrect classes, using labels only for analysis. Holding the states and allocations fixed, the fraction of gated responsibility assigned to incorrect classes is

$$\mathrm{WRM}_g = \frac{\sum_i g_i \varepsilon_i}{\sum_i g_i}, \qquad \mathrm{WRM}_g - \mathrm{WRM}_1 = \frac{\mathrm{Cov}(g, \varepsilon)}{\mathbb{E}[g]}, \tag{13}$$

where the expectation and covariance are empirical over the diagnostic collection. This identity follows by expanding $\mathbb{E}[g\varepsilon]$. It specifies the relevant condition: gating lowers the fraction of incorrect update mass when larger gates are associated with smaller allocation errors. Compatibility measures agreement, so this condition is an empirical property to examine rather than a guarantee of the construction. Moreover, Eq. equation 13 holds for fixed states and allocations; end-to-end online effects also depend on how gating changes the subsequent state trajectory.

B Full Experimental Protocol

Datasets and streams. The core evaluation comprises Caltech101, DTD, EuroSAT, FGVC Aircraft, Food101, Oxford Flowers, Oxford Pets, Stanford Cars, SUN397, UCF101, four Office-Home domains (Venkateswara et al., 2017), VisDA-2017 (Peng et al., 2017), and six DomainNet domains (Peng et al., 2019). Aircraft uses the official variant test list; the other nine classification datasets use the test portions of the Zhou split files (Zhou et al., 2022b;a). Office-Home uses all images in each domain, VisDA-2017 uses its 55,388-image validation domain, and DomainNet uses the official domain-specific test lists with a shared 345-class mapping. Office-Home is therefore evaluated as a full-domain stream. Dataset sizes and class counts are provided in Appendix D. Six domains support detailed diagnostics: EuroSAT, Food101, VisDA-2017, DomainNet Quickdraw, SUN397, and DomainNet Real.

Model and preprocessing. We use the official OpenAI CLIP ViT-B/16 checkpoint (Radford et al., 2021) at a resolution of 224×224. Images undergo bicubic resizing, center cropping, RGB conversion, and CLIP normalization. Text prototypes use dataset-specific prompts. Caltech101, Stanford Cars, and SUN397 use seven templates, with each encoded template normalized before averaging and the resulting prototype normalized again. The remaining core datasets use a single dataset-specific template (Appendix D). All 21 core evaluations use one image feature for prediction and updating. The separately reported ImageNet-A/R/V2 experiments (Hendrycks et al., 2021b;a; Recht et al., 2019) use one base view and 63 AugMix views (Hendrycks et al., 2020), select the six lowest-entropy views, and average their features and CLIP logits for prediction; the selected views also supply the updates. This view-selection protocol follows the augmentation- and confidence-based recipe popularized by test-time prompt tuning (Shu et al., 2022; Feng et al., 2023).

Online execution and parameter selection. Each candidate starts from a fresh Gaussian state and replays the complete stream with batch size one. Core experiments use global seed 1 and the canonical dataset order, with no additional sample-order shuffle. Feature caches and stream order are shared within component comparisons. Current predictions precede state updates, and labels never enter the per-sample prediction, compatibility, gate, or Gaussian update computations.

We perform **dataset-specific oracle tuning on the full target test stream for mechanism analysis**. After a candidate completes its replay, labels are used to compute accuracy and select the best configuration for that dataset. There is no separate validation set, and the selected configuration is not subjected to an additional independent replay. Thus, the reported accuracies characterize the method under target-stream parameter selection; they do not estimate held-out performance with independently chosen hyperparameters. This scope is consistent with degradation-free studies that deliberately avoid test-time hyperparameter adaptation (Zhao et al., 2023a), and with the sensitivity of test-time adaptation to unsuitable hyperparameter choices reported in the literature (Niu et al., 2022; Zhao et al., 2023a;b). Historical ImageNet results additionally involve fixed-order tuning and are reported separately from the canonical-order core ablations.

Implementation. Experiments use a single NVIDIA A100 80GB PCIe GPU, an Intel Xeon Gold 6326 CPU, and 64 GiB system memory, with Python 3.10.16, PyTorch 2.4.1+cu121, and torchvision 0.19.1+cu121. CLIP feature extraction uses CUDA FP16 weights; cached features, logits, text prototypes, Gaussian states, and candidate replay use FP32. The update-view setting is `selected`, which reduces to the sole feature for the core datasets.

Parameter records and diagnostic configurations. Appendix E records dataset-specific optimal-performance configurations for all 21 evaluations, covering allocation, compatibility (τ, α, and their ratio), gate power δ, geometry (β, γ), fusion (η, ρ), and Gaussian-state settings, selected from disclosed search ranges over the same fields. These dataset-specific configurations differ from those used in the component-isolation and sensitivity studies (Tables 9–12); they are not a shared parameter setting for all ablations.

Configurations and metrics. Rank-free disables the prediction prior, update gate, and geometry term. Prediction enables only the prior; Update enables only the gate; Pure Rank enables both ordinal operations with geometry disabled. Geometry Only enables the geometric residual with both ordinal operations disabled. PACE enables all three components. This distinction allows the contribution of geometry to be measured separately from the two uses of compatibility.

We report top-1 accuracy and unweighted means across datasets/domains. Accuracy differences are in percentage points (pp). Paired correction statistics use integer sample counts. Compatibility reliability is assessed through conditional means, AUROC, and Spearman correlation. Adaptation diagnostics use responsibility-weighted error fractions and gate distributions. Main accuracies use two-decimal display precision; Tables 10–2 display three decimals, with differences computed before display rounding.

C Extended Experimental Analyses

C.1 Context from Published Test-Time Adaptation Results

Table 6 adds the cross-domain results reported in Table 1 of Han et al. (2026). The ten dataset identities overlap with our classification benchmark, and all rows use a CLIP ViT-B/16 visual backbone. The published rows cover zero-shot CLIP, TPT (Shu et al., 2022), DiffTPT (Feng et al., 2023), TDA (Karmanov et al., 2024), BoostAdapter (Zhang et al., 2024a), HisTPT, ZERO (Wang et al., 2024a), DOTA, and two results using prompt enhancement (PE). We transcribe their per-dataset accuracies and published averages. The PACE row is computed from Table 1; it is selected by the full-test-stream oracle procedure described above. Prompt construction, stream order, and parameter selection have not been matched across studies, so the table provides numerical context rather than a controlled comparison.

Table 7 retains the broader DOTA reference values (Han et al., 2026), including Office-Home, VisDA-2017, DomainNet, and the separately tuned ImageNet variants. The VisDA-2017 value of

Table 6: Published cross-domain top-1 accuracy (%) from Table 1 of Han et al. (2026), alongside the PACE values from Table 1. Air, Cal, Car, Eur, Flw, Food, Pet, and UCF abbreviate Aircraft, Caltech101, Cars, EuroSAT, Flowers102, Food101, Pets, and UCF101. PE denotes prompt enhancement. The literature rows retain their published averages; the PACE mean is recomputed over the same ten dataset identities under full-target-stream oracle tuning. Protocols differ, so row differences are descriptive.

Method	Air	Cal	Car	DTD	Eur	Flw	Food	Pet	SUN	UCF	Mean
Zero-shot CLIP (reported)	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
TPT (reported)	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT (reported)	25.60	92.49	67.01	47.00	43.13	70.10	87.23	88.22	65.74	62.67	65.47
TDA (reported)	23.91	94.24	67.28	47.40	58.00	71.42	86.14	88.63	67.62	70.66	67.53
BoostAdapter (reported)	27.45	94.77	69.30	45.69	61.22	71.66	87.17	89.51	68.09	71.93	68.68
HisTPT (reported)	26.90	94.50	69.20	48.90	49.70	71.20	89.30	89.10	67.20	70.10	67.60
ZERO (reported)	25.21	93.66	68.04	46.12	34.33	67.68	86.53	87.75	65.03	67.77	64.21
DOTA (reported)	26.25	94.16	69.56	47.64	62.78	75.23	87.08	92.01	69.80	72.54	69.71
DMN + PE (reported)	30.03	95.38	67.96	55.85	59.43	74.49	85.08	92.04	70.18	72.51	70.30
DOTA + PE (reported)	29.82	94.85	69.06	55.97	58.35	77.06	87.07	92.40	70.97	74.86	71.04
PACE (oracle tuned)	27.18	94.32	69.71	51.06	73.22	76.94	87.06	92.91	69.56	74.20	71.62

Table 7: Reported DOTA reference and current PACE accuracy (%). The first four rows summarize the core benchmark families; the fifth is the unweighted mean of all 21 core evaluations. ImageNet-A/R/V2 results are listed separately because they use the historical fixed-order tuning protocol. Means and differences are recomputed from dataset rows, rather than from rounded family averages.

Evaluation	DOTA (reported)	PACE (oracle tuned)	Difference (pp)
Ten classification datasets	69.705	71.616	+1.911
Office-Home (4 domains)	83.835	85.495	+1.660
VisDA-2017	86.110	88.250	+2.140
DomainNet (6 domains)	60.325	61.990	+1.665
Core mean (21 evaluations)	70.498	72.301	**+1.804**
ImageNet-A (fixed-order tuning)	61.080	61.950	+0.870
ImageNet-R (fixed-order tuning)	81.000	81.620	+0.620
ImageNet-V2 (fixed-order tuning)	64.500	65.120	+0.620

86.11% is also recovered in the matched efficiency run (Section C.13); other entries remain external reference values. The PACE values use the oracle protocol above, so their differences from the reference values do not estimate improvement under a common model-selection procedure.

Numerically, PACE has a higher mean in each core family. On the ten classification datasets, the average difference from the DOTA reference is 1.911 pp; across all 21 core evaluations it is 1.804 pp. PACE is 0.02 pp lower on Food101 and 0.24 pp lower on SUN397. The ImageNet-A/R/V2 differences are +0.87, +0.62, and +0.62 pp, respectively. These reference comparisons complement the component analysis, while matched baseline reruns and independent parameter selection remain necessary for deployment-oriented comparisons.

Table 8 adds the natural-shift results from Table 2 of Han et al. (2026). Its mean covers ImageNet, ImageNet-A, ImageNet-R, and ImageNet-Sketch; our historical PACE experiments cover only ImageNet-A/R/V2 and use a separate multi-view, fixed-order tuning procedure. The DOTA paper reports 61.50% on ImageNet-A and 81.21% on ImageNet-R, whereas Table 7 gives 61.08% and 81.00%. The zero-shot CLIP row in both published tables follows Radford et al. (2021).

For ImageNet-V2 (Recht et al., 2019), Table 5 of Han et al. (2026) reports 64.67% for TDA and 64.50% for DOTA with ViT-B/16; DOTA reaches 65.20% on the last half of that stream. We adopt 64.50% as the DOTA reference in Table 7.

Table 6 of Han et al. (2026) also reports all-stream and last-half accuracies for nine classification datasets. On DTD, TDA changes from 47.40% over the full stream to 46.22% over the last

Table 8: Published natural-distribution-shift top-1 accuracy (%) from Table 2 of Han et al. (2026). Mean is the unweighted average of the four displayed datasets and does not include ImageNet-V2. All values are reported by the source paper; no method in this table was rerun in our environment.

Method	ImageNet	ImageNet-A	ImageNet-R	ImageNet-S	Mean
Zero-shot CLIP	68.34	49.89	77.65	48.24	61.03
TPT	68.98	54.77	77.06	47.94	62.19
DiffTPT	70.30	55.68	75.00	46.80	61.95
TDA	69.51	60.11	80.24	50.54	65.10
ZERO	69.31	59.61	77.22	48.40	63.64
DOTA	70.69	61.50	81.21	51.84	66.31

Table 9: Paired prediction changes and windowed temporal gains for the diagnostic domains. NCR is the paired net correction; the temporal columns show all-stream / late-half gains of Update and Full over Rank-free (pp); '–' marks domains without windowed records. Negative-NCR rows are the last two entries.

Dataset / domain	N	$W \to C$	$C \to W$	NCR (pp)	Update	Full
EuroSAT	8,100	141	50	+1.123	+5.136 / +7.176	+6.444 / +8.855
Food101	30,300	457	162	+0.974	+1.020 / +1.294	+1.020 / +1.175
VisDA-2017	55,388	480	170	+0.560	+0.881 / +1.152	+1.311 / +1.672
DomainNet Quickdraw	51,750	874	310	+1.090	+0.920 / +1.287	+1.469 / +1.762
SUN397	19,850	6	4	+0.010	–	–
DomainNet Real	52,041	453	161	+0.561	+0.880 / +1.184	+1.141 / +1.430
Oxford Flowers	2,463	2	5	−0.122	–	–
Office-Home Product	4,439	7	16	−0.203	–	–

half, while DOTA changes from 47.64% to 50.47%. These are published summary accuracies, not matched runs against the PACE temporal diagnostics below.

C.2 How Compatibility Changes Current Predictions

We align Rank-free and Prediction outputs by sample index and count beneficial wrong-to-correct transitions, $N_{W\to C}$, and harmful correct-to-wrong transitions, $N_{C\to W}$. Their difference defines the net correction rate:

$$\mathrm{NCR} = 100\frac{N_{W\to C} - N_{C\to W}}{N}. \tag{14}$$

The paired records cover 331,415 samples. In every dataset, the four transition counts sum to N, and the recovered Rank-free and Prediction accuracies agree with Table 1 within its two-decimal rounding precision. Across the stream collection, there are 3,012 wrong-to-correct and 1,119 correct-to-wrong transitions, yielding 1,893 additional correct predictions and a pooled NCR of 0.571 pp. The unweighted dataset mean NCR is 0.436 pp, and 19 of 21 datasets have positive NCR. The pooled and macro values differ because the datasets have different sizes; all rates are recomputable from the integer counts.

EuroSAT gains 91 correct predictions and DomainNet Quickdraw gains 564. Oxford Flowers and Office-Home Product are the two negative cases, confirming that compatibility does not make every intervention beneficial. The positive aggregate result reflects the balance of corrections and harmful changes, rather than an assumption that agreement always identifies the true class.

Compatibility-conditioned correction. The compatibility-decile diagnostic in Figure 4 sorts samples by $\overline{C}_i$ and partitions them into ten approximately equal-frequency bins, D1–D10, from lowest to highest compatibility. Bin counts differ by at most one; numerical boundaries may differ across datasets. For bin b with n_b samples, the four transition counts sum to n_b, and

$$R_{W\to C,b} = 100\frac{N_{W\to C,b}}{n_b}, \qquad R_{C\to W,b} = 100\frac{N_{C\to W,b}}{n_b}, \qquad \mathrm{NCR}_b = R_{W\to C,b} - R_{C\to W,b}.$$

Global NCR is computed from pooled transition counts, equivalently the n_b-weighted mean of bin NCRs. Equal-frequency bins give comparable sample sizes for examining where corrections occur;

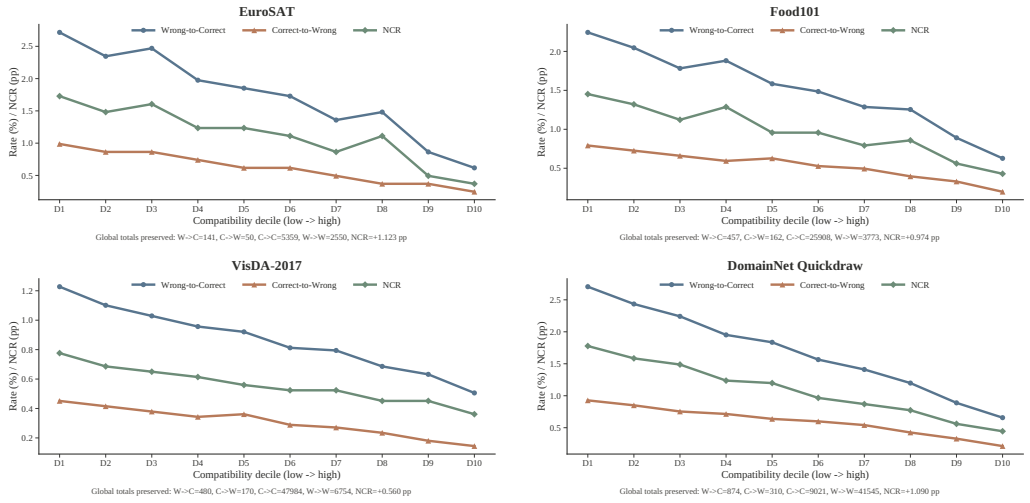


Figure 4: Prediction correction across ten equal-frequency compatibility deciles, from low to high compatibility. Panels show EuroSAT, Food101, VisDA-2017, and DomainNet Quickdraw in reading order. Rates use the count within each decile; global annotations use pooled counts. These quantile bins differ from the fixed-width reliability bins in Figure 5.

Table 10: Compatibility and Rank-free prediction correctness, displayed to three decimals. Gap is calculated from the displayed conditional means; the final row is the unweighted mean over the six domains. AUROC and Spearman are sample-level measures.

Dataset / domain	C_{correct}	C_{wrong}	Gap	AUROC	Spearman
EuroSAT	0.650	0.580	+0.070	0.611	0.187
Food101	0.710	0.667	+0.043	0.573	0.093
VisDA-2017	0.674	0.620	+0.054	0.587	0.107
DomainNet Quickdraw	0.626	0.549	+0.077	0.619	0.161
SUN397	0.660	0.618	+0.042	0.567	0.111
DomainNet Real	0.729	0.684	+0.045	0.576	0.101
Mean	0.675	0.620	+0.055	0.589	0.127

they do not imply equal uncertainty or fixed compatibility intervals. Figure 4 shows the decile plots. The equal-frequency protocol fixes each bin size to one-tenth of the stream size, and per-bin boundaries and transition counts reproduce the pooled totals; correction gains concentrate in the lowest-compatibility deciles (e.g., $+1.73$ pp for the lowest EuroSAT decile versus $+0.37$ pp for the highest).

C.3 Does Samplewise Compatibility Reflect Reliability?

We evaluate $\overline{C}_i$ against the correctness of the Rank-free prediction, with labels used only for this offline diagnostic. AUROC and Spearman correlation are computed from individual samples. The diagnostic in Table 10 uses ten fixed-width intervals of width 0.1 over $[0, 1]$, with 1 included in the final interval. Each interval records sample count, accuracy, and mean compatibility. These bins visualize absolute compatibility and may have very different sample sizes. In all six diagnostic domains, correct predictions have higher mean compatibility than incorrect predictions (Table 10). AUROC is above 0.5 and Spearman correlation is positive throughout, a pattern that parallels the confidence–correctness association exploited in failure prediction (Zhu et al., 2022). The mean AUROC and Spearman correlation over the six domains are 0.589 and 0.127, respectively.

The overall upward trends support compatibility as an ordering signal. For example, EuroSAT accuracy increases from 45.31% in $[0.2, 0.3)$ to 80.92% in $[0.9, 1.0]$; the corresponding Quickdraw accuracies are 9.27% and 34.34%. Very small bins are less stable, such as EuroSAT's four samples in the lowest interval. The modest AUROC and low absolute accuracy of high-compatibility Quick-

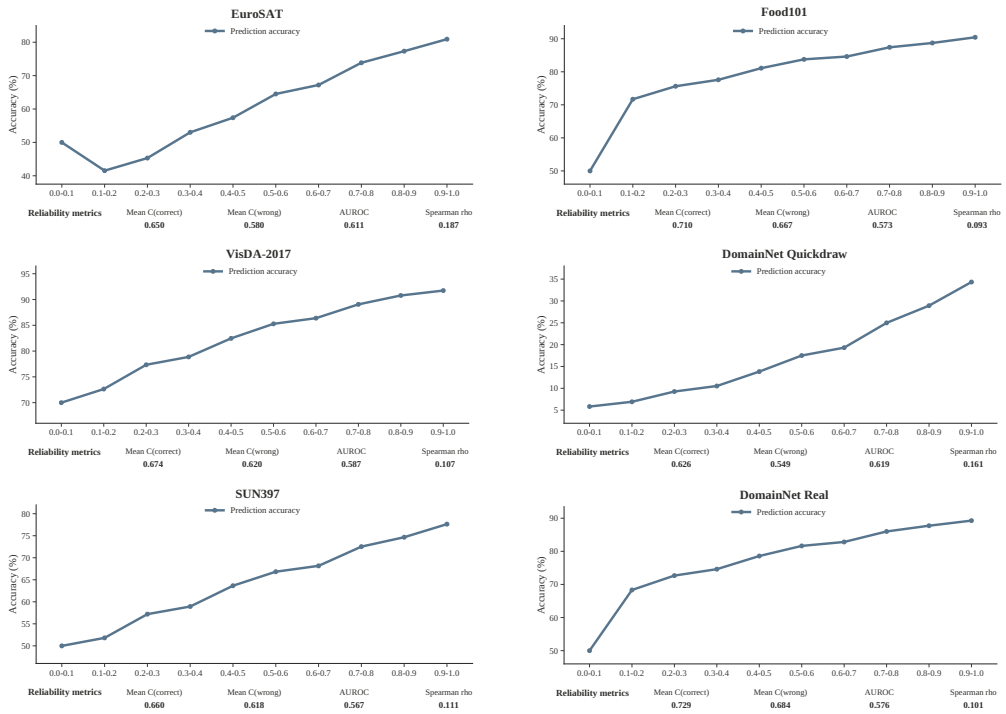


Figure 5: Compatibility and empirical accuracy. Panels show EuroSAT, Food101, VisDA-2017, DomainNet Quickdraw, SUN397, and DomainNet Real in reading order. Lines show accuracy within fixed-width intervals of size 0.1; histograms show counts. Sample-level reliability metrics are displayed to three decimals and are not reconstructed from bin centers.

draw samples also show that compatibility should not be interpreted as a calibrated probability of correctness.

C.4 Responsibility Quality and Gate Distributions

For the actual update responsibilities $\omega_{ik} = g_i a_{ik}$, we measure

$$\mathrm{WRM} = \frac{\sum_i \sum_{k \neq y_i} \omega_{ik}}{\sum_i \sum_k \omega_{ik}}, \qquad \mathrm{CRM} = 1 - \mathrm{WRM}\,. \tag{15}$$

WRM is the fraction of admitted responsibility assigned to incorrect classes. Its normalization distinguishes selective weighting from uniformly reducing all update weights. For the gate-distribution analysis, a sample belongs to the correct-allocation group when $\arg\max_k a_{ik} = y_i$; the other samples form the wrong-allocation group. These groups concern update allocation, not final prediction correctness.

The gate improves responsibility quality in five of six domains. Table 2 shows lower WRM for Update on EuroSAT, Food101, VisDA-2017, DomainNet Quickdraw, and DomainNet Real. Base denotes the ungated update baseline (uniform gate $g_i = 1$, geometry disabled), and Update additionally enables the responsibility gate; both configurations otherwise share their settings. Mean WRM falls from 0.223 to 0.203, a reduction of 1.968 pp computed before display rounding.

Distributional separation explains the gate's selectivity. Figure 6 adds the six gate CDFs to the aggregate responsibility analysis. Over most of the displayed range, the wrong-allocation CDF lies above the correct-allocation CDF in the five favorable domains, indicating that wrong allocations tend to receive smaller gates. SUN397 shows strongly overlapping curves with a slight reversal, consistent with its unfavorable gate-mean ordering. The plot-level conditional means are summarized in Table 2.

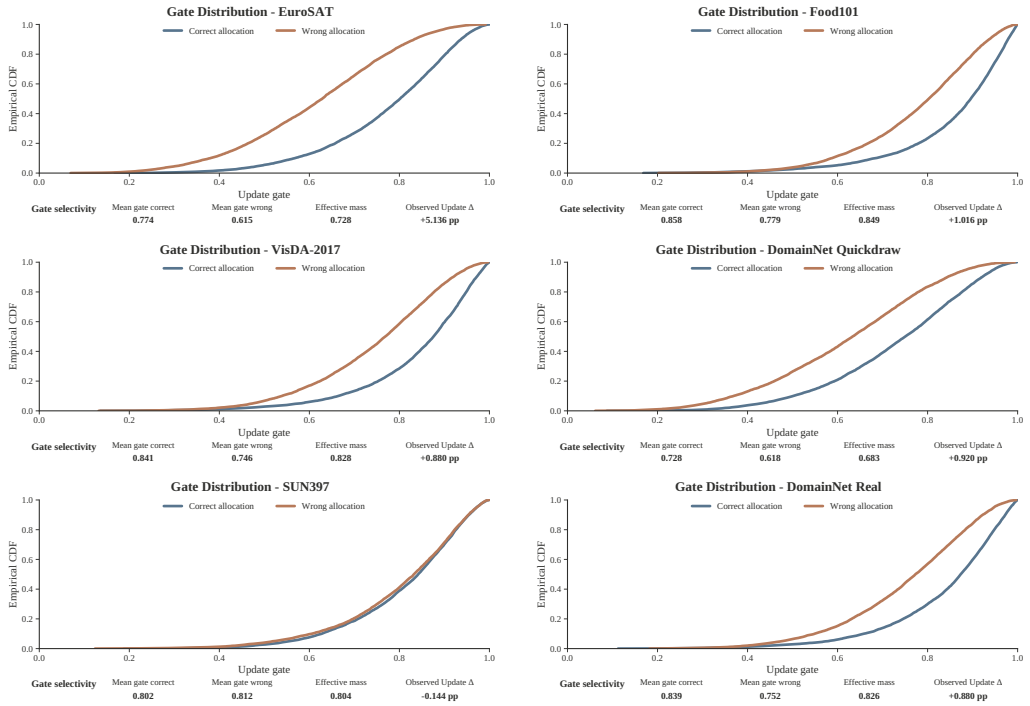


Figure 6: Update-gate CDFs for EuroSAT, Food101, VisDA-2017, DomainNet Quickdraw, SUN397, and DomainNet Real in reading order. Correct-allocation samples satisfy $\arg\max_k a_{ik} = y_i$; all others form the wrong-allocation group. A higher CDF at a fixed gate value means more samples receive gates below that value; the conditional means are reported in Table 2.

Mean Gate(c) and Gate(w) across these six plots are 0.807 and 0.720. The largest separation occurs on EuroSAT (0.774 versus 0.615). SUN397 instead assigns a slightly higher mean gate to wrong allocations (0.812 versus 0.802), and its WRM rises from 0.245 to 0.247. Its Update accuracy also decreases by 0.14 pp. Despite this, compatibility remains positively associated with prediction correctness on SUN397. The distinction demonstrates why update allocations must be evaluated directly: an association with prediction correctness alone does not ensure beneficial weighting of class responsibilities.

Effective update mass. Since $\sum_k a_{ik} = 1$, admitted responsibility sums to g_i for each sample. Table 2 reports

$$M_{\text{total}} = \sum_i \sum_k \omega_{ik}, \qquad M_{\text{eff}} = \frac{M_{\text{total}}}{N} = \frac{1}{N} \sum_i g_i \in [0, 1]. \tag{16}$$

Effective mass is the average update weight admitted per test sample. Normalization removes direct scaling with dataset size; the value can still depend on the stream and adaptation settings.

These measurements concern the allocation of evidence to the state. They do not by themselves establish smaller class-mean or covariance estimation error. Eq. equation 13 characterizes the fixed-state condition for reducing WRM; full online trajectories also include the effects of changed states and allocations.

C.5 Does Gate Selection Matter at Matched Update Mass?

Reducing every update weight can itself change the Gaussian state and the effective count that enters the fusion coefficient in Eq. equation 8. We therefore compare the compatibility gate with a constant gate, a shuffled gate, and gates based on entropy, maximum class probability, and the top-two probability margin, the confidence signals commonly used for sample selection in reliable adaptation (Yu et al., 2024). The alternative gates are computed from the zero-shot posterior: an

entropy gate uses $1 - H(p_i)/\log K$, and the MaxProb and Margin gates use $p_{i,(1)}$ and $p_{i,(1)} - p_{i,(2)}$; each is converted through a power calibration whose exponent is chosen by bisection so that the total admitted mass matches that of the compatibility gate, the constant gate is set to the matched mass on every sample, and the shuffled gate permutes the gate sequence over sample indices with a fixed seed while preserving both the gate values and the total mass. This study isolates the *Update Only* branch: geometry and the prediction prior are disabled, while allocation and other reported settings are held fixed. For each domain, the gated variants are scaled to the same total update mass as the compatibility gate; the ungated Rank-free reference has $g_i = 1$. Table 11 reports the results, including the complete mass, late-half, and mean-fusion-coefficient records.

Table 11: Gate controls with matched total update mass in four diagnostic domains. Each domain column gives gains over its ungated Rank-free reference (pp) or WRM; lower WRM is better. The gated variants share $M_{\text{eff}} = 0.728, 0.828, 0.683, 0.804$ on EuroSAT, VisDA-2017, DomainNet Quickdraw, and SUN397, respectively; No gate has $M_{\text{eff}} = 1$ and Mean λ is the average dynamic-branch fusion coefficient. The PACE gate is the compatibility gate within the Update Only configuration, rather than the full PACE configuration.

Metric	Gate	EuroSAT	VisDA	Quickdraw	SUN397
All ΔAcc	No gate	0.00	0.00	0.00	0.00
	Constant	+1.35	+0.21	+0.19	+0.03
	Shuffled	+0.48	+0.05	+0.04	−0.02
	Entropy	+3.82	+0.70	+0.59	+0.05
	MaxProb	+3.54	+0.91	+0.67	+0.07
	Margin	+4.37	+0.81	+0.83	+0.01
	PACE gate	**+5.13**	+0.88	**+0.92**	−0.14
Late ΔAcc	No gate	0.00	0.00	0.00	0.00
	Constant	+1.70	+0.28	+0.27	+0.01
	Shuffled	+0.62	+0.08	+0.09	−0.04
	Entropy	+4.91	+0.93	+0.78	+0.02
	MaxProb	+4.65	+1.19	+0.90	+0.03
	Margin	+5.66	+1.06	+1.12	−0.02
	PACE gate	+7.18	+1.20	+1.26	−0.26
WRM	No gate	0.286	0.132	0.415	0.245
	Constant	0.278	0.129	0.406	0.244
	Shuffled	0.284	0.132	0.413	0.246
	Entropy	0.259	0.121	0.394	0.243
	MaxProb	0.263	**0.118**	0.389	**0.242**
	Margin	0.251	0.120	0.382	0.245
	PACE gate	**0.241**	0.119	**0.376**	0.247
Mean λ	No gate	0.612	0.744	0.481	0.662
	Constant	0.557	0.711	0.423	0.613
	Shuffled	0.556	0.710	0.421	0.612
	Entropy	0.558	0.711	0.425	0.613
	MaxProb	0.557	0.711	0.424	0.614
	Margin	0.559	0.712	0.426	0.612
	PACE gate	0.556	0.711	0.423	0.613

Constant weighting yields smaller gains than the compatibility gate on EuroSAT, VisDA-2017, and Quickdraw; shuffling the gate-to-sample correspondence reduces gains further. Confidence-based gates narrow the gap: PACE has the largest all-stream gain and lowest WRM on EuroSAT and Quickdraw, while MaxProb has a slightly higher gain and lower WRM on VisDA-2017. SUN397 is a counterexample: the compatibility gate raises WRM from 0.245 to 0.247 and reduces accuracy by 0.14 pp. Equal total mass therefore does not make the gates interchangeable.

Gating also affects the mean fusion coefficient. A second control fixes a common fusion-weight schedule for Constant and PACE gate, giving accuracy gains of 1.18 versus 4.42 pp on EuroSAT, 0.20 versus 0.36 pp on VisDA-2017, 0.16 versus 0.65 pp on Quickdraw, and 0.02 versus -0.09 pp on SUN397 (Table 17); both variants read the same schedule fixed by the compatibility-gate replay, while their own gates continue to drive the state updates. This control supports a selection effect in the first three domains even when the overall fusion schedule is shared, with the negative SUN397 result remaining visible.

C.6 Prediction Correction and Allocation at the Same State

We next ask whether a prediction corrected by the rank prior is automatically a reliable source for updating the Gaussian state. Before each update, the same-state diagnostic records the Rank-free and Prediction outputs alongside compatibility, gate, and class allocation from the same state. Labels are used afterward to form the four paired prediction-transition groups of Eq. equation 14. The same-state diagnostic uses the paired records of Table 9, whose four-domain sample counts match those records and sum to 135,088. Table 4 reports the pooled group summaries; per-domain values appear in Table 18.

The $W \to C$ samples benefit from the prediction prior but retain substantially more wrong allocation mass than the $C \to C$ group (0.296 versus 0.108). Their mean gate is also lower (0.629 versus 0.825). Thus, correcting a sample's current label does not imply that its class responsibility is equally reliable for future state updates. The $W \to W$ group has the largest wrong allocation mass (0.437). These group averages directly connect the two roles of compatibility, while a confidence-stratified, sample-level analysis is still needed to isolate its information beyond baseline confidence.

C.7 Windowed Temporal Diagnostic

The windowed diagnostic divides each diagnostic stream into ten consecutive windows, W1–W10, and tracks gains over Rank-free within each window; all-stream and late-half summaries for Update and Full appear in the temporal columns of Table 9, and the per-window curves in Figures 8–10 (Appendix H). Recent analyses likewise argue that the utility of online adaptation must be established along the temporal axis (Sreeram et al., 2026).

In Table 9, the all-stream entry is the gain over the complete stream and the late-half entry is the unweighted mean of the per-window gains over W6–W10. The per-window curves show larger late-half gains on Update in five domains and retain a negative Update gain on SUN397 (shown in Figure 10); per-window boundaries and correct-prediction counts are recorded with the diagnostic sweep. The component-level failure does not change the aggregate positive effect of the full method.

C.8 Which Compatibility Structure Matters?

The compatibility ablation compares normalized rank displacement with standardized score distance, top-1 agreement, and a randomized semantic ranking on six datasets (Table 5). Score distance standardizes both score vectors per sample before taking the exponential of their absolute difference; top-1 agreement grants full compatibility only when the two sources select the same leading class; the random control permutes the Gaussian class ranking deterministically per sample with a fixed seed. Rank compatibility has the highest mean accuracy, exceeding score distance, top-1 agreement, and random rank by 0.312, 0.577, and 1.762 pp, respectively. It exceeds random rank on all six datasets and top-1 agreement on five. These comparisons support using the relationship across class rankings: disrupting that relationship is consistently detrimental, while retaining only the leading class generally loses useful information.

The advantage over score distance is less uniform. Rank wins on four datasets, but score distance is higher by 0.02 pp on Aircraft and 0.05 pp on Infograph. EuroSAT contributes the largest rank–score difference (+1.31 pp); excluding it, the mean difference is +0.112 pp. Top-1 agreement also exceeds rank by 0.04 pp on OH Clipart. Thus, the evidence favors ordinal compatibility on average without establishing that it dominates other agreement measures in every domain. These single-run differences are descriptive, not significance estimates.

C.9 How Much Does the Anchor Contribute?

We extend the geometry ablation with text-only and visual-only anchors, corresponding to $\beta = 1$ and $\beta = 0$ in Eq. equation 5; the visual-only anchor is fixed at the start of the replay, whereas the mixed anchor tracks the evolving class means. On these six datasets, Full exceeds Pure Rank by 0.068 pp on average, compared with 1.690 pp over Geometry Only, and the mixed anchor exceeds the text-only and visual-only choices by 0.098 and 0.058 pp on average. This agrees with the 21-stream result that the ordinal operations account for most of the improvement. The mixed anchor is the best of the three anchors on EuroSAT, VisDA-2017, and Quickdraw; text-only is slightly better on OH Clipart, and visual-only is slightly better on Aircraft and Infograph. The mixed anchor offers a small average benefit, while the present results do not establish its necessity on every dataset.

C.10 Allocation and Update Mass

The allocation study compares the four choices of ζ in Eq. equation 10 while retaining the rank-gating mechanism. Table 16 reports each rule across the same six diagnostic domains, together with the unweighted means in its final block; the best rule is not selected separately for each domain. CLIP-dominant and equal mixtures have similar mean accuracy (70.533% and 70.525%), followed by CLIP alone (70.452%) and Gaussian allocation (70.328%). Within each domain, the range across the four rules is 0.11–0.46 pp. All four exceed the corresponding Rank-free accuracy in Table 1, although a matched gate-off control for each rule is needed to isolate the gate's incremental contribution.

The best rule varies by domain: mix08clip has the highest accuracy on EuroSAT, VisDA-2017, and DN Real, while mix05 leads on Food101, Quickdraw, and SUN397. In each domain, the best-accuracy rule also has the lowest WRM among these four records. Within each domain, all other parameters are held fixed and only the allocation rule changes; each rule replays from a fresh state without per-rule reselection. The gate formula can be held fixed while realized gate values change, because allocation affects the online state and hence later compatibility. Appendix F reports those values, their unweighted means, and total update mass $M_{\text{total}} = \sum_{i,k} \omega_{ik}$.

C.11 Sensitivity of Prediction and Update Control

The prediction correction depends on α/τ in Eq. equation 6. We denote this ratio by r to distinguish it from the corrected posterior q_{ik}; the supplied plots label the ratio as q. Figure 7 evaluates $r, \delta \in \{0, 0.5, 1, 2, 4\}$ on EuroSAT, VisDA-2017, and Quickdraw. The $(0, 0)$ accuracies equal the Geometry Only entries in Table 1, so the grid is interpreted with geometry retained. Its origin is not the fully Rank-free configuration.

All three grids peak at $(r, \delta) = (2, 1)$, with accuracies of 73.22%, 88.25%, and 19.50%. The neighboring region $r \in \{1, 2\}$ and $\delta \in \{1, 2\}$ stays within 0.49, 0.04, and 0.13 pp of the respective maxima. This shows a neighborhood of favorable settings in these grids, rather than evidence confined to one cell. Increasing either control beyond this region can reduce accuracy. The shared maximizing cell is an observation on the evaluated streams, not an independently selected default for new domains.

At $r = 2$, increasing δ from 0 to 1 improves full-stream accuracy by 3.16, 0.48, and 0.45 pp, while reducing WRM by 0.045, 0.013, and 0.039, respectively. Increasing δ to 2 further lowers WRM and gives the best late-half accuracy in all three sweeps, although full-stream accuracy peaks at $\delta = 1$. At $\delta = 4$, both accuracies decline relative to $\delta = 2$, and WRM rises slightly. Mean gates fall to 0.401, 0.508, and 0.319. The results are consistent with a trade-off between limiting incorrect evidence and retaining sufficient update mass; stronger attenuation is not uniformly beneficial.

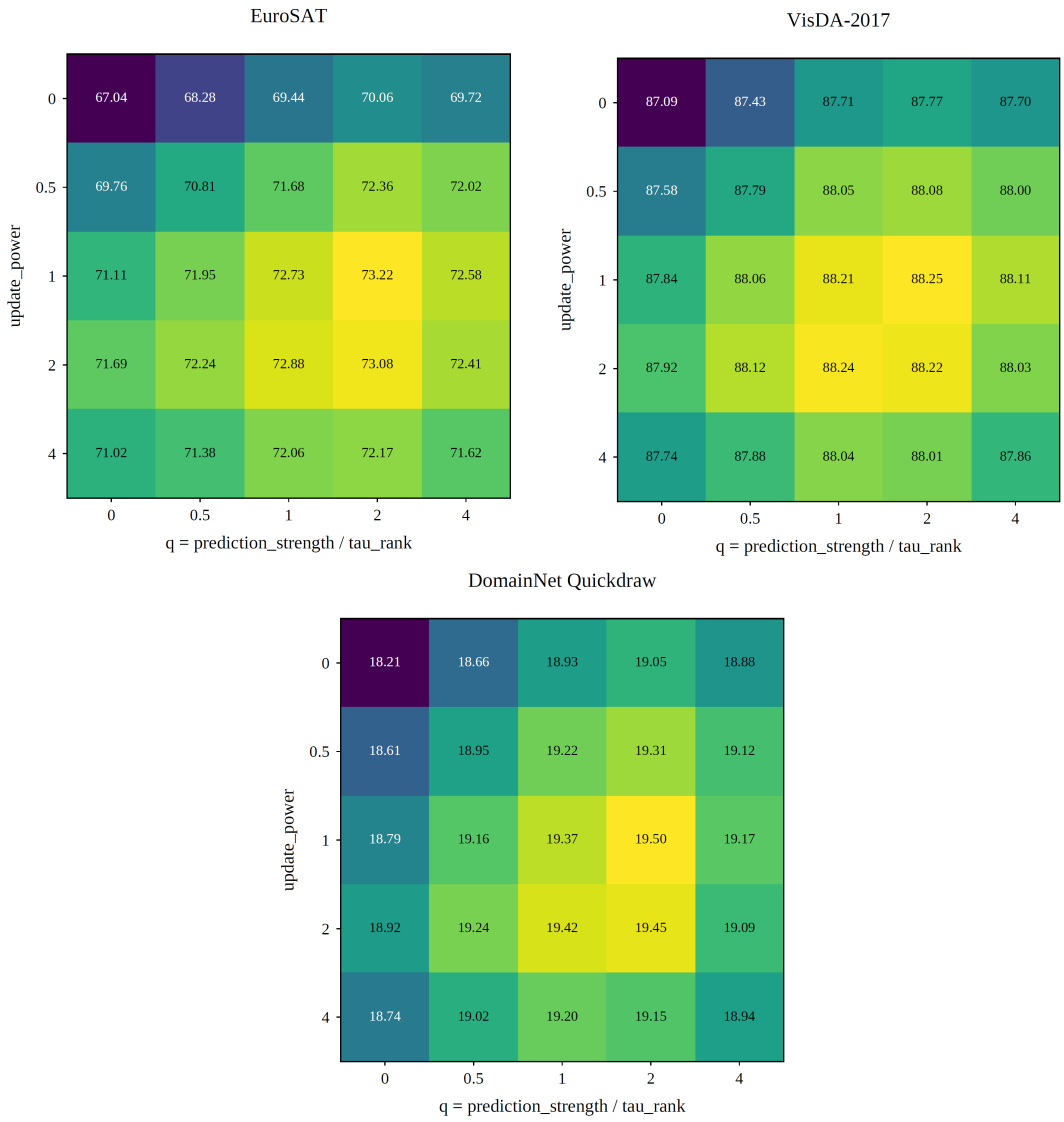


Figure 7: Parameter sensitivity on EuroSAT, VisDA-2017, and DomainNet Quickdraw, in reading order. Each cell gives top-1 accuracy (%). The horizontal coordinate is $r = \alpha/\tau$ (labeled q in the original panels); the vertical coordinate is δ. Color is local to each panel; numerical values support cross-panel interpretation.

The ratio r fully specifies the ordinal prediction penalty for a fixed pre-update state. It does not fully specify adaptation: changing τ also changes C_i and therefore g_i, even if α/τ is fixed. In this sweep, $\tau = 0.20$ is fixed and the grid uses $\alpha \in \{0, 0.10, 0.20, 0.40, 0.80\}$, so both the ratio and the individual settings are reported.

C.12 Variation Across Stream Orders

Table 12 extends the canonical-order analysis to three supplied orders on six domains: the canonical order (seed 1) and two shuffled orders (seeds 2027 and 3407), with all hyperparameters frozen across orders. Full and Pure Rank exceed Rank-free in all 18 dataset–order pairs. Full's macro-average gain over the six domains is 2.003, 1.970, and 1.932 pp for Orders 1, 2, and 3. Its within-domain sample standard deviation ranges from 0.031 to 0.196 pp. These observations show that the gains persist across the three evaluated orders, while retaining the target-stream model-selection qualification.

Table 12: Accuracy across three stream orders, reported as mean ± sample standard deviation (%, $n = 3$). Statistics are recomputed from the displayed order accuracies.

Dataset	Rank-free	Prediction	Update	Pure Rank	Full
EuroSAT	66.74 ± 0.20	67.86 ± 0.18	71.72 ± 0.22	72.98 ± 0.24	73.06 ± 0.20
Food101	86.04 ± 0.06	87.00 ± 0.06	87.05 ± 0.05	87.07 ± 0.04	87.07 ± 0.03
VisDA-2017	86.91 ± 0.12	87.48 ± 0.11	87.78 ± 0.14	88.18 ± 0.12	88.22 ± 0.11
DN Quickdraw	18.00 ± 0.14	19.07 ± 0.17	18.90 ± 0.17	19.38 ± 0.17	19.43 ± 0.17
SUN397	68.91 ± 0.06	68.94 ± 0.06	68.79 ± 0.04	69.04 ± 0.04	69.48 ± 0.08
DN Real	84.81 ± 0.12	85.35 ± 0.12	85.67 ± 0.11	85.93 ± 0.11	85.95 ± 0.11

Table 13: Matched efficiency measurements on VisDA-2017 validation using the same A100, backbone, preprocessing, prompts, and stream order. Accuracy is in percent. These are measured implementation results for this 12-class stream; parameter-selection differences remain subject to the stated oracle-tuning scope.

Method	Time (min)	Latency (ms/image)	Throughput (image/s)	Peak VRAM (GB)	Accuracy
Zero-shot CLIP	5.5	5.96	167.84	0.340	85.64
Original DOTA	10.7	11.59	86.27	0.399	86.11
PACE	8.9	9.64	103.72	0.410	88.25

The SUN397 update-only failure also persists: its gain is -0.14, -0.02, and -0.21 pp across the three orders. Its Full configuration improves by +0.64, +0.56, and +0.50 pp, respectively. Thus, the component failure is not confined to the canonical order, and positive full-method performance does not imply that every component helps in isolation. The three orders quantify observed order variation; they are not confidence intervals over independently selected datasets or configurations. Stream composition and ordering are likewise documented drivers of online adaptation behavior (Wang et al., 2024b).

C.13 Measured Inference Efficiency

We measure end-to-end inference on the 55,388-image, 12-class VisDA-2017 validation stream. Computational cost is a recognized evaluation axis for test-time adaptation (Alfarra et al., 2024), and selective updates are the usual lever for reducing adaptation latency (Cha et al., 2026). Zero-shot CLIP, Original DOTA, and PACE use the same CLIP ViT-B/16 backbone, preprocessing, text templates, and unshuffled sample order, with batch size one and a single view. All three run on the A100 80GB PCIe system specified above. Table 13 gives the supplied timing, throughput, peak device memory, and accuracy values.

Relative to the measured Original DOTA implementation, PACE reduces total runtime by 1.8 minutes (approximately 16.82%), increases throughput from 86.27 to 103.72 images/s, and uses an additional 0.011 GB of peak device memory. Its recorded accuracy is 2.14 pp higher. The speed difference concerns the implementations: PACE maintains Gaussian statistics through effective class masses and first- and second-order accumulators (C, S, Q in the implementation), reducing repeated centered-difference construction and temporary tensors. Both implementations retain a full 512×512 shared covariance and precision-matrix inversion. Compatibility operations act mainly across the 12 classes here; this timing advantage does not establish a general speed benefit from rank gating or predict the cost at larger class counts.

D Dataset and Prompt Details

Table 14 summarizes the core evaluation streams and their reported reference accuracies. Office-Home uses the complete image collection of each domain; VisDA uses validation; the other splits are described in Section 4.1. The VisDA-2017 DOTA value is also recovered in the matched efficiency run in Section C.13. PACE values repeat the Full column of Table 1 to make each reference comparison traceable.

Table 14: Core evaluation streams and reported reference accuracies.

Dataset / domain	Classes	Images	Stream	DOTA (reported)	PACE
Caltech101	101	2,465	Zhou test	94.16	94.32
DTD	47	1,692	Zhou test	47.64	51.06
EuroSAT	10	8,100	Zhou test	62.78	73.22
FGVC Aircraft	100	3,333	Official variant test	26.25	27.18
Food101	101	30,300	Zhou test	87.08	87.06
Oxford Flowers	102	2,463	Zhou test	75.23	76.94
Oxford Pets	37	3,669	Zhou test	92.01	92.91
Stanford Cars	196	8,041	Zhou test	69.56	69.71
SUN397	397	19,850	Zhou test	69.80	69.56
UCF101	101	3,783	Zhou test	72.54	74.20
Office-Home Art	65	2,427	Full domain	83.97	86.44
Office-Home Clipart	65	4,365	Full domain	70.08	71.00
Office-Home Product	65	4,439	Full domain	90.36	92.07
Office-Home Real World	65	4,357	Full domain	90.93	92.47
VisDA-2017	12	55,388	Validation	86.11	88.25
DomainNet Clipart	345	14,604	Official test list	73.64	74.68
DomainNet Infograph	345	15,582	Official test list	51.98	53.24
DomainNet Painting	345	21,850	Official test list	69.24	71.63
DomainNet Quickdraw	345	51,750	Official test list	17.96	19.50
DomainNet Real	345	52,041	Official test list	83.60	85.97
DomainNet Sketch	345	20,916	Official test list	65.53	66.92

For Caltech101, Stanford Cars, and SUN397, the seven templates are: `itap of a {}.`, `a bad photo of the {}.`, `a origami {}.`, `a photo of the large {}.`, `a {} in a video game.`, `art of the {}.`, and `a photo of the small {}.` The strings are preserved from the confirmed implementation. DTD uses `{} texture.`; EuroSAT uses `a centered satellite photo of {}.`; Aircraft, Food101, Oxford Flowers, and Oxford Pets use `a photo of a {}.`, `a type of aircraft.`, `a photo of {}.`, `a type of food.`, `a photo of a {}.`, `a type of flower.`, and `a photo of a {}.`, `a type of pet.`, respectively. UCF101 uses `a photo of a person doing {}.` Office-Home, VisDA, and DomainNet use `a photo of a {}.`

E Optimal-Performance Configurations

Table 15 archives the dataset-specific optimal-performance settings for all 21 evaluations, covering the compatibility, geometry, and fusion fields together with the Gaussian-state fields. These settings differ from those of the component-isolation and sensitivity studies, and should not be substituted for those experimental configurations. All rows use `update_views=selected`; in the single-view core setting this is the sole feature. Configurations are selected from disclosed search ranges over the same fields: the four allocation rules, $\tau \in [0.15, 0.50]$, $r = \alpha/\tau \in [0.5, 2]$, $\delta \in [0.5, 2]$, $\beta \in [0.25, 0.8]$, $\gamma \in [0.03, 0.12]$, $\eta \in [0.2, 2.8]$, $\rho \in [0.01, 0.09]$, and the recorded Gaussian-state grids.

Fields `prediction_strength`, `update_power`, and `residual_strength` correspond to α, δ, γ. The ratio α/τ is denoted by r to distinguish it from the corrected posterior q_{ik}. Rules `mix05clip`, `mix08clip`, and `clip` correspond to $\zeta = 0.5, 0.8, 1$ in Eq. equation 10; `mix05clip` is the equal-mixture rule also labeled `mix05` in the allocation study. Gaussian-state fields keep their recorded names; no conversion of `sigma` to a variance or standard deviation is inferred here.

F Detailed Allocation Results

Table 16 retains each fixed allocation rule separately. M_{total} is the recorded sum of responsibilities; it is not averaged across datasets of different sizes.

Table 15: Optimal-performance compatibility, geometry, fusion, and Gaussian-state settings for all 21 evaluations. These are not common settings of the component-isolation ablations.

Dataset	Allocation	τ	α	$r = \alpha/\tau$	δ	β	γ	η	ρ	epsilon	sigma	init_count	stat_decay
Aircraft	mix08clip	0.2	0.15	0.750	0.75	0.7	0.05	1.2	0.055	0.0001	0.001	0.5	1
Caltech101	mix05clip	0.3	0.3	1.000	1	0.25	0.04	0.4	0.09	0.0015	0.002	0.1	0.99
Cars	clip	0.3	0.225	0.750	0.5	0.35	0.04	1.2	0.05	0.00005	0.004	1	1
DTD	mix08clip	0.5	0.5	1.000	1.5	0.7	0.1	1	0.06	0.00005	0.001	0.1	1
EuroSAT	mix05clip	0.2	0.3	1.500	2	0.8	0.12	2.8	0.05	0.00008	0.002	0.25	1
Flower102	mix08clip	0.25	0.125	0.500	0.75	0.3	0.05	1	0.04	0.00005	0.001	0.5	1
Food101	mix08clip	0.2	0.2	1.000	1	0.35	0.03	0.2	0.015	0.00002	0.002	1	1
Pets	mix08clip	0.25	0.125	0.500	0.5	0.3	0.04	0.6	0.035	0.00008	0.001	0.5	1
SUN397	mix05clip	0.25	0.25	1.000	1.5	0.45	0.06	0.5	0.01	0.0001	0.004	0.25	1
UCF101	mix08clip	0.5	0.375	0.750	1	0.5	0.07	1	0.055	0.00005	0.001	0.1	1
OH Art	mix08clip	0.25	0.25	1.000	1	0.5	0.06	1.2	0.04	0.00008	0.002	0.25	1
OH Clipart	mix05clip	0.2	0.3	1.500	1.5	0.6	0.08	1.4	0.045	0.00008	0.002	0.25	1
OH Product	mix08clip	0.25	0.1875	0.750	1	0.4	0.05	0.8	0.03	0.00005	0.002	0.5	1
OH Real World	mix08clip	0.25	0.25	1.000	1	0.45	0.05	1	0.035	0.00005	0.002	0.5	1
VisDA-2017	mix08clip	0.2	0.4	2.000	1	0.55	0.08	1.2	0.03	0.00005	0.002	0.25	1
DN Clipart	mix08clip	0.25	0.1875	0.750	0.75	0.45	0.05	0.8	0.035	0.00008	0.002	0.5	1
DN Infograph	mix05clip	0.25	0.25	1.000	1	0.55	0.07	1.2	0.04	0.00008	0.002	0.25	1
DN Painting	mix08clip	0.25	0.25	1.000	1	0.5	0.06	1	0.035	0.00005	0.002	0.5	1
DN Quickdraw	mix05clip	0.2	0.4	2.000	1	0.65	0.1	1.6	0.045	0.0001	0.004	0.1	1
DN Real	mix08clip	0.25	0.25	1.000	1	0.45	0.05	0.8	0.03	0.00005	0.002	0.5	1
DN Sketch	mix08clip	0.3	0.225	0.750	1	0.5	0.06	1	0.04	0.00008	0.002	0.25	1

Table 16: Per-domain allocation results. Accuracy and late-half accuracy are percentages. Mean gates are rounded to three decimals, and total mass is recorded to integer precision.

Dataset	Rule	Top-1	Late-half	WRM	Mean g	M_{total}
EuroSAT	clip	72.91	73.56	0.247	0.731	5,921
	mix08clip	73.10	73.79	0.239	0.736	5,962
	mix05	72.98	73.68	0.242	0.739	5,986
	omega	72.64	73.20	0.257	0.727	5,889
Food101	clip	87.00	87.28	0.111	0.829	25,119
	mix08clip	87.04	87.30	0.108	0.832	25,210
	mix05	87.06	87.35	0.107	0.834	25,270
	omega	86.93	87.14	0.115	0.826	25,028
VisDA-2017	clip	88.12	88.52	0.123	0.812	44,975
	mix08clip	88.20	88.66	0.118	0.815	45,141
	mix05	88.17	88.61	0.120	0.817	45,252
	omega	88.01	88.38	0.127	0.808	44,754
DN Quickdraw	clip	19.36	20.00	0.384	0.685	35,449
	mix08clip	19.51	20.20	0.374	0.690	35,708
	mix05	19.54	20.24	0.371	0.693	35,863
	omega	19.22	19.83	0.392	0.681	35,242
SUN397	clip	69.43	69.69	0.249	0.792	15,721
	mix08clip	69.39	69.70	0.247	0.794	15,761
	mix05	69.48	69.77	0.246	0.797	15,820
	omega	69.37	69.62	0.252	0.789	15,662
DN Real	clip	85.89	86.18	0.132	0.810	42,153
	mix08clip	85.96	86.29	0.128	0.813	42,309
	mix05	85.92	86.24	0.130	0.815	42,413
	omega	85.80	86.08	0.137	0.807	41,997
Unweighted means over the six domains:						
	clip	70.452	70.872	0.208	0.777	–
	mix08clip	70.533	70.990	0.202	0.780	–
	mix05	70.525	70.982	0.203	0.783	–
	omega	70.328	70.708	0.213	0.773	–

G Matched-Mass Gate Controls and Joint Diagnostic

The matched-mass records are consolidated into Table 11, including the late-half gains and the mean fusion coefficient; the late-half entry is the unweighted mean over the last five windows.

Table 18 reports the per-domain values behind Table 4. The four group counts in each domain sum to that domain's sample count; the total is 135,088. Group statistics are displayed to three decimals.

Table 17: Fixed common fusion-weight schedule control. Entries are all-stream accuracy gains (pp) over the corresponding ungated reference.

Dataset	Constant gate	PACE gate	PACE − Constant
EuroSAT	+1.18	+4.42	+3.24
VisDA-2017	+0.20	+0.36	+0.16
DN Quickdraw	+0.16	+0.65	+0.49
SUN397	+0.02	-0.09	-0.11

Table 18: Per-domain same-state prediction–update diagnostic. $W \to C$ and $C \to W$ refer to the effect of Prediction relative to Rank-free; wrong mass is the mean allocation to classes other than the true label.

Dataset	Transition	N	Compatibility	Gate	Wrong mass
EuroSAT	$W \to C$	141	0.540	0.620	0.280
	$C \to W$	50	0.543	0.660	0.300
	$C \to C$	5,359	0.651	0.770	0.100
	$W \to W$	2,550	0.582	0.647	0.590
VisDA-2017	$W \to C$	480	0.580	0.720	0.220
	$C \to W$	170	0.529	0.760	0.260
	$C \to C$	47,984	0.671	0.850	0.080
	$W \to W$	6,754	0.623	0.681	0.453
DN Quickdraw	$W \to C$	874	0.510	0.580	0.340
	$C \to W$	310	0.572	0.610	0.350
	$C \to C$	9,021	0.632	0.760	0.180
	$W \to W$	41,545	0.551	0.669	0.425
SUN397	$W \to C$	6	0.580	0.720	0.270
	$C \to W$	4	0.592	0.760	0.310
	$C \to C$	13,677	0.660	0.800	0.160
	$W \to W$	6,163	0.620	0.813	0.437

H TEMPORAL WINDOW ILLUSTRATIONS

These six refined panels display the windowed temporal diagnostic whose all-stream and late-half summaries appear in the temporal columns of Table 9 (Section C.7).

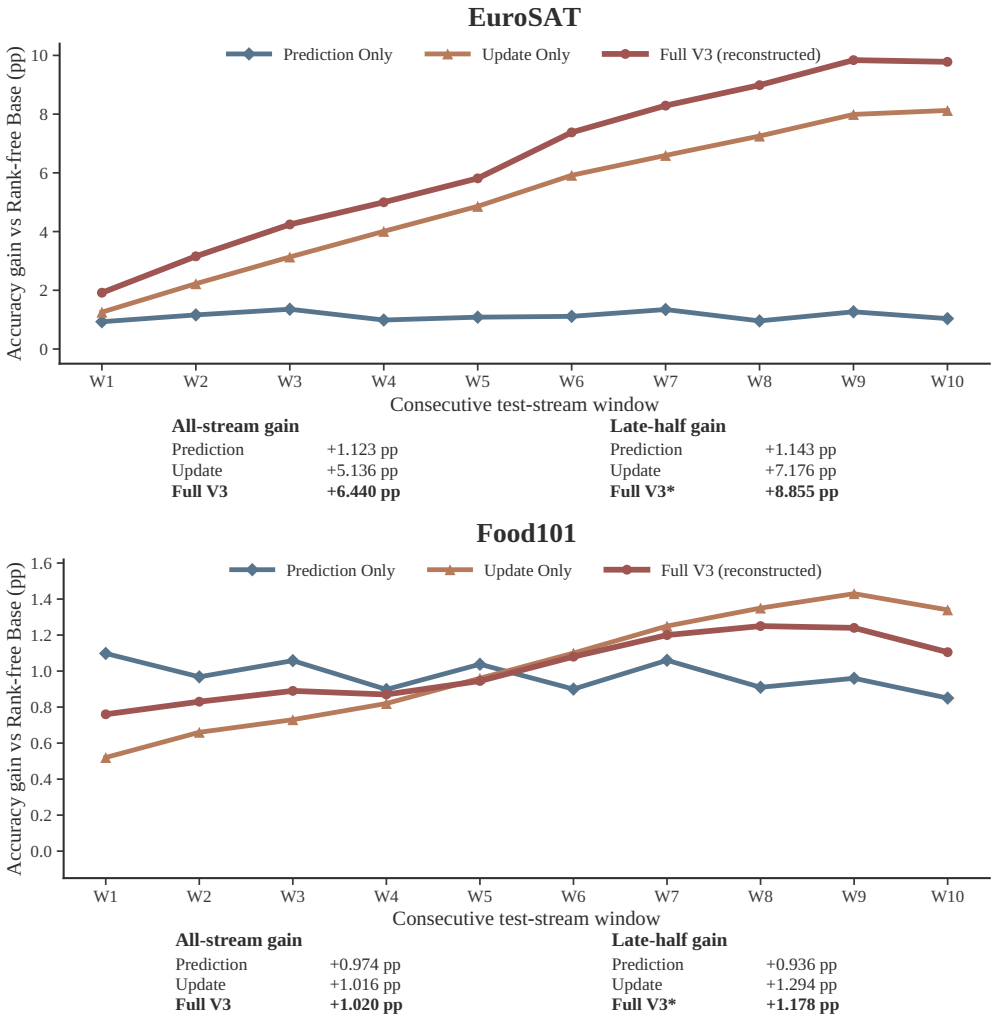
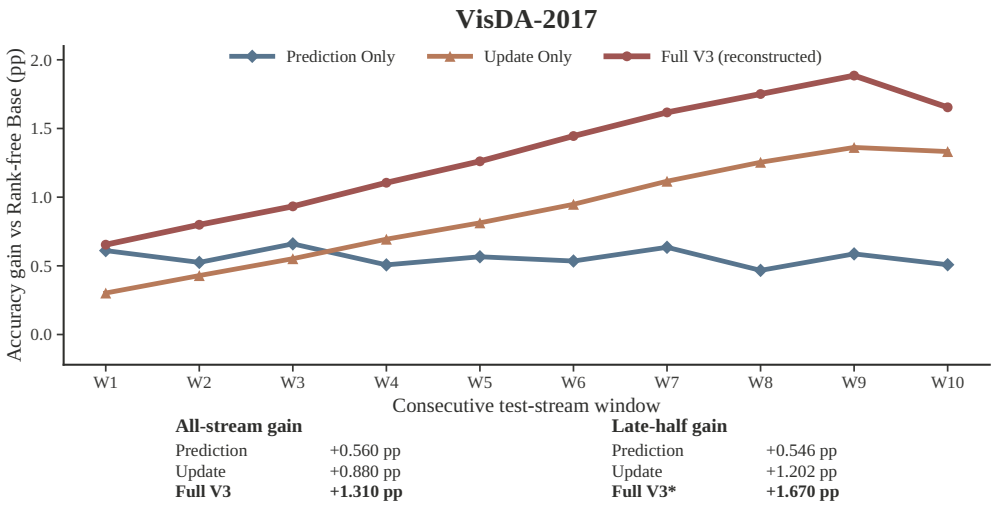


Figure 8: Window-gain panels for EuroSAT and Food101. Each point depicts the gain over Rank-free within one of ten consecutive windows, with the Full configuration labeled in each panel. The all-stream and late-half values are repeated in the temporal columns of Table 9.



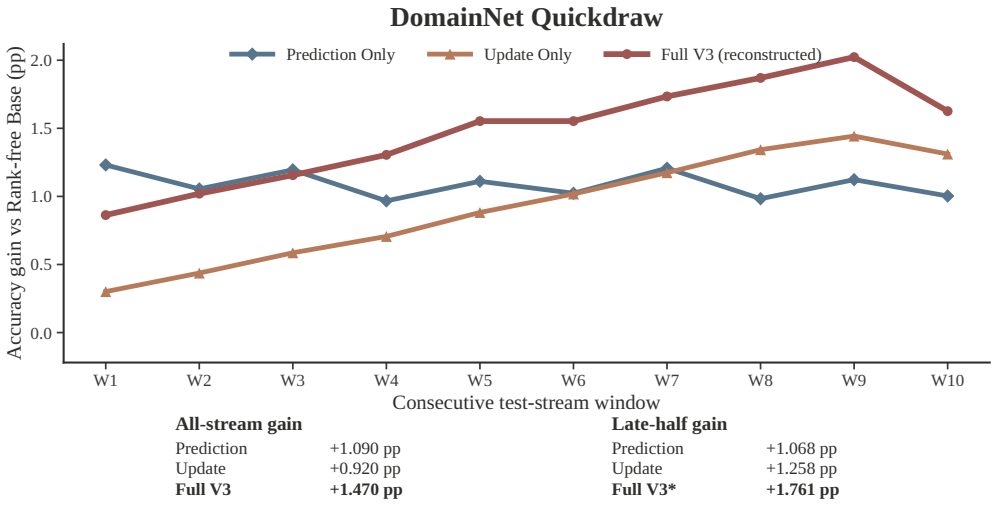
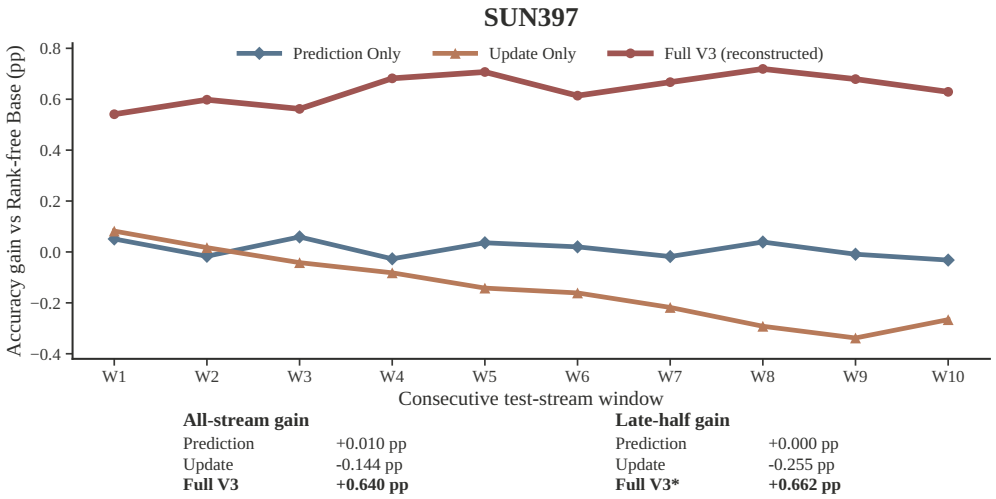


Figure 9: Window-gain panels for VisDA-2017 and DomainNet Quickdraw. The curves use the same protocol and labeling as Figure 8.



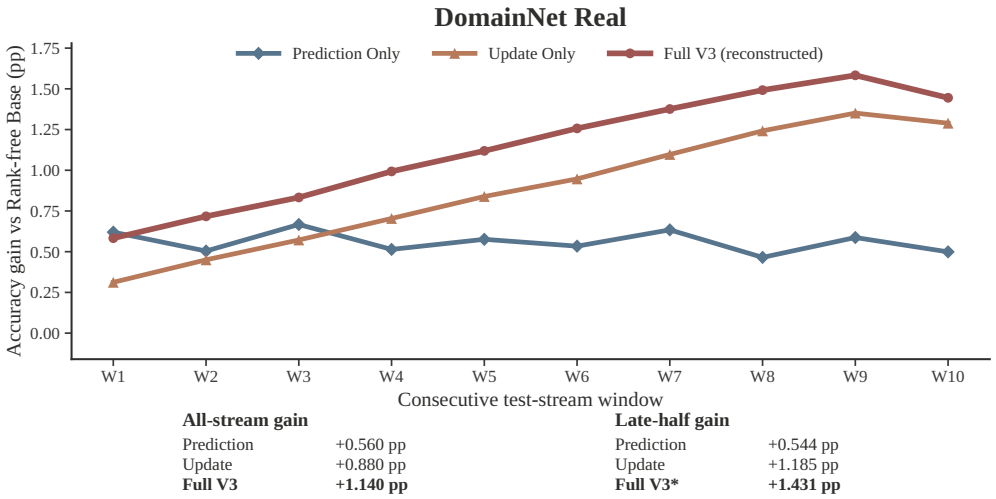


Figure 10: Window-gain panels for SUN397 and DomainNet Real. SUN397's Update curve is negative in later windows, consistent with its negative aggregate update effect.