RELIABILITY-AWARE SELF-DISTILLATION FOR TEST-TIME ADAPTATION OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) improves the robustness of vision-language models under distribution shifts by exploiting unlabeled test data. However, most existing methods derive supervision from the model's current predictions, which may themselves be unreliable under distribution shift and reinforce prediction errors during adaptation. We revisit masked self-distillation as a source of test-time supervision and observe that teacher reliability can be effectively indicated by consistency under mild internal perturbations. Based on this observation, we propose **Reliability-Aware Self-Distillation (RASD)**. Specifically, RASD constructs multiple weak internal probes and aggregates their class-wise evidence into a vulnerability-aware consensus that captures both evidence strength and perturbation sensitivity. The consensus validates full-view predictions for teacher selection and provides a soft correction target, while the selected full-view predictions supervise strongly masked student views through self-distillation. To further stabilize continuous adaptation, we introduce a batch-level marginal-entropy regularization that mitigates prediction collapse. Extensive experiments demonstrate that RASD consistently improves over recent VLM-TTA methods on common corruption benchmarks, while retaining strong generalization across natural ImageNet shifts, diverse recognition datasets, and cross-domain settings.

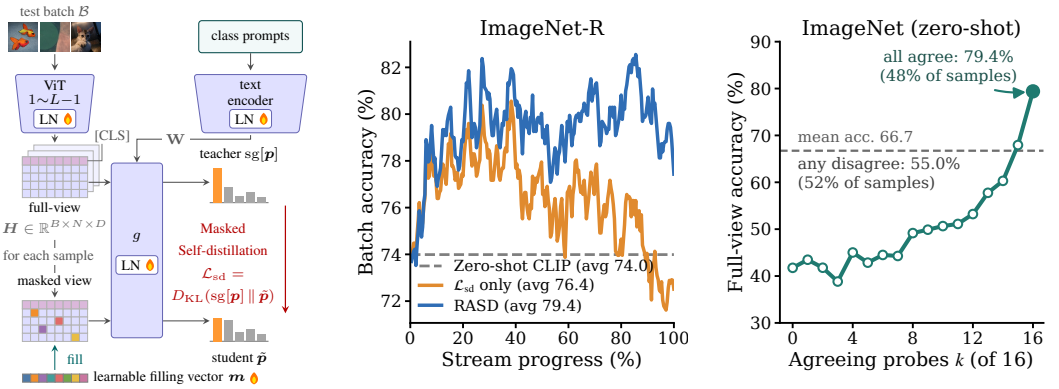
1 INTRODUCTION

Vision-language models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have demonstrated remarkable zero-shot recognition ability by aligning visual and textual representations through large-scale pre-training. However, their performance can degrade substantially when test data deviate from the pre-training distribution, as commonly encountered under image corruptions, domain shifts, and natural distribution changes. This vulnerability poses a critical challenge for deploying such models in open-world environments, where the test distribution is often unknown and continuously evolving. Test-time adaptation (TTA) provides a practical solution by adapting a pre-trained model directly on unlabeled test data, without requiring access to the source training set or additional annotations. The key challenge, however, lies in how to extract reliable supervision from the unlabeled test stream to effectively adapt the model under distribution shift.

Existing TTA methods mainly construct supervision from the model's current predictions. One line of work sharpens the predictive distribution through entropy minimization or confidence-aware sample selection (Wang et al., 2021; Shu et al., 2022). Another line relies on pseudo-labels to optimize prompts, update class prototypes, or align visual and textual representations (Hakim et al., 2025; Maharana et al., 2025). More recent methods further exploit prediction-conditioned statistics or memory-based objectives to improve adaptation robustness (Zhou et al., 2025; Lafon et al., 2025). Despite their different formulations, these strategies largely determine test-time supervision from the current output itself, whose reliability may already be affected by distribution shift. We therefore explore whether different internal views of the same sample can provide complementary evidence to assess the reliability of prediction-based supervision.

To explore this possibility, we consider constructing supervision from different internal views of the same test sample. Specifically, we ask whether the full-view prediction can supervise a feature-masked counterpart of the same sample, forming the masked self-distillation paradigm illustrated in Figure 1(a). This formulation exploits the relation between complete and perturbed internal rep-

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.



(a) Masked self-distillation paradigm at test-time.

(b) Masked self-distillation is effective but may drift.

(c) Internal consistency indicates teacher reliability.

Figure 1: **From self-distillation to a reliable teacher.** Masked self-distillation provides a test-time adaptation signal but may drift, while internal perturbation consistency indicates teacher reliability.

resentations to provide an adaptation signal without explicit pseudo-label construction. Encouragingly, Figure 1(b) shows that naive $\mathcal{L}_{\text{sd}}$ can initially improve the standard full-view accuracy over zero-shot CLIP on ImageNet-R, demonstrating the potential of masked self-distillation for test-time adaptation. However, its batch-wise accuracy exhibits a clear downward drift as adaptation proceeds and eventually falls below zero-shot CLIP, suggesting that blindly distilling every prediction may progressively reinforce unreliable supervision. Interestingly, we find that teacher reliability can be reflected by mild internal perturbations. As shown in Figure 1(c), the accuracy of the original prediction increases steadily with the number of weakly masked probes that agree with it, reaching 79.4% when all probes agree, compared with 55.0% when any disagreement occurs. These observations suggest that the effectiveness of test-time self-distillation is closely tied to teacher reliability, which can be reflected by prediction consistency under internal perturbations.

Beyond binary prediction agreement, each internal probe provides a full class-wise response, revealing how the evidence for individual classes varies under perturbation. Building on this richer signal, we propose Reliability-Aware Self-Distillation (RASD) for test-time adaptation of vision-language models. RASD aggregates the probe-wise class log-odds into a vulnerability-aware consensus distribution q, such that classes supported by strong yet stable evidence receive greater confidence. The consensus is then used to validate the current full-view prediction: only predictions whose decisions are supported by the perturbation-derived consensus are retained as teachers. For these selected samples, $\mathbf{q}_i$ provides a stability-aware soft target for refining the full-view predictive distribution, while the full-view prediction simultaneously supervises a feature-masked student through self-distillation. Finally, to prevent the predictions from gradually collapsing to a few classes during adaptation, we introduce a batch-level entropy regularization whose threshold is determined by the number of classes and the current batch size. Our main contributions are summarized as follows:

- We establish masked self-distillation as a source of test-time supervision for VLMs and identify teacher reliability as its key bottleneck. We further show that prediction consistency under mild internal perturbations provides a useful cue to teacher reliability.
- We develop a vulnerability-aware consensus that aggregates class-wise evidence across internal perturbations to validate prediction-based supervision, enabling reliability-guided teacher selection and stability-aware correction for masked self-distillation.
- Extensive experiments across common corruptions, natural ImageNet shifts, image recognition datasets, and cross-domain benchmarks demonstrate the effectiveness of RASD and its strong performance against recent VLM TTA methods.

2 Related Work

Test-Time Adaptation of Vision-Language Models. Test-time adaptation (TTA) improves the robustness of pre-trained vision-language models under distribution shifts by adapting them on un-

labeled test data during inference. Existing VLM-TTA methods exploit test-time information from different perspectives. TPT (Shu et al., 2022) optimizes textual prompts using augmented test views, while CLIPTTA (Lafon et al., 2025) directly updates CLIP representations with a batch-aware contrastive objective. Other approaches accumulate information from the test stream: TDA (Karmanov et al., 2024) maintains positive and negative feature caches, whereas Ramen (Bao et al., 2026) actively retrieves domain-consistent and prediction-balanced historical samples under mixed-domain shifts. Recent studies further investigate robustness under common corruptions. Mint (Bao et al., 2025) restores corruption-degraded inter-class feature variance, BITTA (Sun et al., 2026) exploits low- and high-entropy samples through complementary learning and unlearning, and LMR (Huang et al., 2026) recovers local semantic margins while mitigating bias accumulation during online adaptation. Different from these directions, RASD focuses on constructing reliable test-time supervision from the model itself. It uses internal perturbation evidence to validate full-view teachers and construct a separate soft correction target, while the selected full-view predictions supervise masked student views.

Self-Distillation for Test-Time Adaptation. Knowledge distillation transfers structured predictive knowledge from a teacher to a student through soft targets (Hinton et al., 2015), while self-distillation constructs such supervision within the model itself. Masked self-distillation has also been widely explored in representation learning, *e.g.,* iBoT (Zhou et al., 2022a). At test time, teacher-based adaptation typically improves supervision through temporal or model-level aggregation: CoTTA (Wang et al., 2022) forms stable targets from weight- and augmentation-averaged predictions, while UnInfo (Adachi et al., 2026) employs an exponential-moving-average teacher for VLM adaptation. More recently, TTD (Chen et al., 2026) introduces a frozen VLM as an external teacher to alleviate the error amplification caused by self-referential supervision. Different from these methods, RASD performs self-distillation entirely within the adapting model, using internal perturbation evidence to select reliable full-view teachers and separately correct the full-view predictive distribution.

Reliable Supervision and Stable Adaptation. Test-time adaptation is sensitive to noisy supervision and error accumulation during online updates. EATA (Niu et al., 2022) improves reliability by selecting low-entropy and non-redundant samples, while SAR (Niu et al., 2023) combines reliable-sample selection with sharpness-aware entropy minimization. Beyond prediction confidence, DeYO (Lee et al., 2024) evaluates sample reliability through prediction changes under object-destructive transformations, and EATA-C (Tan et al., 2025) exploits consistency between full and partial networks. For VLMs, LMR (Huang et al., 2026) alleviates erroneous top-1 reinforcement by protecting plausible near-top candidates and further introduces adaptive margin control and bias correction for stream-level stabilization. Nevertheless, its sample-level adaptation remains anchored to the model's current predictive geometry. In contrast, RASD explicitly evaluates whether the current prediction is supported by multiple internal perturbations before using it as test-time supervision, while separately regularizing prediction diversity during continuous adaptation.

3 Preliminaries and Motivation

CLIP (Radford et al., 2021) performs zero-shot classification by matching an image with the textual descriptions of candidate classes. Given a test image x_i, the visual encoder f_v produces an ℓ_2-normalized image representation $\mathbf{z}_i$. For each class c, its class name is inserted into a prompt template and encoded by the text encoder f_t into a normalized text representation $\mathbf{w}_c$. The classification score is computed from their scaled cosine similarity:

$$s_{i,c} = \mathrm{sim}(\mathbf{z}_i, \mathbf{w}_c)/\tau_{\mathrm{clip}}, \qquad \mathbf{p}_i = \mathrm{softmax}(\mathbf{s}_i), \tag{1}$$

where τ_{clip} denotes the temperature of CLIP, and $\hat{y}_i = \arg\max_c p_{i,c}$ is the zero-shot prediction.

When the test distribution deviates from the pre-training distribution, the pretrained image–text alignment can be disrupted, leading to degraded zero-shot predictions. We therefore consider online test-time adaptation (TTA), where unlabeled test samples arrive sequentially and the model is adapted directly on the test stream without access to source data or target labels.

In the absence of supervision, existing TTA methods typically rely on the model's current predictions, *e.g.,* by sharpening the predictive distribution through entropy minimization (Wang et al., 2021), selecting reliable samples via prediction confidence (Niu et al., 2022; 2023), or treating

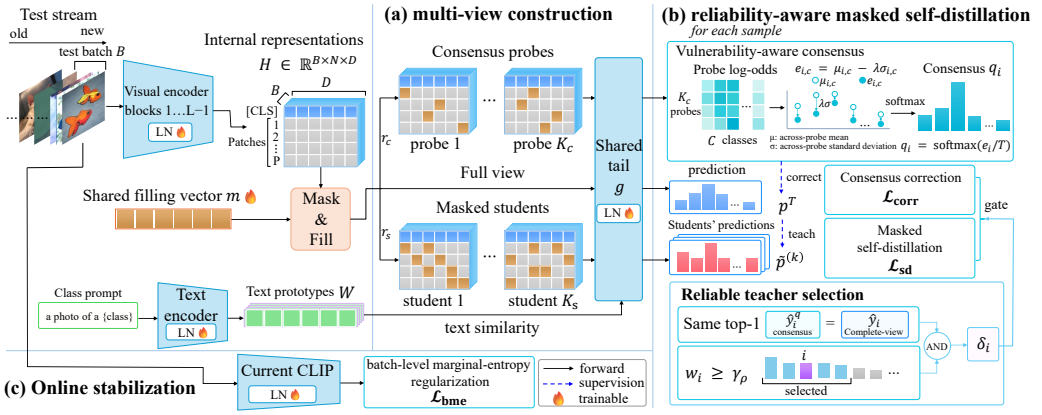


Figure 2: **Overview of RASD.** (a) Weak internal probes form a vulnerability-aware consensus, while stronger masks construct student views. (b) The consensus guides teacher selection and $\mathcal{L}_{\text{corr}}$, whereas selected full-view predictions supervise masked students through $\mathcal{L}_{\text{sd}}$. (c) Batch-level marginal-entropy regularization maintains prediction diversity during online adaptation.

predicted classes as pseudo-labels for subsequent adaptation (Hakim et al., 2025). Despite their different formulations, these strategies still construct supervision from predictions that may already be affected by distribution shift. This motivates us to explore whether different internal views of the same sample can provide complementary evidence for validating prediction-based supervision.

To this end, we consider self-distillation between full and perturbed internal views of the same sample. For a ViT-based CLIP, let $\mathbf{H}_i \in \mathbb{R}^{N\times D}$ denote the token representations entering the last Transformer block, where $N = P + 1$ consists of one `[CLS]` token and P patch tokens, and D is the token dimension. We denote the remaining computation from $\mathbf{H}_i$ to the classification logits by g, such that $\mathbf{s}_i = g(\mathbf{H}_i)$. By perturbing $\mathbf{H}_i$, we can construct different internal views of the same image. Specifically, we obtain a masked view $\widetilde{\mathbf{H}}_i$ by masking a subset of its feature dimensions, with $\widetilde{\mathbf{s}}_i = g(\widetilde{\mathbf{H}}_i)$. A straightforward self-distillation objective is then:

$$\mathcal{L}_{\text{sd}}^{\text{naive}} = T^2 D_{\text{KL}}\left(\text{sg}\left[\text{softmax}(\mathbf{s}_i/T)\right] \,\|\, \text{softmax}(\widetilde{\mathbf{s}}_i/T)\right), \tag{2}$$

where T is the distillation temperature and $\text{sg}[\cdot]$ denotes stop-gradient. This objective encourages the model to preserve the full-view prediction under masked perturbations of its internal representation. As shown in Figure 1(b), masked self-distillation can initially improve the standard full-view prediction over zero-shot CLIP, indicating that consistency between full and perturbed internal views provides a useful test-time adaptation signal.

The effectiveness of this supervision, however, depends critically on the teacher itself. Eq. (2) implicitly treats every full-view prediction as a valid target, although its correctness is unknown at test time. Under distribution shift, an unreliable prediction may propagate erroneous supervision to its masked counterpart and be progressively reinforced during adaptation, consistent with the performance drift observed in Figure 1(b). This raises a central question: *how can we determine whether the current prediction is sufficiently reliable to serve as a teacher?* Interestingly, Figure 1(c) shows that full-view predictions exhibiting higher consistency across mild internal perturbations are substantially more accurate. This observation suggests that perturbation consistency provides a useful cue to teacher reliability.

4 Reliability-Aware Self-Distillation

We propose Reliability-Aware Self-Distillation (RASD) to validate and refine prediction-based test-time supervision using evidence from internal perturbations. As illustrated in Figure 2, RASD aggregates multiple weakly perturbed probes into a vulnerability-aware consensus (Sec. 4.1), which guides teacher selection and soft prediction correction (Sec. 4.2). The selected full-view predictions further supervise strongly masked student views (Sec. 4.3), while a batch-level entropy regularization mitigates prediction collapse during continuous online adaptation (Sec. 4.4).

4.1 VULNERABILITY-AWARE CONSENSUS FROM INTERNAL PROBES

To obtain a reliability signal from internal perturbations, RASD constructs K_c weakly perturbed probe views for each test sample x_i. Starting from the last-block representation $\mathbf{H}_i$, each probe applies a small feature-dimension perturbation independently to every patch token while keeping the `[CLS]` token unchanged. Specifically, for the k-th probe and the n-th patch token, a set of feature dimensions $\Omega_{i,n}^{(k)}$ is sampled uniformly with $\left|\Omega_{i,n}^{(k)}\right| = \mathrm{round}(r_c D)$, where r_c denotes the probe masking ratio. The perturbed representation is defined element-wise as:

$$\widetilde{H}_{i,n,d}^{(k)} = \begin{cases} m_d, & d \in \Omega_{i,n}^{(k)}, \\ H_{i,n,d}, & d \notin \Omega_{i,n}^{(k)}, \end{cases} \tag{3}$$

where $\mathbf{m} \in \mathbb{R}^D$ is a shared learnable filling vector. The same $\mathbf{m}$ is shared across all probe and student views. Each perturbed representation is forwarded through the tail mapping g, yielding $\widetilde{\mathbf{s}}_i^{(k)} = g\left(\widetilde{\mathbf{H}}_i^{(k)}\right)$, where $k = \{1, \ldots, K_c\}$. The probe branches are evaluated without gradient and recompute only the tail of the visual encoder, avoiding the computation of preceding visual blocks.

To characterize the support for a particular class across probes, a suitable evidence score should reflect not only the response of that class itself, but also its separation from all competing classes. We therefore express the evidence for class c under the k-th probe by the multiclass log-odds:

$$\ell_{i,c}^{(k)} = \widetilde{s}_{i,c}^{(k)} - \log \sum_{j \neq c} \exp\left(\widetilde{s}_{i,j}^{(k)}\right). \tag{4}$$

This quantity measures the support for class c against the aggregate evidence of all competing classes. The K_c probes therefore form an empirical evidence distribution for each class. Its mean and standard deviation are given by:

$$\mu_{i,c} = \frac{1}{K_c} \sum_{k=1}^{K_c} \ell_{i,c}^{(k)}, \qquad \sigma_{i,c} = \sqrt{\frac{1}{K_c} \sum_{k=1}^{K_c} \left(\ell_{i,c}^{(k)} - \mu_{i,c}\right)^2}. \tag{5}$$

Here, $\mu_{i,c}$ captures the overall support for class c, while $\sigma_{i,c}$ measures the vulnerability of this support to internal perturbations. RASD thus defines the vulnerability-aware evidence as $e_{i,c} = \mu_{i,c} - \lambda\sigma_{i,c}$, where λ controls the penalty on perturbation sensitivity. Consequently, high evidence is assigned to classes that remain consistently supported across the local perturbation neighborhood.

Collecting the evidence of all C classes as $\mathbf{e}_i = [e_{i,1}, \ldots, e_{i,C}]$, we obtain the vulnerability-aware consensus distribution $\mathbf{q}_i = \mathrm{softmax}(\mathbf{e}_i/T)$, which summarizes the strength and perturbation stability of the class-wise evidence and is detached from gradient computation in subsequent adaptation.

4.2 RELIABLE TEACHER SELECTION AND CONSENSUS CORRECTION

Self-distillation requires a reliable full-view prediction to serve as the teacher. Given the consensus $\mathbf{q}_i$, we first quantify how coherently the internal probes support its decision, and then use this support to determine whether the current full-view prediction should provide supervision.

Let $\hat{y}_i^q = \arg\max_c q_{i,c}$ denote the consensus decision. We measure its agreement with the individual probe predictions as $a_i = \frac{1}{K_c} \sum_{k=1}^{K_c} \mathbb{I}\left[\arg\max_c \widetilde{s}_{i,c}^{(k)} = \hat{y}_i^q\right]$. While a_i captures decision-level consistency across probes, the entropy of $\mathbf{q}_i$ reflects how clearly the aggregated evidence separates the candidate classes. We therefore define: $c_i = 1 - \frac{\mathcal{H}(\mathbf{q}_i)}{\log C}$, where $\mathcal{H}(\cdot)$ denotes Shannon entropy. The two signals are combined into an internal support score $w_i = a_i c_i$. A large w_i indicates that the internal probes provide both consistent and confident support for the consensus decision.

We next connect this internal support to the original full-view prediction. Let $\hat{y}_i = \arg\max_c p_{i,c}$ denote its predicted class. Agreement between the full-view and consensus decisions serves as the primary criterion for teacher selection, while w_i provides an additional selective filter controlled by ρ. Accordingly, the teacher-selection indicator is defined as follows:

$$\delta_i = \mathbb{I}\left[\hat{y}_i = \hat{y}_i^q\right] \mathbb{I}\left[w_i \geq \gamma_\rho\right], \tag{6}$$

where γ_ρ denotes the top-ρ threshold of w_i in the current batch. With $\rho = 1$, selection depends only on decision agreement. Smaller ρ further filters samples with weak internal support.

Beyond teacher selection, the consensus $\mathbf{q}_i$ further provides class-wise guidance for refining the retained predictions. We therefore introduce the consensus correction objective as follows:

$$\mathcal{L}_{\text{corr}} = \frac{1}{\sum_{i\in\mathcal{B}} \delta_i} \sum_{i\in\mathcal{B}} \delta_i T^2 D_{\text{KL}} \left(\text{sg}[\mathbf{q}_i] \parallel \text{softmax}(\mathbf{s}_i/T)\right). \tag{7}$$

Since the selected samples satisfy $\hat{y}_i = \hat{y}_i^q$, the consensus does not introduce a conflicting top-1 target. Instead, $\mathcal{L}_{\text{corr}}$ refines the soft class distribution via the perturbation-aware evidence in $\mathbf{q}_i$.

4.3 Reliability-Aware Masked Self-Distillation

Using the full-view predictions selected via δ_i as teachers, we perform self-distillation on perturbed internal views to preserve these internally supported predictions under feature masking. For each test sample x_i, we construct K_s masked student views from the same last-block representation $\mathbf{H}_i$. Following Sec. 4.1, each student independently masks a fraction r_s of feature dimensions in every patch token and replaces them with the shared filling vector $\mathbf{m}$.

The student masking strength is controlled independently by r_s. In general, stronger masking can be used to construct more challenging student views, while the probe perturbations are kept sufficiently mild to assess prediction stability.

Let $\widehat{\mathbf{H}}_i^{(l)}$ denote the l-th masked student representation, with $\widehat{\mathbf{s}}_i^{(l)} = g\left(\widehat{\mathbf{H}}_i^{(l)}\right)$ and $l = \{1, \ldots, K_s\}$. For each retained sample, the full-view prediction is detached and used as the teacher distribution. The resulting masked self-distillation objective is:

$$\mathcal{L}_{\text{sd}} = \frac{1}{\sum_{i\in\mathcal{B}} \delta_i} \sum_{i\in\mathcal{B}} \delta_i \frac{1}{K_s} \sum_{l=1}^{K_s} T^2 D_{\text{KL}} \left(\text{sg}\left[\text{softmax}(\mathbf{s}_i/T)\right] \parallel \text{softmax}\left(\widehat{\mathbf{s}}_i^{(l)}/T\right)\right). \tag{8}$$

Matching the masked student predictions to these selected teachers encourages the shared adaptable parameters to maintain their predictions under stronger masking perturbations of the internal representation. The shared filling vector $\mathbf{m}$ is also optimized through the student branches.

4.4 Online Stabilization and Overall Adaptation

The preceding objectives improve the reliability of supervision at the sample level. Nevertheless, as the model is continuously updated along the test stream, prediction bias may still accumulate and gradually reduce the diversity of the predicted classes. We therefore introduce a batch-level regularization to maintain sufficient prediction diversity during online adaptation. Since the diversity that can be meaningfully represented by a finite test batch depends on both the number of classes and the batch size, we define a batch-aware diversity reference accordingly.

For each incoming batch $\mathcal{B}_t$, we compute the marginal prediction $\bar{\mathbf{p}}_t = \frac{1}{|\mathcal{B}_t|} \sum_{x_i\in\mathcal{B}_t} \mathbf{p}_i$. We then define the batch-aware entropy reference and the batch-level marginal-entropy regularization as:

$$h_t = \begin{cases} \log C, & C \le |\mathcal{B}_t|, \\ \frac{1}{2}\left(\log C + \log|\mathcal{B}_t|\right), & C > |\mathcal{B}_t|, \end{cases}, \quad \mathcal{L}_{\text{bme}} = \left[h_t - \mathcal{H}(\bar{\mathbf{p}}_t)\right]_+. \tag{9}$$

When the number of classes exceeds the batch size, directly using either $\log C$ or $\log|\mathcal{B}_t|$ can impose an overly strong or conservative diversity constraint. We thus use their midpoint, providing a moderate reference that accounts for both the label-space size and the available batch size.

Based on the above, the overall adaptation objective can be summarized as follows:

$$\mathcal{L} = \mathcal{L}_{\text{corr}} + \mathcal{L}_{\text{sd}} + \mathcal{L}_{\text{bme}}. \tag{10}$$

For each test batch, RASD first outputs the standard full-view prediction with the current model and then adapts on the same batch. The updated model is carried forward to the next batch. Only the LayerNorm affine parameters of the visual and text encoders and the filling vector $\mathbf{m}$ are updated.

Table 1: Per-corruption accuracy (%) on CIFAR-10-C, CIFAR-100-C, and ImageNet-C at severity 5 using CLIP ViT-B/16. Avg denotes the mean over the 15 corruption types. Best and second-best results are highlighted in bold and underline, respectively.

Method	Venue	Gauss.	Shot	Impu.	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brit.	Cont.	Elas.	Pixl.	Jpeg	Avg
CIFAR-10-C																	
CLIP (ZS)	ICML'21	38.01	41.70	54.44	71.72	40.87	67.96	73.59	73.91	77.42	70.24	84.45	62.41	53.85	47.66	59.41	61.18
TENT	ICLR'21	16.22	19.31	44.74	80.19	24.53	76.19	82.05	83.35	83.31	81.01	91.25	81.85	62.71	67.66	57.74	63.47
RoTTA	CVPR'23	39.17	42.90	55.41	72.18	41.32	68.02	74.01	74.38	78.01	70.80	84.80	63.19	54.62	49.32	60.15	61.89
SAR	ICLR'23	47.98	53.60	60.56	74.30	47.56	73.15	76.43	77.91	79.88	75.66	86.79	71.62	58.34	62.03	64.71	67.37
TPT	NeurIPS'22	37.74	42.24	60.57	72.88	44.80	69.69	75.37	75.96	78.84	72.12	85.68	62.04	58.90	55.14	62.64	63.64
WATT-P	NeurIPS'24	44.70	49.66	60.48	74.35	46.33	72.21	76.28	78.18	80.97	74.90	87.66	68.63	58.35	56.13	63.47	66.15
WATT-S	NeurIPS'24	57.22	61.86	67.63	78.61	53.07	77.85	80.14	81.84	83.46	80.38	89.67	78.45	65.76	68.63	67.67	72.81
CLIPArTT	WACV'25	51.06	55.52	61.26	75.96	49.50	73.32	77.78	78.76	80.16	75.64	87.67	71.31	62.67	62.46	63.73	68.54
BCA	CVPR'25	38.61	41.54	56.85	73.35	41.21	66.98	73.55	72.47	77.56	70.33	86.01	63.01	54.01	49.04	62.63	61.81
TCA	ICCV'25	38.06	41.28	55.68	72.25	38.98	66.91	74.09	74.36	76.59	70.54	84.32	63.05	52.14	48.27	61.93	61.23
BATCLIP	ICCV'25	61.13	64.09	65.76	80.51	54.96	80.65	81.94	83.04	84.19	80.84	88.95	82.15	69.16	62.68	66.64	73.85
Mint	NeurIPS'25	60.44	64.36	69.57	78.70	52.70	80.43	80.57	81.72	82.23	79.89	89.94	79.13	65.31	62.82	66.83	72.98
BITTA	ICLR'26	61.92	66.43	64.97	79.63	58.14	80.02	80.61	81.72	83.71	81.10	86.76	81.69	70.89	64.91	68.38	74.06
LMR	ACMMM'26	66.83	67.98	72.22	82.83	64.01	82.81	83.39	85.06	86.06	84.71	**91.92**	86.15	73.23	71.51	69.94	77.91
RASD (Ours)	–	**70.63**	**74.53**	**79.55**	**84.46**	**72.54**	**84.90**	**86.26**	**87.59**	**86.91**	**87.62**	91.72	**89.24**	**77.15**	**82.52**	**74.00**	**81.97**
CIFAR-100-C																	
CLIP (ZS)	ICML'21	19.60	21.41	25.32	42.44	20.07	43.19	47.91	48.43	49.72	41.58	57.05	34.51	29.17	23.93	32.44	35.78
TENT	ICLR'21	8.01	8.82	9.26	51.46	8.35	52.26	55.42	54.45	45.52	50.78	65.48	52.21	36.32	41.92	38.90	38.61
RoTTA	CVPR'23	20.65	22.22	26.17	42.48	20.26	42.90	47.88	48.75	49.92	41.86	57.00	34.52	29.27	25.08	32.88	36.12
SAR	ICLR'23	25.30	27.19	32.78	47.12	23.42	47.16	51.70	51.94	52.48	48.77	61.54	44.50	32.26	33.67	38.06	41.19
TPT	NeurIPS'22	17.95	19.51	27.13	43.53	20.08	42.65	48.63	49.11	49.48	42.14	57.35	33.26	31.13	27.59	32.75	36.15
WATT-P	NeurIPS'24	20.53	22.22	27.30	43.14	17.51	42.37	48.17	47.31	49.34	41.49	57.07	35.29	27.75	25.83	31.89	35.81
WATT-S	NeurIPS'24	21.20	23.11	28.23	44.16	18.45	43.44	49.14	48.16	50.05	42.35	57.82	36.43	28.36	26.85	32.93	36.71
CLIPArTT	WACV'25	24.43	26.65	33.74	47.61	25.49	47.12	52.32	51.87	52.01	47.22	62.04	42.57	33.99	32.98	36.68	41.11
BCA	CVPR'25	23.75	26.05	30.44	47.29	28.24	47.82	50.24	51.66	52.89	44.78	59.99	37.64	31.75	27.88	37.63	39.87
TCA	ICCV'25	22.12	23.47	27.97	46.28	26.87	45.77	49.76	50.01	51.11	42.66	58.29	36.88	30.78	26.52	35.71	38.28
BATCLIP	ICCV'25	24.91	27.73	33.66	50.11	26.27	48.49	54.85	52.35	51.62	48.38	63.27	45.21	34.74	32.38	37.31	42.09
Mint	NeurIPS'25	29.68	33.72	40.64	50.89	27.66	49.38	55.33	52.49	52.20	49.94	65.71	48.42	36.28	33.17	38.27	44.25
BITTA	ICLR'26	26.60	29.11	35.87	50.03	27.07	49.26	54.87	51.92	52.11	48.60	63.00	46.85	35.18	32.56	37.51	42.70
LMR	ACMMM'26	38.91	41.44	48.47	56.79	38.61	**57.15**	**61.36**	**60.70**	**61.07**	**59.00**	**70.36**	**63.17**	45.35	51.58	**45.39**	53.29
RASD (Ours)	–	**40.60**	**42.94**	**49.16**	**57.16**	**40.05**	56.95	61.10	60.18	60.64	58.81	69.54	62.75	**47.34**	**52.98**	45.08	**53.69**
ImageNet-C																	
CLIP (ZS)	ICML'21	13.26	14.18	13.50	24.23	15.74	24.49	22.64	33.09	31.05	37.63	55.62	17.12	13.45	33.05	33.69	25.52
TENT	ICLR'21	5.14	5.70	7.44	25.22	19.34	26.80	24.16	33.56	30.42	37.74	54.24	22.50	13.90	35.02	36.08	25.15
RoTTA	CVPR'23	11.34	12.96	12.32	23.38	15.50	24.66	22.90	32.56	30.02	35.98	54.32	17.20	12.80	31.06	33.46	24.78
SAR	ICLR'23	17.96	20.46	20.68	25.72	23.04	29.52	26.04	34.92	32.74	39.00	55.00	27.14	19.64	36.66	37.50	29.73
TPT	NeurIPS'22	8.48	9.46	10.20	23.98	15.16	25.10	24.00	33.94	32.12	37.08	55.64	16.54	13.68	34.06	33.58	24.87
CLIPArTT	WACV'25	20.79	22.31	21.73	28.33	20.79	32.29	30.48	38.05	34.56	43.56	56.81	30.58	25.01	38.95	39.90	32.57
BCA	CVPR'25	13.12	13.11	12.51	25.21	17.88	27.25	24.72	33.07	31.48	38.17	54.79	22.47	14.10	34.74	36.53	26.61
TCA	ICCV'25	11.56	12.79	11.41	24.02	14.65	25.14	23.04	31.94	30.23	36.11	54.26	18.98	13.33	30.88	33.51	24.79
BATCLIP	ICCV'25	19.32	21.38	19.60	26.58	21.94	30.88	29.02	36.48	32.00	40.98	56.72	26.14	23.74	37.67	38.34	30.72
Mint	NeurIPS'25	20.26	21.84	21.54	27.26	22.26	31.36	27.70	34.04	30.64	40.88	55.02	26.06	20.58	37.88	38.88	30.41
BITTA	ICLR'26	20.76	22.46	20.72	26.50	22.22	30.02	28.98	35.48	31.14	39.68	53.72	27.24	22.26	35.94	36.58	30.25
LMR	ACMMM'26	23.06	25.46	24.88	30.22	29.48	35.98	33.86	41.60	38.10	46.22	58.56	34.10	32.92	43.66	43.06	36.08
RASD (Ours)	–	**28.37**	**30.81**	**30.11**	**32.02**	**32.47**	**38.29**	**37.91**	**44.23**	**42.21**	**49.70**	**60.62**	**38.49**	**40.87**	**47.97**	**46.32**	**40.03**

5 EXPERIMENTS

We evaluate RASD on common corruption benchmarks and further examine its generalization across natural ImageNet shifts, image recognition datasets, and cross-domain benchmarks. Unless otherwise specified, all experiments use CLIP ViT-B/16, with detailed settings provided in Appendix C.

5.1 MAIN RESULTS ON COMMON CORRUPTIONS

Table 1 reports the results on CIFAR-10-C, CIFAR-100-C, and ImageNet-C at severity 5 with the ViT-B/16 as the visual backbone. RASD achieves the highest average accuracy on all three benchmarks, reaching 81.97%, 53.69%, and 40.03%, respectively. Compared with the strong baseline LMR, the corresponding improvements are 4.06%, 0.40%, and 3.95%. The gains are pronounced on CIFAR-10-C and ImageNet-C, while the comparison on CIFAR-100-C is considerably closer. Nevertheless, RASD consistently maintains strong performance across a broad range of corruption types, indicating that its advantage is not restricted to a particular form of visual degradation.

These results are consistent with the motivation of RASD. Under corruption, the visual evidence supporting the current prediction can become unstable, making prediction-based test-time supervision prone to error reinforcement. Rather than directly trusting the current output, RASD exploits

Table 2: Accuracy (%) on ten image recognition datasets using CLIP ViT-B/16. Avg denotes the mean accuracy across all datasets. Best results are in bold and second-best results are underlined.

Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Avg
CLIP (ZS)	24.72	93.35	65.50	43.91	48.36	70.77	85.73	89.10	63.53	67.46	65.24
WATT-S	24.48	93.18	65.66	44.60	54.12	71.46	83.34	88.01	64.32	65.61	65.48
BATCLIP	25.21	93.02	65.12	44.62	48.95	71.38	85.60	88.12	64.66	66.93	65.36
BCA	**28.59**	94.69	66.86	**53.49**	56.63	**73.12**	85.97	90.43	68.41	67.59	68.59
BITTA	25.67	**95.47**	**69.13**	51.68	46.68	72.34	86.44	88.69	**72.82**	71.61	68.05
LMR	26.67	94.24	67.64	45.09	74.75	72.55	**87.07**	90.57	66.59	70.24	69.54
RASD (Ours)	27.30	94.28	67.59	48.58	**78.78**	72.39	86.93	**92.18**	67.68	**72.43**	**70.81**

Table 3: Accuracy (%) across 16 domains from OfficeHome, PACS, VLCS, and Terra Incognita using CLIP ViT-B/16. Best and second-best results are in bold and underlined, respectively.

Dataset	Domain	ZS	WATT-S	TCA	BATCLIP	BITTA	LMR	**Ours**
OfficeHome	Art	83.07	84.10	81.95	82.94	82.82	83.72	**84.88**
	Clipart	68.02	69.67	69.19	68.29	68.16	71.16	**73.61**
	Product	89.14	89.98	90.09	90.02	89.66	90.70	**91.94**
	Real World	89.81	90.68	89.86	89.92	89.92	90.61	**91.19**
PACS	Art	97.46	97.66	96.88	97.95	97.85	**98.10**	97.56
	Cartoon	99.06	99.10	98.98	**99.15**	98.89	99.10	99.10
	Photo	**99.94**	**99.94**	99.88	**99.94**	**99.94**	**99.94**	**99.94**
	Sketch	88.16	91.75	89.16	89.06	89.34	91.47	**88.67**
VLCS	Caltech101	**100.00**	99.76	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	LabelMe	68.63	70.26	67.75	68.38	68.38	68.13	**69.76**
	SUN09	74.82	75.94	**76.55**	76.04	75.13	76.35	75.23
	VOC2007	88.55	86.48	88.06	**89.44**	88.65	89.14	87.86
TerraInc	L38	20.34	30.76	28.88	38.97	44.62	44.11	**56.26**
	L43	33.94	**37.99**	37.31	32.63	31.34	28.17	33.94
	L46	27.98	31.00	29.37	29.05	33.08	18.54	**35.53**
	L100	53.15	37.54	41.05	55.73	**59.21**	51.40	58.68
Avg	–	73.88	74.54	74.06	75.47	76.06	75.04	**77.76**

multiple internal perturbations to assess and refine the supervision used for adaptation, while masked self-distillation encourages the model to preserve reliable predictions under stronger perturbations. This design provides a more robust adaptation signal when the input evidence is degraded.

5.2 Generalization to Diverse Distribution Shifts

Cross-dataset generalization. We further evaluate RASD on ten recognition datasets spanning substantially different visual domains and semantic spaces, including object-centric, texture, scene, remote-sensing, and action recognition tasks. As shown in Table 2, RASD achieves the highest average accuracy of 70.81%, surpassing LMR by 1.27%. Particularly strong gains are observed on EuroSAT, Pets, and UCF101, where RASD reaches 78.78%, 92.18%, and 72.43%, respectively. The strong average performance across these datasets indicates that the improvements of RASD are not limited to a particular recognition task.

Natural ImageNet shifts. Table 4 reports the results on ImageNet-A, ImageNet-V2, ImageNet-R, and ImageNet-Sketch. RASD achieves the highest average accuracy of 61.73%, outperforming LMR by 0.99% and zero-shot CLIP by 4.55%. The gains are particularly pronounced on ImageNet-A and ImageNet-R, where RASD reaches 55.15% and 79.46%, respectively, compared with 52.35% and 77.88% for LMR. On ImageNet-V2 and ImageNet-Sketch, RASD remains competitive with the strongest baseline, resulting in consistently strong performance across the four natural ImageNet shifts.

Table 4: Accuracy (%) on natural ImageNet distribution shifts using CLIP ViT-B/16. IN-A, IN-V2, IN-R, and IN-S denote ImageNet-A, ImageNet-V2, ImageNet-R, and ImageNet-Sketch, respectively.

Method	IN-A	IN-V2	IN-R	IN-S	AVG
CLIP (ZS)	47.77	60.86	73.99	46.10	57.18
WATT-S	48.11	60.80	74.43	46.79	57.53
CLIPArTT	50.03	61.20	76.11	47.89	58.81
BATCLIP	48.07	60.02	74.28	46.91	57.32
Mint	51.16	60.34	75.81	48.24	58.89
BITTA	48.43	55.70	70.00	36.97	52.78
LMR	52.35	62.20	77.88	**50.51**	60.74
RASD (Ours)	**55.15**	**62.35**	**79.46**	49.95	**61.73**

Cross-domain generalization. Table 3 evaluates RASD across 16 domains from OfficeHome, PACS, VLCS, and Terra Incognita. RASD achieves the highest overall average accuracy of 77.76%, outperforming BITTA and LMR by 1.70% and 2.72%, respectively. Strong improvements are observed on OfficeHome and the more challenging Terra Incognita benchmark. In particular, RASD reaches 56.26% on Terra Incognita-L38, compared with 44.62% for BITTA and 44.11% for LMR, while remaining competitive on the relatively saturated PACS and VLCS benchmarks. Overall, the consistent improvements across these benchmarks demonstrate that the benefit of RASD extends beyond common corruptions to substantially different forms of distribution shift.

5.3 Ablation Studies

Effect of each component. Table 5 evaluates the contribution of each adaptation objective on CIFAR-10-C and CIFAR-100-C. With reliability-based teacher selection enabled, masked self-distillation alone already substantially improves over zero-shot CLIP, showing that distilling internally supported full-view predictions to masked student views provides an effective adaptation signal. Consensus correction alone yields more moderate gains, while combining it with self-distillation further improves the accuracy to 74.01% and 43.96%, indicating that the perturbation-derived consensus provides complementary soft guidance beyond teacher selection. Introducing the batch-level marginal-entropy regularization brings a further substantial improvement, with $\mathcal{L}_{\text{sd}} + \mathcal{L}_{\text{bme}}$ reaching 79.15% and 50.48%, and $\mathcal{L}_{\text{corr}} + \mathcal{L}_{\text{bme}}$ reaching 80.62% and 52.25%. The full objective achieves the best performance of 81.97% and 53.69%, showing the complementary effects of reliable self-distillation, consensus-based correction, and online stabilization.

Table 5: Ablation of the adaptation objectives on CIFAR-10-C and CIFAR-100-C at severity 5 using CLIP ViT-B/16.

Setting	CIFAR-10-C	CIFAR-100-C
Zero-shot CLIP	61.18	35.78
$+ \mathcal{L}_{\text{sd}}$	72.85	40.66
$+ \mathcal{L}_{\text{corr}}$	62.01	39.59
$+ \mathcal{L}_{\text{sd}} + \mathcal{L}_{\text{corr}}$	74.01	43.96
$+ \mathcal{L}_{\text{sd}} + \mathcal{L}_{\text{bme}}$	79.15	50.48
$+ \mathcal{L}_{\text{corr}} + \mathcal{L}_{\text{bme}}$	<u>80.62</u>	<u>52.25</u>
Full objective	**81.97**	**53.69**

Sensitivity to batch size and adaptation steps. Figure 3 examines the effect of the batch size B and the number of adaptation steps S on CIFAR-10-C and CIFAR-100-C. Increasing the batch size consistently improves the performance on both benchmarks, with relatively larger gains observed at smaller batch sizes and diminishing improvements as B increases. The performance becomes largely stable from $B = 128$ onward, and further increasing the batch size provides little benefit. A similar trend is observed for the number of adaptation steps: accuracy improves steadily from $S = 1$ to $S = 10$, while increasing to $S = 15$ results in minor changes. Based on these results, we use $B = 128$ and $S = 10$ as the default settings in the experiments.

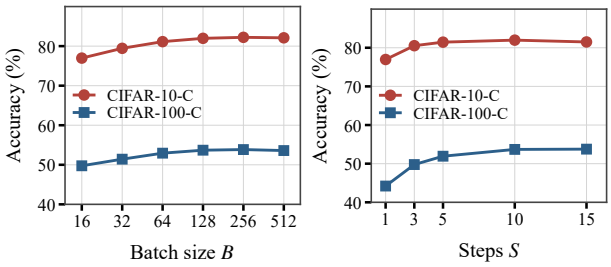


Figure 3: Ablations of batch size and adaptation steps per batch on CIFAR-10-C and CIFAR-100-C.

6 Conclusion

We have presented RASD, a reliability-aware self-distillation framework for test-time adaptation of vision-language models. Starting from the observation that masked self-distillation is effective but sensitive to unreliable teachers, RASD exploits mild internal perturbations to assess teacher reliability and aggregates their class-wise evidence into a vulnerability-aware consensus. The consensus supports reliable teacher selection and soft prediction correction, while the selected full-view predictions supervise strongly masked student views through self-distillation. A batch-aware marginal-entropy regularization further stabilizes continuous adaptation by mitigating excessive prediction concentration. Extensive experiments demonstrate the best average performance across common corruption benchmarks, natural ImageNet shifts, image recognition datasets, and cross-domain benchmarks. Our results show that internal perturbation responses provide an effective source of reliable test-time supervision without requiring external teachers, annotations, or source data.

AI USE STATEMENT

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

ETHICS STATEMENT

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

REPRODUCIBILITY STATEMENT

The method and all hyper-parameters are given in Sec. C.2. The configurations of the compared methods are in Appendix D. The datasets and evaluation protocol are in Appendix C.1. All experiments use a fixed random seed with deterministic computation enabled. We will release the complete code, configuration files, and run scripts, including the scripts that generate all tables and figures upon acceptance.

REFERENCES

Kazuki Adachi, Shin'ya Yamaguchi, and Tomoki Hamagami. Uniformity first: Uniformity-aware test-time adaptation of vision-language models against image corruption. *Transactions on Machine Learning Research (TMLR)*, 2026.

Wenxuan Bao, Ruxi Deng, and Jingrui He. Mint: A simple test-time adaptation of vision-language models against common corruptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Wenxuan Bao, Yanjun Zhao, Xiyuan Yang, and Jingrui He. Ramen: Robust test-time adaptation of vision-language models with active sample selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9632–9642, 2026.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, pp. 472–489, 2018.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, pp. 446–461, 2014.

Xiao Chen, Jiazhen Huang, Zhiming Liu, Qinting Jiang, Fanding Huang, Jingyan Jiang, and Zhi Wang. Test-time distillation for continual model adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7593–7604, June 2026.

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

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

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1657–1664, 2013.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021a.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yan Huang, Guowei Wang, Xu Wang, Kangjun Liu, and Xin Lin. Local margin restoration for test-time adaptation of vision-language models. *arXiv preprint arXiv:2608.02216*, 2026.

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

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

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 554–561, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Marc Lafon, Gustavo Adolfo Vargas Hakim, Clément Rambour, Christian Desrosiers, and Nicolas Thome. CLIPTTA: Robust contrastive vision-language test-time adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 (ICLR)*, 2024.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5542–5550, 2017.

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

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

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729, 2008.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning (ICML)*, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*, 2023.

David Osowiechi, Mehrdad Noori, Gustavo Adolfo Vargas Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. WATT: Weight average test time adaptation of CLIP. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 48015–48044, 2024.

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

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

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

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14274–14289, 2022.

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

Jingwei Sun, Jianing Zhu, Jiangchao Yao, Gang Niu, Masashi Sugiyama, and Bo Han. Bilateral information-aware test-time adaptation for vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Mingkui Tan, Guohao Chen, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Peilin Zhao, and Shuaicheng Niu. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47(8):6274–6289, 2025.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5018–5027, 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.

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

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

Zixin Wang, Dong Gong, Sen Wang, Zi Huang, and Yadan Luo. Is less more? Exploring token condensation as training-free test-time adaptation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 144–154, 2025.

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

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15922–15932, 2023.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022a.

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

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

APPENDIX

CONTENTS

A Algorithm **15**

B Analytical Properties of RASD **15**

B.1 Analysis of Masked Self-Distillation 15

B.2 Additional Analytical Properties 17

C Experimental Setups **18**

C.1 Dataset Details 18

C.2 Implementation Details 18

D Settings of the Compared Methods **19**

E Additional Results **21**

A ALGORITHM

Algorithm 1 summarizes the complete online adaptation procedure of RASD. For each incoming batch, the current model first produces the full-view predictions used for evaluation. The same batch is then used for adaptation: internal probes build the vulnerability-aware consensus for teacher selection and correction, while strongly masked student views are optimized through self-distillation. In parallel, a batch-aware marginal-entropy regularization maintains sufficient prediction diversity to mitigate class collapse during continuous online updates. The updated model is then carried forward to the next batch, and the predictions from all batches are collected in stream order.

Algorithm 1 Online adaptation procedure of RASD

Require: Pretrained CLIP (f_v, f_t), class prompts $\{t_c\}_{c=1}^{C}$, test stream $\{\mathcal{B}_t\}_{t=1}^{N_b}$, adaptation steps S, and hyperparameters $K_c, r_c, K_s, r_s, \lambda, T, \rho$
Ensure: Predictions $\mathcal{P}$ over the complete test stream

```
Initialize P ← ∅ and model parameters θ_1
for t = 1, ..., N_b do
    Predict: obtain the standard full-view predictions {p_i^out}_{x_i ∈ B_t} with the current model θ_t
    Append {p_i^out} to P in stream order
    Set the batch-aware entropy reference h_t by Eq. (9)
    for s = 1, ..., S do
        Re-encode text prototypes and obtain {H_i}_{x_i ∈ B_t}
        Compute the full-view predictions {p_i}_{x_i ∈ B_t}
        Consensus probing: construct K_c gradient-free probe views and obtain q_i and w_i
         following Secs. 4.1–4.2
        Teacher selection and correction: select teachers using δ_i in Eq. (6) and compute
         L_corr by Eq. (7)
        Masked self-distillation: construct K_s masked student views with gradients and
         compute L_sd by Eq. (8)
        Online stabilization: compute the batch marginal prediction p̄_t = |B_t|^{-1} Σ_{x_i ∈ B_t} p_i
         and compute L_bme by Eq. (9) using the entropy reference h_t
        Update the adaptable parameters using the overall objective in Eq. (10)
    end for
    Carry the updated model forward as θ_{t+1}
end for
return P
```

B ANALYTICAL PROPERTIES OF RASD

This section provides analytical insights into several key components of RASD. We first analyze masked self-distillation and characterize its effect on decision stability under internal masking. We then discuss additional properties of the vulnerability-aware evidence, consensus correction, and marginal-entropy stabilization.

B.1 ANALYSIS OF MASKED SELF-DISTILLATION

Consider a retained sample x_i. Let $\mathbf{p}_i^T = \mathrm{softmax}(\mathbf{s}_i/T)$ denote the temperature-scaled prediction of the full-view teacher. For a random student mask $\xi \sim \mathcal{M}_{r_s}$, let $\hat{\mathbf{p}}_{i,\xi}^T = \mathrm{softmax}(\hat{\mathbf{s}}_{i,\xi}/T)$ denote the corresponding masked-view prediction.

Proposition 1 (Decision stability under masked self-distillation). Assume that the full-view teacher has a unique top-1 class $y_i = \arg\max_c p_{i,c}^T$, and define its probability margin as:

$$\Delta_i^T = p_{i,y_i}^T - \max_{c \neq y_i} p_{i,c}^T > 0. \tag{11}$$

Then the probability that a random masked student changes the full-view decision satisfies:

$$\Pr_{\xi\sim\mathcal{M}_{r_s}}\left[\arg\max_c \hat{p}^T_{i,\xi,c} \neq y_i\right] \leq \frac{2\,\mathbb{E}_\xi\left[D_{\mathrm{KL}}\left(\mathbf{p}_i^T \| \hat{\mathbf{p}}_{i,\xi}^T\right)\right]}{(\Delta_i^T)^2}. \tag{12}$$

Proof. For brevity, let $\mathbf{p} = \mathbf{p}_i^T$, $\hat{\mathbf{p}} = \hat{\mathbf{p}}_{i,\xi}^T$, $y = y_i$, and $\Delta = \Delta_i^T$. Consider any mask ξ for which the masked student changes the full-view decision. Then there exists a class $j \neq y$ such that $\hat{p}_j \geq \hat{p}_y$. Since y is the unique top-1 class of the teacher, we obtain:

$$p_y - p_j \geq p_y - \max_{c\neq y} p_c = \Delta. \tag{13}$$

Meanwhile,

$$\begin{aligned} p_y - p_j &= (p_y - \hat{p}_y) + (\hat{p}_y - \hat{p}_j) + (\hat{p}_j - p_j) \\ &\leq (p_y - \hat{p}_y) + (\hat{p}_j - p_j) \\ &\leq |p_y - \hat{p}_y| + |p_j - \hat{p}_j| \\ &\leq \|\mathbf{p} - \hat{\mathbf{p}}\|_1, \end{aligned} \tag{14}$$

where the first inequality follows from $\hat{p}_y - \hat{p}_j \leq 0$. Combining Eqs. (13) and (14) yields:

$$\|\mathbf{p} - \hat{\mathbf{p}}\|_1 \geq \Delta. \tag{15}$$

By Pinsker's inequality,

$$D_{\mathrm{KL}}(\mathbf{p}\|\hat{\mathbf{p}}) \geq \frac{1}{2}\|\mathbf{p} - \hat{\mathbf{p}}\|_1^2. \tag{16}$$

Therefore, any masked view that changes the teacher decision must satisfy:

$$D_{\mathrm{KL}}(\mathbf{p}\|\hat{\mathbf{p}}) \geq \frac{\Delta^2}{2}. \tag{17}$$

Equivalently, for every sampled mask ξ,

$$\mathbb{I}\left[\arg\max_c \hat{p}_c \neq y\right] \leq \frac{2D_{\mathrm{KL}}(\mathbf{p}\|\hat{\mathbf{p}})}{\Delta^2}. \tag{18}$$

Taking expectation over $\xi \sim \mathcal{M}_{r_s}$ gives

$$\Pr_{\xi}\left[\arg\max_c \hat{p}^T_{i,\xi,c} \neq y_i\right] \leq \frac{2\,\mathbb{E}_\xi\left[D_{\mathrm{KL}}\left(\mathbf{p}_i^T \| \hat{\mathbf{p}}_{i,\xi}^T\right)\right]}{(\Delta_i^T)^2}, \tag{19}$$

which proves Eq. (12). □

Relation to the masked self-distillation objective. Define the expected per-sample masked self-distillation objective as

$$\overline{\mathcal{L}}_{\mathrm{sd},i} = T^2\mathbb{E}_\xi\left[D_{\mathrm{KL}}\left(\mathbf{p}_i^T \| \hat{\mathbf{p}}_{i,\xi}^T\right)\right]. \tag{20}$$

Proposition 1 immediately yields

$$\Pr_{\xi}\left[\arg\max_c \hat{p}^T_{i,\xi,c} \neq y_i\right] \leq \frac{2\overline{\mathcal{L}}_{\mathrm{sd},i}}{T^2(\Delta_i^T)^2}. \tag{21}$$

The inner K_s-sample average in Eq. (8) is a Monte-Carlo estimator of $\overline{\mathcal{L}}_{\mathrm{sd},i}$ for each retained sample, while Eq. (8) further averages these sample-wise estimates over the selected teachers.

Interpretation. Proposition 1 gives masked self-distillation a decision-stability interpretation: reducing the expected teacher–student divergence tightens an upper bound on the probability that an internal masking perturbation changes the current full-view decision. The usefulness of such consistency depends on the reliability of the full-view teacher, since an unreliable prediction can likewise be propagated to its masked views. This motivates the reliability-aware teacher validation and consensus correction introduced in Secs. 4.1–4.2.

B.2 Additional Analytical Properties

We further provide several analytical properties that clarify the behavior of the vulnerability-aware consensus, consensus correction, and marginal-entropy stabilization used in RASD.

Log-odds interpretation of class-wise evidence. For any logits $\mathbf{s} \in \mathbb{R}^C$ and $\mathbf{p} = \mathrm{softmax}(\mathbf{s})$, the class-wise evidence in Eq. (4) satisfies:

$$\ell_c = s_c - \log \sum_{j \neq c} \exp(s_j) = \log \frac{p_c}{1 - p_c}. \tag{22}$$

Thus, ℓ_c measures the support for class c relative to the aggregate probability mass of all competing classes. Accordingly, the score $\mu_{i,c} - \lambda\sigma_{i,c}$ favors classes that maintain strong evidence with limited variation across internal perturbations.

Bounded consensus correction. For a selected sample, let $\mathbf{p}_i^T = \mathrm{softmax}(\mathbf{s}_i/T)$ and define

$$\mathcal{L}_{\mathrm{corr},i} = T^2 D_{\mathrm{KL}}\left(\mathbf{q}_i \| \mathbf{p}_i^T\right), \tag{23}$$

where $\mathbf{q}_i$ is detached during optimization. Its gradient with respect to the full-view logits is

$$\frac{\partial \mathcal{L}_{\mathrm{corr},i}}{\partial s_{i,c}} = T\left(p_{i,c}^T - q_{i,c}\right). \tag{24}$$

Since both $\mathbf{p}_i^T$ and $\mathbf{q}_i$ are probability distributions,

$$\left|\frac{\partial \mathcal{L}_{\mathrm{corr},i}}{\partial s_{i,c}}\right| \leq T, \qquad \left\|\nabla_{\mathbf{s}_i} \mathcal{L}_{\mathrm{corr},i}\right\|_1 \leq 2T. \tag{25}$$

Hence, the perturbation-derived consensus induces a bounded soft correction of the current predictive distribution.

Interpretation of batch-aware entropy stabilization. To interpret Eq. (9), we define the entropy-effective number of predicted classes as:

$$N_{\mathrm{eff}}(\bar{\mathbf{p}}_t) = \exp(\mathcal{H}(\bar{\mathbf{p}}_t)).$$

This quantity characterizes the effective class coverage of the batch-level predictive distribution in the entropy sense.

Let $B_t = |\mathcal{B}_t|$. The entropy reference in Eq. (9) can be equivalently expressed through the reference class scale

$$R_t = \exp(h_t) = \begin{cases} C, & C \leq B_t, \\ \sqrt{CB_t}, & C > B_t. \end{cases}$$

Accordingly,

$$\mathcal{L}_{\mathrm{bme}} = \left[\log \frac{R_t}{N_{\mathrm{eff}}(\bar{\mathbf{p}}_t)}\right]_+.$$

Therefore, $\mathcal{L}_{\mathrm{bme}}$ penalizes the logarithmic deficit between the reference class scale R_t and the entropy-effective class coverage of the current predictions. When $C \leq B_t$, the reference reduces to the full label-space scale C. When $C > B_t$, $R_t = \sqrt{CB_t}$ lies between the batch scale and the full class scale, *i.e.,* $B_t < R_t < C$, providing an intermediate diversity requirement rather than being limited solely by the batch size. The regularization becomes inactive once $N_{\mathrm{eff}}(\bar{\mathbf{p}}_t) \geq R_t$. This provides a batch-adaptive stabilization mechanism whose diversity requirement automatically scales with both the number of classes and the available test samples.

C Experimental Setups

C.1 Dataset Details

Corruption benchmarks. CIFAR-10-C and CIFAR-100-C (Hendrycks & Dietterich, 2019) apply 15 corruptions to the test sets of CIFAR-10/100 (Krizhevsky & Hinton, 2009) (noise: Gaussian, shot, impulse; blur: defocus, glass, motion, zoom; weather: snow, frost, fog, brightness; digital: contrast, elastic, pixelate, JPEG) at 5 severities with 10,000 images per severity. ImageNet-C (Hendrycks & Dietterich, 2019) applies the same 15 corruptions to the ImageNet (Deng et al., 2009) validation set with 50,000 images per severity. Following prior work (Maharana et al., 2025; Lafon et al., 2025), we evaluate at severity 5 and report the average over the 15 corruptions.

Fine-grained benchmarks. FGVC-Aircraft (Maji et al., 2013) (100 aircraft variants), Caltech101 (Fei-Fei et al., 2007), Stanford Cars (Krause et al., 2013) (196 car models), DTD (Cimpoi et al., 2014) (47 textures), EuroSAT (Helber et al., 2019) (10 satellite land-use classes), Flower102 (Nilsback & Zisserman, 2008) (102 flower species), Food101 (Bossard et al., 2014) (101 food classes), Oxford Pets (Parkhi et al., 2012) (37 pet breeds), SUN397 (Xiao et al., 2010) (397 scene classes), and UCF101 (Soomro et al., 2012) (frames of 101 actions). We use the test splits and class names of CoOp (Zhou et al., 2022b). The CoOp Caltech101 split removes the background category and contains 100 classes. The test-set sizes are 3,333 (Aircraft), 2,465 (Caltech101), 8,041 (Cars), 1,692 (DTD), 8,100 (EuroSAT), 2,463 (Flower102), 30,300 (Food101), 3,669 (Pets), 19,850 (SUN397), and 3,783 (UCF101).

Natural ImageNet shifts. ImageNet-A (Hendrycks et al., 2021b) contains 7,500 natural adversarial examples of 200 classes that are easy for humans yet systematically mislead standard classifiers through unusual poses, backgrounds, and imaging conditions. ImageNet-V2 (Recht et al., 2019) re-collects 10,000 test images following the original ImageNet protocol and represents a mild shift. ImageNet-R (Hendrycks et al., 2021a) contains 30,000 artistic renditions (paintings, sketches, cartoons, etc.) of 200 classes and tests style shift. ImageNet-Sketch (Wang et al., 2019) contains 50,889 black-and-white sketches of 1,000 classes and tests texture and shape shift.

Domain-generalization benchmarks. OfficeHome (Venkateswara et al., 2017) has 65 classes and about 15,500 images in the domains Art, Clipart, Product, and Real World. PACS (Li et al., 2017) has 7 classes in Art Painting, Cartoon, Photo, and Sketch. VLCS (Fang et al., 2013) consists of the 5 classes shared by Caltech101, LabelMe, SUN09, and PASCAL VOC2007. Terra Incognita (Beery et al., 2018) consists of camera-trap images of 10 classes at four locations (L38, L43, L46, L100); its shift stems mainly from background, illumination, and location-specific factors, with severe class imbalance, occlusion, and motion blur. We follow the DomainBed (Gulrajani & Lopez-Paz, 2021) splits and treat each domain as an independent test stream.

Evaluation protocol. Images are resized and center-cropped to 224×224 and normalized with the CLIP mean and standard deviation. Adaptation is online and label-free. For every test batch, the current parameters produce the reported prediction, and the same batch then supplies the unlabeled adaptation signal. The adaptable parameters, optimizer state, and any method-specific memory are reset before each corruption type and before each domain, so updates do not carry across these streams. The reported metric is top-1 accuracy. Corruption accuracy is the unweighted mean over the 15 types at severity 5, using the full CIFAR test sets (10,000 images per corruption) and the full ImageNet validation set (50,000 images per corruption). Domain averages are unweighted means over domains. The random seed is 42.

C.2 Implementation Details

Updated parameters and optimization. RASD updates the scale and bias of all LayerNorms in the visual encoder (including `ln_pre`, `ln_1`/`ln_2` of every block, and `ln_post`) and in the text encoder (including `ln_final`); all other backbone parameters are frozen. The filling vector $\boldsymbol{m}$ uses a separate Adam optimizer with learning rate 10^{-3} and no weight decay, and is initialized on the first batch of each stream as the mean of the visible patch tokens. LayerNorm parameters use Adam without weight decay at the learning rates in Table A. The model stays in evaluation mode, so no

Table A: Templates and hyperparameters of RASD. "Learning rate" lists the visual-encoder and text-encoder LayerNorm rates; a single value means the two encoders share it. λ is the fragility penalty, and λ_{sd}, λ_{corr}, and λ_{bme} weight the three terms of the adaptation objective. T=2 except on DTD, where T=1.

Dataset	Template	Learning rate	K_s	r_s	r_c	λ	λ_{sd}	λ_{corr}	λ_{bme}
Aircraft	"A photo of a {}, a type of aircraft."	10^{-4}	16	15%	0.5%	1	1	0.5	2.2
Caltech101	"A photo of a {}."	2×10^{-4} / 7×10^{-4}	16	15%	0.5%	1	1	1	1
Cars	"A photo of a {}."	1.5×10^{-4} / 2×10^{-4}	16	15%	0.5%	1	0.5	0.5	1
DTD	"{} texture."	2.5×10^{-4}	16	15%	0.5%	1	1	0.5	1
EuroSAT	"A centered satellite photo of {}."	2.4×10^{-4} / 7×10^{-4}	2	10%	0.5%	2	1	1	1
Flower102	"A photo of a {}, a type of flower."	10^{-4} / 4×10^{-4}	16	15%	0.5%	1	0.5	0.5	1.69
Food101	"A photo of a {}, a type of food."	5×10^{-5} / 3×10^{-5}	2	8%	0.8%	2.3	1	1	1
Pets	"A photo of a {}, a type of pet."	4×10^{-4} / 3×10^{-4}	16	15%	0.5%	1	1	0	1.6
SUN397	"A photo of {}."	10^{-4}	16	15%	0.5%	1	0	1	1
UCF101	"A photo of a person doing {}."	2×10^{-4}	2	8%	0.8%	1	0.5	0.5	3
ImageNet-A	"a photo of a {}."	2.3×10^{-4}	16	15%	0.5%	1	1.8	1	1
ImageNet-V2	"a photo of a {}."	1.2×10^{-4}	4	15%	0.7%	1	1	1	1
ImageNet-R	"a photo of a {}."	10^{-4}	2	12%	0.5%	1	1	1	0.8
ImageNet-Sketch	"a photo of a {}."	5×10^{-5} / 3×10^{-5}	2	8%	0.8%	2.4	1	1	1.4
OfficeHome	"a photo of a {}"	3×10^{-4}	16	15%	0.5%	1	1	1	1
PACS	"a photo of a {}"	5×10^{-5}	16	15%	0.5%	1	1	1	1
VLCS	"a photo of a {}"	5×10^{-6}	2	12%	5%	1	0.96	1	1
Terra Incognita	"a photo of a {}"	1.8×10^{-4}	16	15%	0.5%	1	1	1	1
CIFAR-10-C	"a photo of a {}."	4×10^{-4} / 10^{-4}	2	15%	10%	1	1	1	1
CIFAR-100-C	"a photo of a {}."	1.5×10^{-4}	16	0.5%	0.5%	1	1	1	1.4
ImageNet-C	"a photo of a {}."	5×10^{-5} / 10^{-5}	2	0.5%	0.5%	1	1	1	1

normalization statistics are updated. The random seed is 42 and deterministic cuDNN computation is enabled.

Probes, masks, and teacher selection. Masks are re-sampled at every adaptation step. Each patch token independently drops $\text{round}(rD)$ feature dimensions (D=768 for ViT-B/16); the `[CLS]` token is never masked. Consensus probes are evaluated without gradient. Student views keep gradients with respect to the input tokens and $\boldsymbol{m}$. All K_c probes of a batch are forwarded together. Throughout the reported experiments, K_c=16, the temperature is T=2 except on DTD where T=1, and teacher selection retains samples whose full-view decision agrees with the consensus and whose internal support lies in the top ρ=0.7 fraction of the batch. The batch size is B=128 and each batch is adapted for S=10 steps. The objective used in the experiments is $\lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{sd}}\mathcal{L}_{\text{sd}} + \lambda_{\text{bme}}\mathcal{L}_{\text{bme}}$. The marginal-entropy threshold is the closed form in Eq. equation 9 at B=128, applied to the marginal prediction of the current batch. On CIFAR-100-C and ImageNet-C the selected student ratio equals the probe ratio; on the other datasets the student mask is stronger.

Dataset-specific settings. Table A lists the prompt template, LayerNorm learning rates, mask settings, fragility penalty λ, and the three loss coefficients. A single learning rate means that the visual and text encoders share it. Fine-grained datasets use the dataset-specific templates of CoOp (Zhou et al., 2022b). The ViT-L/14 CIFAR-C experiments reuse the ViT-B/16 hyperparameters of the corresponding dataset.

D Settings of the Compared Methods

Every baseline uses the OpenAI CLIP ViT-B/16 checkpoint and the same test streams as RASD. Adaptable parameters and method-specific memory are reset before each corruption type and each domain. We take hyperparameters from the officially released configuration files or scripts. A benchmark that is absent from a release uses the transferred configuration named in the corresponding paragraph, with no search on target labels. Unless a method fixes its own prompts, the text template is the one used by zero-shot CLIP on that dataset (Table A).

Zero-shot CLIP. The visual and text encoders stay frozen. Class names are inserted into the dataset template and encoded once. This is the reference row marked CLIP (ZS).

TENT. TENT (Wang et al., 2021) minimizes the entropy of the current prediction and updates only the affine parameters of the visual-encoder normalization layers, for one step per batch. On CIFAR-10-C and CIFAR-100-C the optimizer is Adam with learning rate 10^{-3} and batch size 200. On ImageNet-C it is SGD with learning rate 2.5×10^{-4}, momentum 0.9, and batch size 64. The prompt is "a photo of a {}.".

RoTTA. RoTTA (Yuan et al., 2023) stores test samples in a memory bank, reweights them by timeliness, and trains against an exponential-moving-average teacher. The CLIP port updates LayerNorm affine parameters. The released configuration uses memory size 64, an update every 64 samples, EMA coefficient ν=0.001, and loss weights λ_t=λ_u=1. The optimizer is Adam with learning rate 10^{-3}. The batch size is 200 on CIFAR-C and 64 on ImageNet-C.

SAR. SAR (Niu et al., 2023) combines reliable-sample selection with sharpness-aware entropy minimization. The entropy gate is E_0=$0.4\ln C$, the recovery threshold is e_0=0.2, the SAM radius is ρ=0.05, and the entropy moving-average factor is 0.9. The optimizer and batch size match TENT: Adam with learning rate 10^{-3} and batch size 200 on CIFAR-C, and SGD with learning rate 2.5×10^{-4} and batch size 64 on ImageNet-C.

TPT. TPT (Shu et al., 2022) tunes a prompt independently on each image and then resets it. Each image is expanded to 64 views, the original view plus 63 augmentations. The 10% of views with the lowest entropy are kept, and one AdamW step is taken with learning rate 5×10^{-3} and weight decay 0.01. The original center-cropped view is evaluated after this step. ImageNet-C uses random resized crops and flips; CIFAR-C additionally uses the official AugMix policy.

WATT-S and WATT-P. WATT (Osowiechi et al., 2024) adapts the visual-encoder LayerNorms under several text templates and averages the resulting weights. Each batch is episodic. The eight released templates are "a photo of a {}", "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 {}.". Both variants take L=2 gradient steps per template with Adam and batch size 128. WATT-S walks through the templates sequentially and repeats the average M=5 times. WATT-P adapts the templates in parallel and repeats the average M=50 times. The released CIFAR corruption scripts set the WATT-S learning rate to 10^{-3}. WATT-P and the OfficeHome, PACS, and VLCS scripts use 10^{-4}. ImageNet-C uses the CIFAR script. Fine-grained recognition and the natural ImageNet shifts, which have no released script, use the domain-generalization learning rate 10^{-4} for WATT-S.

CLIPArTT. CLIPArTT (Hakim et al., 2025) builds a text prompt from the top-K=3 predicted class names and minimizes a soft image–text contrastive loss at temperature τ=0.01. It updates the visual-encoder LayerNorms with Adam at learning rate 10^{-3} for 10 steps, then resets the weights. The batch size is 128. The same configuration is used on corruption benchmarks and on the natural ImageNet shifts.

BCA. BCA (Zhou et al., 2025) maintains class prototypes and a cluster-to-class prior, and updates them after predicting each image. The batch size is 1 and the classifier temperature is 100. The ten recognition datasets use the released ViT-B/16 configuration files, including the authors' CuPL descriptions in addition to the dataset template. Table B lists the two confidence gates and their initial counts. CIFAR-C and ImageNet-C are outside that release. On those benchmarks we use the out-of-distribution setting reported in the paper, τ=0.3, n_1=30000, and n_2=10, for both gates, with the dataset template only.

TCA. TCA (Wang et al., 2025) is training-free. It condenses visual tokens and adjusts the prediction with a class-wise reservoir of cached features. Images are processed one at a time, with a single view. The released executable uses the `Ours-0.1` pruning schedule on Transformer blocks 3, 6, and 9, and four coreset centers. The official dataset files do not cover CIFAR-C, ImageNet-C, or the four domain-generalization benchmarks. Those runs use the released ImageNet cache settings: reservoir size 3, scale 0.1, λ=3, and β=5.

Table B: Released ViT-B/16 gate settings of BCA on the ten recognition datasets. τ_1, n_1 and τ_2, n_2 are the two confidence gates and their initial counts.

Dataset	τ_1	n_1	τ_2	n_2
Aircraft	0.10	55000	0.10	10
Caltech101	0.05	20000	0.65	1
Cars	0.45	70000	0.40	5
DTD	0.05	45000	0.30	85
EuroSAT	0.30	40000	0.65	95
Flower102	0.30	15000	0.60	5
Food101	0.45	50000	0.25	95
Pets	0.35	20000	0.60	5
SUN397	0.40	70000	0.10	40
UCF101	0.35	50000	0.35	10

BATCLIP. BATCLIP (Maharana et al., 2025) updates the LayerNorms of both encoders for one step and aligns image features with text class prototypes. The prompt is "a photo of a {}.". CIFAR-10-C uses AdamW with learning rate 10^{-3}, weight decay 0.01, and batch size 200. CIFAR-100-C uses Adam with learning rate 5×10^{-4}, weight decay 0.01, and batch size 200. ImageNet-C uses AdamW with learning rate 5×10^{-4}, weight decay 0.01, and batch size 64. Recognition, natural-shift, and domain benchmarks have no separate released file, so they use the ImageNet-C configuration.

Mint. Mint (Bao et al., 2025) updates the visual normalization layers so that corruption-degraded inter-class feature variance is restored. A class-conditional feature accumulator with prior strength 10,000 is kept inside a domain, while the network weights are restored after the post-update prediction of each batch. The text encoder is an average of seven official templates: "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 {}.". There is one Adam step per batch, and the batch size is 20. CIFAR-10-C and CIFAR-100-C use learning rate 0.007. ImageNet-C and the natural ImageNet shifts use learning rate 0.015. The official ImageNet-C script evaluates a 5,000-image subset; we keep its hyperparameters and evaluate the full 50,000-image validation set.

BITTA. BITTA (Sun et al., 2026) uses low-entropy samples for learning and high-entropy samples for unlearning, with the unlearning step multiplied by 1.2. Normalization parameters of both encoders are updated and the logit scale stays frozen. The prompt is "a photo of a {}.". The optimizer is AdamW with β_1=0.9, weight decay 0.01, and one step per batch. CIFAR-10-C uses learning rate 10^{-3} and batch size 200. CIFAR-100-C uses learning rate 5×10^{-4} and batch size 200. ImageNet-C uses learning rate 5×10^{-4} and batch size 64. Recognition, natural-shift, and domain benchmarks use this ImageNet-C configuration.

LMR. LMR (Huang et al., 2026) restores the margin around plausible near-top classes and corrects stream-level bias. Only the visual-encoder LayerNorms are updated. The text encoder is frozen, the prompt is "a photo of a {}.", and each batch takes one Adam step at learning rate 10^{-3} with no weight decay. The batch size is 64. The shared coefficients are a positive-set gap of 0.03, a confidence threshold of 0.3, five hard negatives, a bias step of 0.05, bias momentum 0.8, and a bias clip of 2.0. CIFAR-10-C uses margin coefficient γ=0.3 and class-prior coefficient 3. CIFAR-100-C and ImageNet-C use γ=4 and class-prior coefficient 1. Recognition, natural-shift, and domain benchmarks use the ImageNet-C configuration.

E Additional Results

Does internal consistency indicate prediction reliability? We further examine whether the consistency revealed by internal probes provides a meaningful indicator of prediction reliability. Before any adaptation, we divide zero-shot CLIP predictions according to whether the full-view prediction agrees with the consensus obtained from the weak internal probes. As shown in Table C, consistent predictions are substantially more accurate than inconsistent ones across different distribution shifts. For example, the accuracy differs by 39.93% on ImageNet-R, 34.66% on Cars, and 46.64% on OfficeHome. Similar gaps are observed on corruption, natural-shift, and recognition benchmarks.

Table C: Accuracy (%) of consistent and inconsistent full-view predictions under internal probing. Statistics are computed on zero-shot CLIP ViT-B/16 before adaptation.

Dataset	Consistent	Inconsistent	Gap
CIFAR-10-C	61.37	37.54	+23.83
ImageNet-V2	61.26	28.97	+32.29
ImageNet-R	74.28	34.35	+39.93
Cars	65.84	31.18	+34.66
Food101	85.87	35.96	+49.91
OfficeHome	82.76	36.12	+46.64

Table D: Sensitivity to the number of consensus probes K_c, probe mask ratio r_c, and student mask ratio r_s. Mean accuracy (%) over the 15 corruptions at severity 5 using CLIP ViT-B/16. † denotes the default setting for each dataset. Best results in each row are in bold.

K_c	4	8	16†	32	
CIFAR-10-C	81.95	81.92	**81.97**	81.95	
CIFAR-100-C	53.67	**53.74**	53.69	53.66	
r_c	1%	5%	10%†	15%	
CIFAR-10-C	81.87	81.90	**81.97**	81.44	
r_c	0.1%	0.5%†	1%		
CIFAR-100-C	**53.75**	53.69	53.53		
r_s	5%	10%	15%†	20%	25%
CIFAR-10-C	81.62	81.91	**81.97**	81.96	81.85
r_s	0.5%†	1%	5%	10%	15%
CIFAR-100-C	53.69	**53.82**	53.79	53.74	53.69

These results provide direct empirical support for using internal perturbation consistency to assess whether a full-view prediction is sufficiently reliable to serve as test-time supervision.

Sensitivity to the number of probes and the mask ratios. Table D evaluates the sensitivity of RASD to the number of consensus probes K_c, the probe mask ratio r_c used for weak internal perturbations, and the student mask ratio r_s used for masked self-distillation. Varying K_c from 4 to 32 causes only minor changes on both CIFAR-10-C and CIFAR-100-C, showing that the consensus is not sensitive to the exact number of probes. The probe mask ratio r_c is also stable within the tested range, with only small performance variations on both benchmarks. Similarly, varying the student mask ratio r_s over a broad range yields consistently comparable results. Overall, RASD exhibits low sensitivity to these hyper-parameters, indicating that its performance is robust to reasonable choices of the probe number and masking strengths.

Effect of updating the text-side LayerNorms. Table E compares updating only the visual-encoder LayerNorms with jointly updating the LayerNorms of both the visual and text encoders. Joint adaptation achieves 81.97%, 53.69%, and 40.03% on CIFAR-10-C, CIFAR-100-C and ImageNet-C, respectively, compared with 81.67%, 52.81%, and 38.91% when updating the visual encoder alone. The consistent gains on both benchmarks indicate that adapting the text-side normalization provides complementary benefits to visual-side adaptation.

Table E: Effect of updating the text-side LayerNorms on the CIFAR-10-C, CIFAR-100-C, and ImageNet-C datasets.

	CIFAR-10-C	CIFAR-100-C	ImageNet-C
Vision LN only	81.67	52.81	38.91
Vision + Text LN	**81.97**	**53.69**	**40.03**

Effect of the batch size. Table F complements Figure 3 with per-corruption results for batch sizes ranging from 16 to 512. Increasing the batch size generally improves performance on both benchmarks, with the gains becoming much smaller once the batch size reaches 128. On CIFAR-10-C, the average accuracy increases from 76.97% at $B = 16$ to 81.97% at the default $B = 128$, and remains around 82% for larger batches. A similar trend is observed on CIFAR-100-C, where the accuracy improves from 49.75% to 53.69% and changes only slightly when further increasing the batch size.

Table F: Per-corruption accuracy (%) under different batch sizes on CIFAR-10-C and CIFAR-100-C at severity 5 using CLIP ViT-B/16. Avg denotes the mean over the 15 corruption types, and † marks the default batch size. Best and second-best results are in bold and underlined, respectively.

Batch size	Gauss.	Shot	Impu.	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brit.	Cont.	Elas.	Pixl.	Jpeg	Avg
CIFAR-10-C																
B=16	65.94	71.00	72.81	79.08	66.60	80.15	80.67	82.85	82.75	82.46	86.48	84.20	72.39	77.12	70.07	76.97
B=32	66.25	72.51	74.31	82.88	69.16	82.41	83.18	86.08	85.36	84.30	89.93	86.72	75.87	81.06	71.65	79.44
B=64	68.75	73.37	78.11	83.35	70.07	84.14	85.29	87.11	87.11	86.73	91.03	89.07	77.20	81.89	73.69	81.13
B=128†	70.63	74.53	**79.55**	84.46	**72.54**	84.90	86.26	87.59	86.91	87.62	91.72	89.24	77.15	82.52	74.00	81.97
B=256	**71.08**	**74.75**	79.04	**84.86**	71.75	**85.34**	86.28	**88.03**	**87.54**	**87.71**	**92.17**	**89.64**	77.43	82.80	**74.83**	**82.22**
B=512	70.92	74.49	78.50	84.82	72.01	85.00	**86.57**	87.51	**87.54**	87.59	92.01	89.21	**77.44**	**83.30**	74.64	82.10
CIFAR-100-C																
B=16	37.10	38.25	44.24	53.81	36.61	53.99	56.25	57.59	56.71	52.93	65.71	58.26	43.80	48.69	42.28	49.75
B=32	38.07	40.76	46.99	54.78	37.87	55.01	58.28	58.83	59.06	56.60	67.44	59.96	45.04	50.05	42.60	51.42
B=64	38.92	41.47	48.19	56.57	39.57	56.78	59.94	60.00	59.97	58.18	68.84	61.63	46.67	52.53	44.88	52.94
B=128†	40.60	**42.94**	**49.16**	57.16	**40.05**	56.95	61.10	60.18	60.64	58.81	69.54	**62.75**	47.34	52.98	45.08	53.69
B=256	**40.99**	42.84	48.51	57.71	39.98	**57.59**	**61.54**	**60.31**	**60.67**	**59.08**	69.34	**62.75**	**47.55**	**53.00**	**45.83**	**53.85**
B=512	40.70	42.49	48.15	**57.83**	39.29	57.54	61.37	60.20	60.58	58.57	**69.93**	61.96	47.26	52.78	45.33	53.60

The per-corruption results show a broadly consistent trend across different corruption types. We therefore use $B = 128$ as the default setting in the main experiments.

Effect of the fill mode. Table G studies how the masked feature dimensions should be replaced when adaptation is performed using only the reliability-aware self-distillation objective $\mathcal{L}_{\text{sd}}$, with teacher selection enabled. Using fixed replacements such as zeros or the `[CLS]` token severely degrades performance, suggesting that such values introduce overly disruptive changes to the masked internal representation. Replacing the masked dimensions with the mean of visible patch tokens is considerably more stable, but still underperforms zero-shot CLIP. In contrast, the learnable filling vector $\boldsymbol{m}$, initialized from the patch mean and optimized together with the adaptable parameters, achieves 64.34% and 38.75% on CIFAR-10-C and CIFAR-100-C, respectively. These results indicate that a learnable fill representation provides a more suitable substitute for masked feature dimensions and leads to more effective masked self-distillation.

Table G: Comparison of fill modes for masked feature dimensions on CIFAR-10-C and CIFAR-100-C at severity 5 using CLIP ViT-B/16. $\boldsymbol{m}$ is initialized with the mean of visible patch tokens.

Fill Mode	CIFAR-10-C	CIFAR-100-C
CLIP (ZS)	61.18	35.78
Zeros	24.97	4.78
CLS token	32.06	11.52
Patch tokens mean	53.15	25.02
Learnable $\boldsymbol{m}$	**64.34**	**38.75**

Effect of the reliability gate. We further examine the effect of the reliability gate δ_i used for teacher selection. As defined in Eq. (6), δ_i first requires agreement between the full-view prediction and the consensus decision, and further filters samples according to the internal support score w_i. We set $\rho = 0.7$, such that at most the top 70% samples with stronger internal support are retained for adaptation. As shown in Table H, removing the gate reduces the average accuracy from 81.97% to 80.24% on CIFAR-10-C and from 53.69% to 53.55% on CIFAR-100-C. These results demonstrate that the reliability gate improves adaptation by filtering less-supported predictions, while also reducing unnecessary computation by restricting subsequent correction and self-distillation to the selected samples.

Table H: Effect of the reliability gate on CIFAR-10-C and CIFAR-100-C. Mean accuracy (%) over the 15 corruptions at severity 5 using CLIP ViT-B/16.

Gate δ_i	CIFAR-10-C	CIFAR-100-C
×	80.24	53.55
✓	**81.97**	**53.69**

Effect of larger visual backbone (*e.g.*, ViT-L/14). Table I further evaluates RASD with the stronger ViT-L/14 backbone on CIFAR-10-C and CIFAR-100-C. RASD achieves the highest average accuracy on both benchmarks, reaching 87.39% and 64.22%, respectively, and outperforming LMR by 1.16% and 1.07%. The improvement is also consistently observed across a broad range of corruption types, showing that the benefit of RASD is preserved when scaling to a stronger visual backbone.

Table I: Per-corruption accuracy (%) on CIFAR-10-C and CIFAR-100-C at severity 5 using CLIP ViT-L/14. Best and second-best results are in bold and underlined, respectively.

Method	Gauss.	Shot	Impu.	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brit.	Cont.	Elas.	Pixl.	Jpeg	Avg
CIFAR-10-C																
CLIP (ZS)	64.35	67.12	75.94	80.46	51.59	80.32	83.04	83.23	84.98	79.17	91.05	84.10	66.51	73.67	72.31	75.86
BATCLIP	74.80	78.41	84.60	87.73	72.06	86.64	89.70	89.80	90.30	87.27	94.87	92.44	79.02	83.70	79.99	84.76
LMR	78.82	81.42	86.39	89.29	73.80	87.56	90.71	**90.54**	**90.72**	89.15	**95.65**	**94.09**	79.67	85.00	80.63	86.23
RASD (Ours)	**80.53**	**82.89**	**87.09**	**90.18**	**79.58**	**88.99**	**90.78**	90.23	89.95	**90.55**	94.60	92.66	**82.64**	**87.50**	**82.63**	**87.39**
CIFAR-100-C																
CLIP (ZS)	35.52	37.67	50.27	50.18	26.01	52.52	57.13	54.90	58.90	49.58	67.14	53.46	36.40	44.83	42.74	47.82
BATCLIP	40.09	43.70	53.67	55.43	30.21	54.03	60.24	57.96	59.04	49.91	71.14	61.46	36.20	47.83	47.56	51.23
LMR	51.72	54.37	63.96	66.61	48.04	65.96	69.98	68.59	68.86	67.31	77.58	**74.97**	53.11	61.09	55.05	63.15
RASD (Ours)	**53.97**	**54.62**	**64.95**	**67.64**	**51.47**	**66.80**	**70.09**	**69.52**	**69.88**	**67.74**	**78.53**	74.82	**54.10**	**62.46**	**56.68**	**64.22**

These results indicate that the proposed reliability-aware adaptation is not specific to ViT-B/16 and generalizes well across different model capacities.