DECOUPLED STREAMS FOR OPEN-SET TEST-TIME ADAPTATION OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-set test-time adaptation requires a model to improve classification on in-distribution (ID) samples while reliably rejecting unknown out-of-distribution (OOD) inputs. Existing approaches for vision-language models optimize these two objectives through the same trainable parameters, implicitly assuming that their updates remain compatible throughout adaptation. We show that this compatibility progressively weakens during adaptation: classification and ID/OOD separation gradients become less aligned, while recent gradient-rebalancing and modular adapter-routing strategies do not consistently improve both objectives jointly. This motivates a simple, yet effective alternative: rather than reconciling the objectives within a shared adaptation space, we separate their optimization. We introduce **DSTA** (Decoupled Streams for Open-Set Test-Time Adaptation), which equips vision-language models with two independent sets of low-rank adapters for classification and ID/OOD separation over a shared frozen backbone, while retaining their interaction through OOD-guided sample selection. In this decoupled regime, DSTA stabilizes the separation stream with a novel margin-based objective that prevents excessive ID/OOD gap increase, and proposes automatic OOD logit-scale calibration to avoid compressed confidence scores. Across clean and corrupted open-set benchmarks, DSTA improves the joint behavior of ID classification and OOD detection, with particularly strong gains in unknown-sample rejection under distribution shift. These benefits remain evident across backbone sizes, supporting parameter-space decoupling as an effective design principle for open-set test-time adaptation. Code will be released upon acceptance.

1 INTRODUCTION

Vision-language models (VLMs) learn a joint representation of visual inputs and natural-language concepts, enabling zero-shot recognition without task-specific supervision. Nevertheless, their predictions can deteriorate when the test distribution differs from the data encountered during pretraining, e.g., due to domain shifts or image corruptions (Hakim et al., 2025; Lafon et al., 2025). Test-time Adaptation (TTA) addresses this by updating a pretrained model using unlabeled samples from the target distribution at inference time. This allows the model to automatically adjust to the observed shift without requiring target data, source data, or ground-truth labels (Wang et al., 2021).

Many TTA methods assume a closed-set scenario, where test samples belong to the classes known by the model (Wang et al., 2021; Niu et al., 2022; 2023; Karmanov et al., 2024). However, real-world test streams can also contain samples from unknown classes. Open-set TTA addresses this challenging setting by requiring the model to improve classification on in-distribution (ID) samples while simultaneously identifying out-of-distribution (OOD) inputs (Lee et al., 2023). These objectives are tightly coupled during adaptation: OOD samples should be excluded from the classification update, yet incorrectly assigning them to known classes can produce erroneous predictions and contaminate subsequent model updates. Reliable OOD detection is therefore essential not only for evaluation, but also for determining which samples can safely contribute to adaptation (Lee et al., 2023). Recent methods address this challenge by combining a task-adaptation objective with mechanisms for estimating or improving ID/OOD separation (Gao et al., 2024; Yu et al., 2024; Dong et al., 2025).

This problem is particularly relevant for VLMs such as CLIP (Radford et al., 2021), whose image-text similarities support both zero-shot classification and confidence-based OOD detection. Existing

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.

open-set TTA approaches for CLIP jointly optimize classification and ID/OOD separation (Sreenivas & Biswas, 2025; Lafon et al., 2025). In our work, we investigate how these objectives interact during test-time optimization. Despite their complementary goals, classification and ID/OOD separation are generally optimized within the same adaptation space (Gao et al., 2024; Lafon et al., 2025): this design implicitly assumes that updates beneficial to one objective are also compatible with the other.

We question this assumption by studying the optimization dynamics of CLIP during open-set TTA. We show that the alignment between the gradients induced by the classification and separation objectives progressively decreases during adaptation. This reveals a previously overlooked source of interference in open-set TTA, that is: *two distinct adaptation problems are optimized through the same trainable parameters.* In this respect, we evaluate gradient-conflict mitigation methods in the open-set TTA setting and show that they do not fully resolve the resulting ID/OOD trade-off. Consequently, rather than attempting to reconcile the two objectives within a shared parameter space, we investigate whether their interaction can be avoided entirely through parameters' decoupling.

This analysis led us to design **DSTA** (**D**ecoupled **S**treams for Open-Set **T**est-Time **A**daptation), which equips CLIP with two independent sets of low-rank adapters (Hu et al., 2022) for classification and ID/OOD separation over a shared frozen backbone. By assigning the two objectives to disjoint trainable parameters, DSTA prevents direct gradient interference while retaining their functional dependency through OOD-guided sample selection. In this decoupled regime, DSTA stabilizes the separation stream with a novel margin-based objective that prevents excessive ID/OOD gap growth, and proposes automatic OOD logit-scale calibration to improve ID/OOD discrimination.

We evaluate DSTA across diverse domains, from standard object recognition to egocentric and exocentric imagery, under clean semantic OOD and corruption-induced shifts. DSTA achieves the strongest aggregate performance across classification and OOD detection, with particularly strong gains in unknown-sample rejection under corruption. Results support our central hypothesis that explicitly isolating the optimization of classification and OOD separation provides an effective foundation for robust open-set adaptation of CLIP models.

The main contributions of our work are the following.

- We identify optimization interference between classification and ID/OOD separation in shared-parameter open-set TTA, showing that their gradients become progressively less aligned during adaptation. We show that representative gradient-rebalancing and modular adapter-routing strategies do not consistently improve both objectives jointly.
- We introduce DSTA, a parameter-decoupled method that prevents classification and ID/OOD separation from competing over the same trainable parameters while preserving OOD-guided sample selection. We propose a margin-based separation objective and automatic OOD logit-scale calibration to stabilize independent separation optimization.
- We achieve state-of-the-art aggregate performance across ID classification and OOD detection in open-set TTA, including settings with covariate perturbations beyond standard clean semantic OOD settings. The benefits remain evident across datasets, backbone sizes and shift types.

2 Related work

Open-set Test-Time Adaptation for VLMs. Open-set TTA extends conventional TTA by detecting OOD samples and limiting their influence during adaptation. Existing methods largely rely on sample filtering: OSTTA (Lee et al., 2023) filters samples whose confidence decreases during adaptation; SoTTA (Gong et al., 2023) stores high-confidence samples selected by maximum predicted probability and adapts on them; STAMP (Yu et al., 2024) combines consistency and entropy filtering with class-balanced memory and self-weighted entropy minimization; and UniEnt (Gao et al., 2024) fits a two-component Gaussian mixture to outlier scores, minimizing entropy on estimated ID samples and maximizing it on OOD samples. CLIPTTA (Lafon et al., 2025) targets VLMs specifically with image-text contrastive adaptation and outlier contrastive exposure for ID/OOD separation. Single-image VLM methods include ROSITA (Sreenivas & Biswas, 2025), combining feature banks and contrastive learning, and Open-IRT (Peng et al., 2025), combining prompt-based OOD filtering with adaptation. These works are not directly comparable to our batch-based setting, which follows established protocols (Yu et al., 2024; Gao et al., 2024; Lafon et al., 2025). Moreover, DSTA focuses on objective interaction, using decoupled adapters for classification and separation updates.

Gradient Conflict Mitigation in Multi-Task Learning. Gradient-conflict mitigation is relevant to open-set TTA when classification and ID/OOD separation share adaptation parameters. PCGrad (Yu et al., 2020) projects conflicting components of negatively aligned task gradients; CAGrad (Liu et al., 2021) constructs a conflict-averse update balancing average and task-specific improvements; and FairGrad (Ban & Ji, 2024) allocates optimization progress through a fairness-based utility formulation. PIVRG (Qin et al., 2025b) prioritizes tasks by performance drops relative to single-task baselines, requiring task-level performance estimates unavailable in unlabeled TTA, while ConsMTL (Qin et al., 2025a) assumes predefined shared and task-specific parameter groups. For empirical comparison, we use methods that can be applied to TTA without additional supervision or changes to the test-time protocol. We are the first to systematically evaluate gradient-conflict mitigation in open-set TTA, and to prevent interference by assigning classification and separation to disjoint trainable parameters.

Modular and Parameter-Efficient Multi-Task Architectures. Prior work controls cross-task parameter sharing through learned or modular structures. Ruder et al. (Ruder et al., 2019) learn the structure of parameter sharing, while HyperFormer (Mahabadi et al., 2021) generates task-specific adapters with a shared task-conditioned hypernetwork and AdapterFusion (Pfeiffer et al., 2021) combines task-specific adapters through a learned composition stage. More recent modular approaches include Polytropon (Ponti et al., 2022), which routes tasks over reusable adapter modules, and MHR (Page-Caccia et al., 2023), which provides finer-grained routing over adapter parameters. We evaluated these methods and show that DSTA better directly isolates classification and ID/OOD separation parameters because their roles are known a priori.

3 Gradient Interference in Open-set TTA

To analyze whether shared optimization induces interference in open-set TTA, we consider a setting where both classification and ID/OOD separation are optimized over the same trainable parameters ψ. Specifically, we define $\mathcal{L}_{\text{cls}}$ as the classification objective on ID samples, and $\mathcal{L}_{\text{sep}}$ as the objective distinguishing ID from OOD inputs, with related gradients $g_{\text{cls}} = \nabla_\psi \mathcal{L}_{\text{cls}}$ and $g_{\text{sep}} = \nabla_\psi \mathcal{L}_{\text{sep}}$. Following prior work on gradient conflict (Yu et al., 2020), we quantify their compatibility through cosine similarity $\cos(g_{\text{cls}}, g_{\text{sep}})$, which measures the extent to which updates that improve one objective also support the other. Larger values indicate stronger compatibility, whereas values closer to zero indicate weaker alignment and reduced mutual support. Using a first-order Taylor expansion, a classification update $\psi' = \psi - \eta g_{\text{cls}}$ ($\eta > 0$ is the step size) rewrites the separation objective as $\mathcal{L}_{\text{sep}}(\psi - \eta g_{\text{cls}}) = \mathcal{L}_{\text{sep}}(\psi) - \eta g_{\text{sep}}^\top g_{\text{cls}} + \mathcal{O}(\eta^2)$. When $g_{\text{sep}}^\top g_{\text{cls}}$ approaches zero, the first-order benefit of a classification update to the separation objective diminishes, indicating progressively weaker mutual support between the two objectives. At zero, the classification update has no effect on the separation objective, while a negative inner product indicates direct gradient conflict. Fig. 1 (top row) shows this interaction in a typical open-set TTA setup, where ImageNet is ID and Places365 is OOD (Cao et al., 2025; Lafon et al., 2025). The cosine similarity remains positive throughout the initial trajectory but progressively decreases after the initial adaptation stages, indicating that the two objectives become less aligned and provide diminishing mutual support during optimization.

We further analyze how the overall optimizer update supports each objective. Let d denote the corresponding descent direction, with parameter update $\psi' = \psi - \eta d$. For objective $i \in \{\text{cls}, \text{sep}\}$, the inner product $g_i^\top d$ characterizes the first-order decrease induced by ψ': larger positive values indicate stronger local improvement of objective i (Liu et al., 2021). Since $g_i^\top d$ is scale-dependent, we normalize it to cosine similarity $\cos(g_i, d)$ to isolate directional alignment and enable fair comparisons across objectives and update rules. This is used to compare the standard shared-parameter update with gradient-conflict mitigation methods, including PCGrad (Yu et al., 2020), CAGrad (Liu et al., 2021) and FairGrad (Ban & Ji, 2024). Because these methods differ in how they combine or modify the two shared gradients, this comparison allows us to assess whether gradient rebalancing improves the alignment between classification and separation updates. Fig. 1 (bottom row) shows the update directions exhibit different degrees of alignment with the classification and separation gradients under the considered optimization strategies. PCGrad (Yu et al., 2020) modifies the task gradients only when their pairwise cosine similarity is negative, resulting in limited rebalancing in this setting. CAGrad (Liu et al., 2021) increases the alignment with the separation gradient, whereas FairGrad (Ban & Ji, 2024) produces a more balanced alignment between the two. These

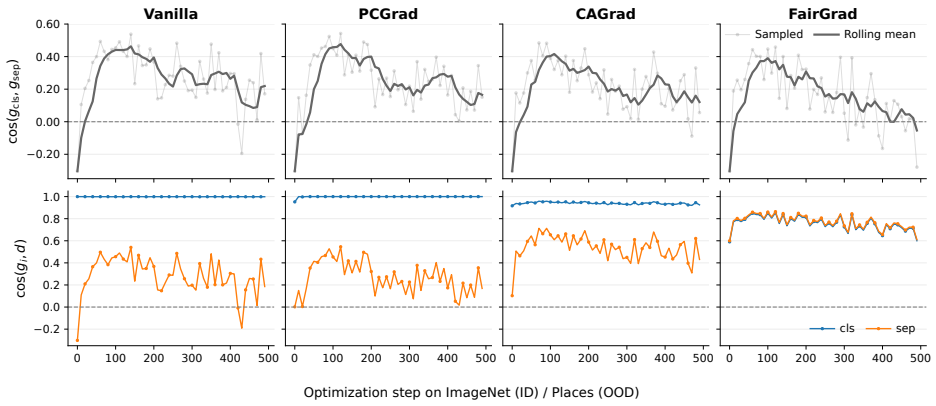

Figure 1: **Gradient interference in shared-parameter open-set TTA.** Classification (*cls*) and ID/OOD separation (*sep*) over the 500 optimization steps on ImageNet (ID)/Places365 (OOD). **Top:** cosine similarity between *cls* and *sep* gradients remains positive but decreases, indicating progressively weaker compatibility. **Bottom:** cosine similarity between each objective gradient g_i and descent direction d: for the vanilla joint update and PCGrad, d remains nearly collinear with g_{cls}, whereas CAGrad and especially FairGrad rebalance it toward g_{sep}.

observations show that gradient-rebalancing methods modify how a shared update supports the two objectives, but still fail to improve mutual alignment between classification and separation gradients.

Finally, we evaluate whether these changes in update alignment translate into improved ID classification and OOD detection. We apply these gradient-rebalancing methods to a recent state-of-the-art shared-parameter method named CLIPTTA (Lafon et al., 2025). We also tested the modular adapter-routing strategies Polytropon (Ponti et al., 2022) and MHR (Page-Caccia et al., 2023), as they have been shown to improve multi-task optimization over earlier approaches.

Tab. 1 reports the results on ImageNet (experimental setup in Sec. 5). While PCGrad yields a minor improvement relative to CLIPTTA, both the remaining gradient rebalancing methods (CAGrad, FairGrad) and the modular adapter-routing strategies (Polytropon, MHR) improve ID/OOD separation at the expense of classification accuracy. Their gains are therefore one-sided, shifting performance toward the separation objective rather than improving both objectives jointly. Ultimately, neither family of methods resolves the underlying conflict and the classification-separation trade-off.

Table 1: Comparison of alternative strategies to mitigate objective interference on ImageNet.

Method	ACC↑	AUC↑	FPR95↓
CLIPTTA	67.6	97.7	9.7
Gradient Conflict Mitigation			
CLIPTTA w/ PCGrad	67.6	97.8	8.9
CLIPTTA w/ CAGrad	66.7	98.4	6.6
CLIPTTA w/ FairGrad	64.0	99.0	4.2
Modular PEFT			
CLIPTTA w/ Polytropon	58.6	99.4	1.7
CLIPTTA w/ MHR	59.5	99.3	1.8
DSTA (ours)	**69.2**	**99.8**	**0.8**

4 The Proposed DSTA Method

Let θ denote the frozen parameters of a pretrained CLIP vision encoder f_v, and C the text prototypes of the known classes produced by the frozen CLIP text encoder f_t. DSTA augments the vision encoder with two independent sets of low-rank adapters (Hu et al., 2022) specialized for classification and ID/OOD separation, denoted by ϕ_{cls} and ϕ_{sep}, respectively. At each adaptation step, only the adapter set associated with the corresponding objective is active and updated, while the other set is temporarily disabled. We describe DSTA in detail below, while Fig. 2 provides an overview.

Separation Stream. Given a test batch x, the separation adapters ϕ_{sep} are activated and the representations $z_{\text{sep}} = f_v(x; \theta, \phi_{\text{sep}})$ are computed first. Inspired by maximum-softmax confidence scoring for OOD detection and test-time sample selection (Hendrycks & Gimpel, 2017; Ming et al., 2022; Gong et al., 2023), we compute the ID-confidence score as the largest log-softmax value over the known classes. Specifically, let $u_{ic} = a\, z_{\text{sep},i}^{\top} C_c$ denote the logit of sample i for known class c, where $z_{\text{sep},i}$ is its separation-stream visual representation, C_c is the corresponding class text prototype, and $z_{\text{sep},i}^{\top} C_c$ measures their image–text

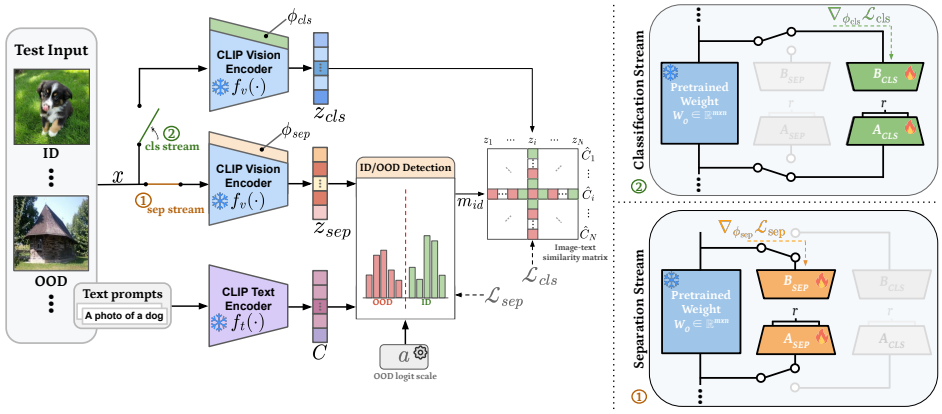

Figure 2: **Overview of DSTA.** Test batches go through two streams over the same frozen backbone. **Separation stream's** adapters ϕ_{sep} are activated in the vision encoder f_v, producing representations z_{sep}, while text encoder f_t maps the known classes to text prototypes C. This stream uses image-text similarities between z_{sep} and C to perform ID/OOD detection, where calibrated OOD logit scale a shapes score distribution used for separation. The resulting mask m_{id} selects reliable ID samples for classification, while optimizing the margin-based objective $\mathcal{L}_{\text{sep}}$. **Classification stream.** ϕ_{sep} are disabled and classification adapters ϕ_{cls} are activated, producing z_{cls}. For each reliable ID sample $z_{\text{cls}}[m_{\text{id}}]$, the most similar class is used as pseudo-label $\hat{C}_i$ to form the image-text similarity matrix optimized by $\mathcal{L}_{\text{cls}}$. **Right: objective-specific adaptation.** For each weight W_0, DSTA has adapters $(A_{\text{sep}}, B_{\text{sep}})$ and $(A_{\text{cls}}, B_{\text{cls}})$. Only the active stream is updated via $\mathcal{L}_{\text{sep}}$ or $\mathcal{L}_{\text{cls}}$, while W_0 is frozen.

similarity. The OOD logit scale a controls the spread of the class probabilities used for ID/OOD scoring, and we will show later how to automatically calibrate it.

Hence, we define the ID-confidence score for sample i and the corresponding soft ID weight as:

$$s_i = \max_c \log \frac{\exp(u_{ic})}{\sum_{k=1}^{C} \exp(u_{ik})}, \qquad w_i = \sigma\left(\frac{s_i - \alpha}{\gamma}\right). \tag{1}$$

A larger s_i indicates higher compatibility with at least one known-class prototype and thus an ID-like sample, as ID inputs tend to match at least one known class with a higher score (Ming et al., 2022). The weight w_i converts this confidence into a soft ID assignment, where α denotes the ID/OOD threshold and γ controls the softness of the partition. At the beginning of each incoming test batch, we set α by applying Otsu thresholding (Otsu, 1979) to the batch confidence scores $\{s_i\}$. Samples with scores above α receive weights closer to one, while those below it receive weights closer to zero, and a smaller γ makes this transition sharper. Samples with large w_i are therefore treated as reliable ID samples by converting the soft assignments into a binary mask $m_{\text{id}} = (\mathbf{1}[w_i > 0.5])_i$, which defines the subset of reliable samples used for classification adaptation.

Margin-based Separation Objective. To prevent the separation stream from increasing the ID/OOD gap during optimization, we construct a bounded objective $\mathcal{L}_{\text{sep}}$ from the soft ID weights w_i and scores s_i of Eq. 1. We define the weighted mean scores of the ID and OOD groups as:

$$\mu_{\text{id}} = \frac{\sum_i w_i s_i}{\sum_i w_i}, \qquad \mu_{\text{ood}} = \frac{\sum_i (1 - w_i) s_i}{\sum_i (1 - w_i)}, \qquad \Delta = \mu_{\text{id}} - \mu_{\text{ood}}, \tag{2}$$

where Δ measures the separation between the two groups, with larger values indicating stronger ID/OOD separation.

To maximize the ID/OOD gap, prior work (Lafon et al., 2025) employs the outlier contrastive exposure objective whose gradient magnitude increases linearly with Δ. Consequently, it pushes for larger separation even after the ID and OOD groups are separated. When using decoupled adapters, this can lead to a problematic behavior as the separation adapters are optimized only for separation, without a competing objective to counteract the continued growth of this gap. As a result, Δ can be driven to extreme values, destabilizing adaptation. To prevent this, we introduce a margin-based smooth L1 objective that limits further separation once a sufficient gap is achieved. Let m denote the

target separation margin and β define the width of the quadratic transition region, so the separation loss is defined as:

$$\mathcal{L}_{\text{sep}}(\Delta) = \begin{cases} 0, & \Delta \geq m, \\ (m-\Delta)^2/(2\beta), & m-\beta < \Delta < m, \\ (m-\Delta) - (\beta/2), & \Delta \leq m-\beta. \end{cases} \tag{3}$$

The loss has three regimes: when separation is insufficient, it applies a constant-magnitude optimization signal to increase the gap; near the target, this signal is smoothly reduced; once the target margin is reached, the separation stream receives no further optimization signal. Unlike the squared objective, its gradient magnitude is bounded by one and vanishes once the target separation is reached. The full derivative is provided in Supp. Mat. A.1.

Automatic OOD Logit-Scale Calibration. The ID-confidence scores of the separation stream (Eq. 1) depend on the logit scale a, that controls the sharpness of the class-probability distribution, and hence the dispersion of confidence scores across samples. A poor initialization of a can compress these scores into a narrow range, weakening ID/OOD discrimination. We therefore calibrate the initial a from the first test batch.

Let $q_{ic} = z_{\text{sep},i}^{\top} C_c$ denote similarity between image i and known-class prototype c, and $\mathbf{q}_i = (q_{i1}, \ldots, q_{iC})$ be the similarity vector with respect to all known classes. For a candidate scale a, the corresponding ID-confidence score is $s_i(a) = \max_c \log[\text{softmax}(a\mathbf{q}_i)]_c$ (Eq. 1), yielding the batch score distribution $\mathcal{S}(a) = \{s_i(a)\}_i$. Thus, a controls how strongly similarity differences are amplified and, consequently, the dispersion of $\mathcal{S}(a)$.

For each candidate a in a predefined grid $\mathcal{G}$, we measure the standard deviation of $\mathcal{S}(a)$ and select the scale whose dispersion is closest to a fixed target σ_{target}: $a^\star = \arg\min_{a\in\mathcal{G}} |\text{Std}(\mathcal{S}(a)) - \sigma_{\text{target}}|$. Importantly, this is a label-free procedure that uses only the confidence scores of the incoming batch. The resulting $a^\star$ initializes the trainable OOD logit scale, which is subsequently refined with the separation stream. Additional details are provided in Supp. Mat. A.5.

Classification Stream. After computing the separation objective, DSTA disables the separation adapters ϕ_{sep}, activates the classification adapters ϕ_{cls}, and computes $z_{\text{cls}} = f_v(x; \theta, \phi_{\text{cls}})$. Classification adaptation is performed only using samples identified as reliable ID samples by the separation stream, which we denote as $z_{\text{cls}}^{\text{id}} = z_{\text{cls}}[m_{\text{id}}]$. Inspired by memory-based TTA (Yuan et al., 2023; Yu et al., 2024; Lafon et al., 2025), we adopt a batch-sized Class-wise Confident Memory (CCM) to store class-balanced high-confidence samples from previous batches, that are used as auxiliary reliable samples. These samples are then added to the set of current reliable ID samples. We optimize the classification adapters using the image-text contrastive objective $\mathcal{L}_{\text{cls}}$ with marginal entropy regularization (Lafon et al., 2025), which has been shown to work well in practice. DSTA applies this objective exclusively to the reliable ID samples selected by the separation stream, allowing the classification update to operate on a cleaner adaptation set while updating ϕ_{cls}.

Decoupled Optimization and Gradient Isolation. The separation and classification streams are parameter-isolated, but functionally connected: information flows from separation to classification through sample selection, while their trainable parameters remain independent. The two objectives are optimized independently over their disjoint parameter sets: $\min_{\phi_{\text{cls}}} \mathcal{L}_{\text{cls}}$ and $\min_{\phi_{\text{sep}}} \mathcal{L}_{\text{sep}}$. The backbone parameters θ and text prototypes C remain frozen. The two optimizations interact only through the ID mask m_{id} produced by the separation stream, which determines the samples used by $\mathcal{L}_{\text{cls}}$. Thus, no weighting coefficient is required to combine $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{sep}}$. The full algorithm and its computational overhead are reported in Supp. Mat. A.3 and Supp. Mat. A.11, respectively.

This parameter decoupling also yields explicit gradient isolation. Let $\phi = [\phi_{\text{cls}}, \phi_{\text{sep}}]$ denote the complete trainable adapter parameters. Since $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{sep}}$ update disjoint parameter sets, their gradients have disjoint support, and therefore $\langle \nabla_\phi \mathcal{L}_{\text{cls}}, \nabla_\phi \mathcal{L}_{\text{sep}} \rangle = 0$. Intuitively, ϕ contains both adapter sets, but $\mathcal{L}_{\text{cls}}$ only acts on ϕ_{cls} and $\mathcal{L}_{\text{sep}}$ only on ϕ_{sep}. The two gradients are then non-zero in different parts of the parameter vector, yielding nil inner product. By construction, classification and ID/OOD separation cannot directly interfere through the trainable adapter parameters in DSTA. This contrasts with gradient-conflict mitigation methods that modify gradients after conflicts arise on shared parameters. The explicit gradient decomposition is provided in Supp. Mat. A.2.

Table 2: **Open-set TTA results with CLIP ViT-B/16.** Panel (a) reports clean results and Panel (b) the corresponding corrupted settings. Best results are in bold and second-best results are underlined.

	CIFAR-10				CIFAR-100				ImageNet				HAC				CCT				Average			
	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑
(a) Clean																								
CLIP	89.3	98.5	5.2	94.0	68.1	86.8	83.5	34.6	66.7	90.1	43.8	68.4	78.1	70.6	90.9	20.6	37.0	63.6	83.2	29.3	67.8	81.9	61.3	49.4
OSTTA	90.9	60.5	72.8	46.7	70.9	43.3	93.8	15.1	66.9	84.9	59.2	58.6	72.8	37.8	96.7	8.5	34.7	53.6	93.2	15.4	67.2	56.0	83.1	28.9
SoTTA	89.5	98.5	4.9	94.2	68.9	88.5	76.3	44.1	66.7	89.3	47.1	66.5	72.3	74.3	80.9	35.8	**39.3**	60.5	87.4	24.7	67.3	82.2	59.3	53.1
STAMP	89.9	98.6	5.5	94.2	67.5	87.7	80.0	39.4	29.7	63.0	80.2	30.0	75.5	68.9	93.0	16.8	35.5	64.8	81.8	30.4	59.6	76.6	68.1	42.2
UniEnt	94.2	99.9	**0.0**	98.0	72.7	97.8	8.7	85.9	65.2	95.4	17.1	79.2	79.0	**82.3**	**65.1**	**53.3**	37.4	65.4	85.7	26.8	69.7	88.2	35.3	68.6
CLIPTTA	94.6	99.8	0.4	97.9	74.9	98.4	7.6	87.4	67.6	97.7	9.7	83.1	79.5	72.8	86.8	28.2	37.4	63.4	83.5	29.1	70.8	86.4	37.6	65.1
DSTA (ours)	**95.0**	**100.0**	**0.0**	**98.3**	**75.5**	**100.0**	**0.1**	**90.2**	**69.2**	**99.8**	**0.8**	**86.8**	**79.7**	73.8	67.7	50.3	37.1	**68.6**	**78.0**	**34.5**	**71.3**	**88.4**	**29.3**	**72.0**
(b) Corrupted																								
CLIP	60.2	88.0	58.0	57.9	35.2	67.0	93.8	14.7	24.6	68.6	89.2	20.3	64.8	73.5	82.5	34.2	20.3	54.9	93.7	13.3	41.0	70.4	83.4	28.1
OSTTA	62.7	53.4	85.2	29.3	34.5	36.3	92.6	15.7	31.2	75.1	79.8	31.6	51.1	33.9	95.1	11.8	17.0	46.8	95.8	9.4	39.3	49.1	89.7	19.6
SoTTA	38.6	52.0	80.6	31.0	22.5	49.1	90.4	17.8	4.0	34.6	93.1	7.1	56.6	64.3	83.9	31.3	**22.3**	53.1	94.3	12.5	28.8	50.6	88.5	19.9
STAMP	60.4	88.0	57.1	58.6	34.5	66.8	93.8	14.6	9.0	53.8	92.9	11.1	62.1	71.5	84.4	31.1	19.3	54.1	94.0	12.7	37.1	66.8	84.4	25.6
UniEnt	78.7	98.6	5.7	89.7	48.9	91.0	31.0	65.3	23.6	44.8	90.8	17.3	**69.8**	**85.6**	54.2	62.5	20.5	**61.0**	**88.3**	**19.9**	48.3	76.2	54.0	50.9
CLIPTTA	79.1	**98.7**	6.1	89.8	50.4	**96.7**	19.2	**70.5**	**39.0**	**89.0**	**43.2**	**55.1**	67.4	76.6	78.4	40.1	20.9	55.5	93.4	13.8	**51.4**	**83.3**	48.1	53.9
DSTA (ours)	**80.5**	**98.7**	**3.1**	**91.3**	**51.5**	85.4	**18.8**	69.0	36.3	82.9	55.0	48.5	67.5	85.0	**49.4**	**64.5**	20.8	58.7	88.8	19.4	51.3	82.1	**43.0**	**58.5**

5 EXPERIMENTAL ANALYSIS

Datasets. Following the standard open-set evaluation protocol (Lafon et al., 2025), for clean semantic shifts, we use CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) as ID datasets and SVHN (Netzer et al., 2011) as OOD, and ImageNet (Deng et al., 2009) as ID with Places365 (Zhou et al., 2016) as OOD. For covariate shifts (Hendrycks & Dietterich, 2019), we use CIFAR-10-C and CIFAR-100-C with SVHN-C as OOD and on ImageNet-C with Places365-C as OOD. For each benchmark, the ID and OOD samples undergo matched corruption types following the same corruption protocol. We broaden this standard evaluation with the Human-Animal-Cartoon (HAC) (Zhao et al., 2026) and Caltech Camera Traps (CCT) (Beery et al., 2018), which cover challenging heterogeneous domains and location-based distribution shifts. Humans (H), Animals (A), and Cartoons (C) of HAC serve as ID datasets with EPIC-KITCHENS (EK) (Damen et al., 2021) as OOD. For CCT, training images constitute the ID set, while images from the trans-validation camera are OOD. Details of the HAC and CCT dataset preparation and corruption procedures are in Supp. Mat. A.10.

Evaluation Setup. We report ID top-1 accuracy (Acc), AUROC (AUC) for ID/OOD discrimination, and FPR95. Classification metrics are calculated on ID samples, while OOD samples are used solely for unknown detection. We report the H-score (Zhao et al., 2026), computed as $\mathrm{H} = 3/(1/\mathrm{Acc} + 1/\mathrm{AUC} + 1/(1 - \mathrm{FPR95}))$, to summarize classification and OOD detection jointly. For HAC, metrics are averaged across its three ID/OOD settings. For corrupted datasets, Acc, AUC, and FPR95 are averaged across corruption types before computing the H-score. Final results are arithmetic means across all settings. DSTA is compared with prior open-set TTA methods, including OSTTA (Lee et al., 2023), SoTTA (Gong et al., 2023), STAMP (Yu et al., 2024), UniEnt (Gao et al., 2024), and CLIPTTA (Lafon et al., 2025), along with zero-shot CLIP (Radford et al., 2021). All methods use the same ID/OOD composition and evaluation protocol.

Implementation Details. We use OpenAI CLIP ViT-B/16 and OpenVision ViT-S/16 (Li et al., 2025) as backbones. Following previous work (Yu et al., 2024; Gao et al., 2024; Lafon et al., 2025), we use batches comprising 128 ID and 128 OOD samples, and adapt each batch for 10 steps using rank-8 LoRA with scaling 16. Classification adapters are placed in the attention projections (Q, K, V, O) and final projection (PROJ), while separation ones are placed in the MLP (FFN1, FFN2) and final projection (PROJ). The soft ID/OOD partition uses $\gamma = 0.005$. OOD calibration uses $\sigma_{\text{target}} = 0.5$ to estimate the OOD logit scale a from incoming test data, yielding dataset-dependent values without fixed dataset-specific tuning. The margin objective uses $m = 3$ with the Smooth L1 choice $\beta = 1$ (Girshick, 2015). More implementation details are reported in Supp. Mat. A.9.

5.1 RESULTS

Does DSTA jointly improve classification and OOD detection? Tab. 2 reports TTA results using CLIP ViT-B/16 backbone. On clean data (panel (a)), DSTA achieves the best aggregate performance across all metrics. Compared with the strongest competing result for each metric, this corresponds to a $+0.5$ point improvement in ACC over CLIPTTA and, relative to UniEnt, a $+0.2$ point improvement

Table 3: **Progressive ablation of DSTA components on ImageNet.** We first isolate the effect of the proposed margin loss under CLIPTTA's original shared-parameter adaptation, and then progressively introduce parameter decoupling and the remaining DSTA components.

	Components		Tuning		ImageNet		
Configuration	**Margin**	**Logit Calib.**	**CLS**	**SEP**	ACC↑	AUC↑	FPR95↓
Shared-parameter baselines							
CLIPTTA			LN (shared)	LN (shared)	67.6	97.7	9.7
CLIPTTA + Margin	✓		LN (shared)	LN (shared)	67.6	97.5	10.3
Decoupled adaptation							
+ LN Decoupling			LN	LN	67.8	55.7	94.2
+ Margin Loss	✓		LN	LN	67.9	99.5	1.4
+ Logit Calibration	✓	✓	LN	LN	68.0	99.7	1.0
+ LoRA	✓	✓	ATTN	MLP	69.0	99.8	0.9
DSTA (ours)	✓	✓	**ATTN+PROJ**	**MLP+PROJ**	**69.2**	**99.8**	**0.8**

in AUC, a 6.0 point reduction in FPR95, and a $+3.4$ point improvement in H-score. The advantage is particularly clear on the standard semantic open-set benchmarks (CIFAR-10, CIFAR-100, and ImageNet), where DSTA improves both ID classification and OOD detection. Relative to CLIPTTA, DSTA improves ACC on all datasets while also increasing AUC and reducing FPR95. The gains are particularly pronounced on CIFAR-100 where DSTA improves H-score from 87.4% to 90.2% while reducing FPR95 from 7.6% to 0.1%, and on ImageNet where it improves ACC from 67.6% to 69.2%, AUC from 97.7% to 99.8%, and reduces FPR95 from 9.7% to 0.8%, yielding a $+3.7$ point improvement in H-score. For HAC and CCT, the relative strengths of DSTA become metric-dependent, revealing a more pronounced classification-separation trade-off. Nevertheless, when performance is aggregated across all datasets, DSTA leads all evaluation metrics. These results show that isolating classification and ID/OOD separation updates improves both objectives, rather than trading performance from one objective to the other.

Does DSTA remain effective under corruptions? As shown in panel (b) of Tab. 2, the classification-separation trade-off becomes more pronounced between the methods. DSTA achieves the best aggregate FPR95 and H-score, reducing FPR95 by 5.1 points and improving H-score by 4.6 points over CLIPTTA, with only a negligible 0.1-point decrease in ACC. This pattern is especially evident on HAC, where, relative to CLIPTTA, DSTA improves AUC by 8.4 points and reduces FPR95 by 29.0 points, resulting in a $+24.4$ point gain in H-score while keeping the ACC essentially unchanged ($+0.1$ points). On CCT, the gains are metric-dependent with DSTA improving the H-score by 5.6 points. Overall, these results show that the advantage of decoupling is preserved under covariate shift. DSTA improves the joint classification-OOD detection objective through stronger unknown-sample rejection, without degrading aggregate ID classification.

Does DSTA remain effective with a smaller backbone? We evaluate DSTA with OpenVision ViT-S/16 to assess whether the benefits of it generalize beyond the ViT-B/16 (full analysis in Supp. Mat. A.7). On clean data, DSTA remains close to the strongest competing method, UniEnt, with only small differences in aggregate metrics. More importantly, DSTA achieves the best aggregate FPR95 and H-score under corruption, reducing FPR95 by 6.0 points and improving H-score by 2.3 points. Compared to CLIPTTA, the corresponding gains are larger, with a 13.2-point reduction in FPR95 and a $+6.5$-point improvement in H-score. These results show that the advantages of DSTA are not specific to the ViT-B/16 and remain evident with a smaller vision encoder.

Ablation study. Tab. 3 analyzes how each component contributes to DSTA on ImageNet. We begin from CLIPTTA, where the classification and separation objectives are optimized jointly through the same LayerNorm (LN) parameters. To isolate the effect of the proposed separation objective, we first replace OCE with the margin loss while keeping this shared-parameter formulation unchanged. This leaves ACC unchanged and slightly degrades OOD detection, showing that the gains of DSTA cannot be attributed to the margin objective alone. We next introduce parameter decoupling. In the *+LN Decoupling* configuration, classification and separation are assigned independent LN parameter sets, so that the two objectives no longer update the same parameters. This causes OOD detection to collapse, with AUC decreasing by 42.0 points and FPR95 increasing by 84.5 points relative to CLIPTTA. This indicates that parameter isolation alone is not sufficient: once the separation stream is optimized independently, the original unbounded OCE objective severely degrades separation. The *+Margin Loss* configuration keeps the same LN decoupling and replaces OCE with the proposed bounded margin objective. This restores OOD separation, increasing AUC by 43.8 points and reducing FPR95 by 92.8 points relative to the decoupled OCE configuration. Automatic logit-scale

Table 4: **Capacity control and adapter allocation on ImageNet.** CLIPTTA++ augments CLIPTTA with additional LoRA capacity, while DSTA variants compare alternative adapter placements.

	Capacity		Tuning		ImageNet		
Configuration	**Rank**	**T. Params**	**CLS**	**SEP**	ACC↑	AUC↑	FPR95↓
CLIPTTA parameter variants							
CLIPTTA	–	0.4M	LN	LN	67.6	97.7	9.7
CLIPTTA++	8	1.4M	ATTN+MLP	ATTN+MLP	68.8	97.1	13.0
CLIPTTA++	8	0.6M	ATTN	ATTN	68.4	94.1	24.9
CLIPTTA++	8	0.8M	MLP	MLP	68.4	96.5	15.3
CLIPTTA++	16	1.2M	ATTN	ATTN	68.8	95.1	20.3
CLIPTTA++	16	1.5M	MLP	MLP	68.6	97.5	11.0
DSTA adapter allocation							
DSTA	8	1.2M	ATTN	ATTN	68.5	95.7	15.6
DSTA	8	1.5M	MLP	MLP	68.3	**99.8**	**0.9**
DSTA	8	1.3M	MLP	ATTN	66.5	95.7	15.6
DSTA (ours)	8	1.3M	**ATTN**	**MLP**	**69.0**	**99.8**	**0.9**

calibration provides a further improvement, increasing AUC by 0.2 points and reducing FPR95 by 0.4 points. We then replace the LN parameters with objective-specific LoRA adapters. In the *+LoRA* configuration, classification adapters are placed in self-attention, whereas separation adapters are in the MLP. This produces the largest classification gains, increasing ACC by 1.0 point while maintaining strong OOD detection. Self-attention directly modulates token interactions and the aggregation of patch information into the classification representation, which may help explain its effectiveness for classification adaptation (Dosovitskiy et al., 2021; Touvron et al., 2022). In contrast, the token-wise MLP transforms features independently at each token, providing a complementary adaptation space for the separation stream (Vaswani et al., 2017). Finally, the full DSTA configuration adds separate output-projection adapters to both streams. This yields a further $+0.2$ point improvement in ACC and reduces FPR95 by 0.1 points, resulting in the best overall configuration.

Capacity control and adapter allocation. We examine whether DSTA's gains can be explained by a larger trainable parameter budget, and whether performance depends on how the objective-specific adapter sets are allocated. We augment CLIPTTA (named CLIPTTA++) with additional LoRA capacity comparable with DSTA. Tab. 4 compares these parameter-budget controls with alternative DSTA adapter allocations on ImageNet. Increasing the trainable capacity of CLIPTTA does not recover DSTA's joint classification and OOD-detection performance. Even when CLIPTTA++ uses 1.4-1.5M trainable parameters, i.e., a budget comparable to or larger than DSTA's 1.3M, its best ACC remains 0.2 points below DSTA, while its OOD detection is substantially weaker. In particular, the CLIPTTA++ variant with the best FPR95 reaches 11.0%, compared with 0.9% for DSTA. This indicates that DSTA's gains cannot be attributed to parameter count alone, and instead depend on the structure and allocation of the adaptation parameters across the two objectives. Adapter allocation is also critical. Assigning MLP adapters to both streams preserves the strong OOD separation of DSTA, matching its AUC and FPR95, but reduces ACC by 0.7 points. Conversely, assigning attention adapters to both streams substantially degrades OOD detection. Notably, swapping the proposed allocation, with MLP adapters for classification and attention adapters for separation, reduces ACC by 2.5 points and AUC by 4.1 points while increasing FPR95 by 14.7 points. These results support the asymmetric allocation used by DSTA, with classification adaptation in self-attention and separation adaptation in the MLP. Together with the CLIPTTA++ controls, they further show that this advantage cannot be reproduced simply by increasing the trainable parameter budget.

6 Conclusions

We introduced DSTA, a decoupled methodology for open-set test-time adaptation of vision-language models. Motivated by the observation that classification and ID/OOD separation become progressively less aligned when optimized through shared parameters, DSTA assigns the two objectives to disjoint adapter sets while preserving their interaction through OOD-guided sample selection. Experiments support this design: DSTA achieves the strongest aggregate performance on clean benchmarks and improves FPR95 and H-score under corruption. Ablations further show that the gains cannot be explained by the separation objective or parameter count alone, but arise from the combination of decoupling, bounded separation, calibration, and objective-specific adapter allocation. These results support parameter-space decoupling as an effective design principle for open-set test-time adaptation. Future work could build on this principle by developing more efficient decoupling strategies and learning when information should be selectively shared across streams.

AI USE STATEMENT

In this work, we used generative AI tools for editing portions of the manuscript to improve grammar, clarity, and readability. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide ingredients for mathematical proofs, assist in writing proofs, propose or refine hypotheses, design or provide feedback on the methodology or experiments, implement methods, translate content, clean or reformat datasets, conduct qualitative or thematic analysis, or interpret results. Any remaining required-disclosure categories were not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve its readability, which is a recommended disclosure category. All AI-assisted text was reviewed and edited by the authors. The authors independently verified the analyses, results, and claims, and take responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing publicly available benchmark datasets and does not involve the collection of new human-subject data. Our study focuses on methodological advances in open-set test-time adaptation. As with other vision-language models, performance may reflect biases present in the pretrained models and evaluation datasets, and results should therefore be interpreted within the scope of the benchmarks considered.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our experiments throughout the main paper and supplementary material. The main text describes the DSTA architecture, optimization objectives, evaluation protocol, and principal implementation settings. The supplementary material provides the complete adaptation algorithm (Sec. A.3), additional details on OOD logit-scale calibration (Sec. A.5), implementation and baseline hyperparameters (Sec. A.9), and dataset preparation and corruption protocols (Sec. A.10). Computational details are reported in Sec. A.11, while Sec. A.12 evaluates reproducibility across three matched random seeds. The source code and experimental configurations will be released upon acceptance.

REFERENCES

Hao Ban and Kaiyi Ji. Fair resource allocation in multi-task learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2715–2731. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/ban24a.html.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pp. 472–489. Springer, 2018. doi: 10.1007/978-3-030-01270-0_28. URL https://doi.org/10.1007/978-3-030-01270-0_28.

Chentao Cao, Zhun Zhong, Zhanke Zhou, Tongliang Liu, Yang Liu, Kun Zhang, and Bo Han. Noisy test-time adaptation in vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=iylpeTI0Ql.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The EPIC-KITCHENS dataset: Collection, challenges and baselines. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):4125–4141, 2021. doi: 10.1109/TPAMI.2020.2991965. URL https://doi.org/10.1109/TPAMI.2020.2991965.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Hao Dong, Eleni N. Chatzi, and Olga Fink. Towards robust multimodal open-set test-time adaptation via adaptive entropy-aware optimization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=hj323oR3rw.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 23975–23984. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02263. URL https://doi.org/10.1109/CVPR52733.2024.02263.

Ross B. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1440–1448. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.169. URL https://doi.org/10.1109/ICCV.2015.169.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/2da53cd1abdae59150e35f4693834f32-Abstract-Conference.html.

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 2025, Tucson, AZ, USA, February 26 - March 6, 2025*, pp. 7092–7101. IEEE, 2025. doi: 10.1109/WACV61041.2025.00689. URL https://doi.org/10.1109/WACV61041.2025.00689.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric P. Xing. Efficient test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14162–14171. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01343. URL https://doi.org/10.1109/CVPR52733.2024.01343.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Marc Lafon, Gustavo Adolfo Vargas Hakim, Clément Rambour, Christian Desrosiers, and Nicolas Thome. CLIPTTA: robust contrastive vision-language test-time adaptation. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/67e2da8af2476c7441c881fc253042a6-Abstract-Conference.html.

Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 16334. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01501. URL https://doi.org/10.1109/ICCV51070.2023.01501.

Xianhang Li, Yanqing Liu, Haoqin Tu, and Cihang Xie. Openvision: A fully-open, cost-effective family of advanced vision encoders for multimodal learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 3977–3987. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00379. URL https://doi.org/10.1109/ICCV51701.2025.00379.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 18878–18890, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 565–576. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.47. URL https://doi.org/10.18653/v1/2021.acl-long.47.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/e43a33994a28f746dcfd53eb51ed3c2d-Abstract-Conference.html.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

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 *The Eleventh International*

Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076. URL https://doi.org/10.1109/tsmc.1979.4310076.

Lucas Page-Caccia, Edoardo Maria Ponti, Zhan Su, Matheus Pereira, Nicolas Le Roux, and Alessandro Sordoni. Multi-head adapter routing for cross-task generalization. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/b295b3a940706f431076c86b78907757-Abstract-Conference.html.

Boyang Peng, Sanqing Qu, Tianpei Zou, Fan Lu, Ya Wu, Kai Chen, Siheng Chen, Yong Wu, and Guang Chen. Ood-barrier: Build a middle-barrier for open-set single-image test time adaptation via vision language models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/dd391150be8cec625434323f6b1f9d14-Abstract-Conference.html.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 487–503. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EACL-MAIN.39. URL https://doi.org/10.18653/v1/2021.eacl-main.39.

Edoardo Maria Ponti, Alessandro Sordoni, and Siva Reddy. Combining modular skills in multitask learning. *CoRR*, abs/2202.13914, 2022. URL https://arxiv.org/abs/2202.13914.

Xiaohan Qin, Xiaoxing Wang, and Junchi Yan. Towards consistent multi-task learning: Unlocking the potential of task-specific parameters. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 10067–10076. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.00941. URL https://openaccess.thecvf.com/content/CVPR2025/html/Qin_Towards_Consistent_Multi-Task_Learning_Unlocking_the_Potential_of_Task-Specific_Parameters_CVPR_2025_paper.html.

Xiaohan Qin, Xiaoxing Wang, and Junchi Yan. Revisiting fairness in multitask learning: A performance-driven approach for variance reduction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 20492–20501. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.01908. URL https://openaccess.thecvf.com/content/CVPR2025/html/Qin_Revisiting_Fairness_in_Multitask_Learning_A_Performance-Driven_Approach_for_Variance_CVPR_2025_paper.html.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The*

Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pp. 4822–4829. AAAI Press, 2019. doi: 10.1609/AAAI.V33I01.33014822. URL https://doi.org/10.1609/aaai.v33i01.33014822.

Manogna Sreenivas and Soma Biswas. Efficient open set single image test time adaptation of vision language models. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=72YVabBErN.

Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV*, volume 13684 of *Lecture Notes in Computer Science*, pp. 497–515. Springer, 2022. doi: 10.1007/978-3-031-20053-3_29. URL https://doi.org/10.1007/978-3-031-20053-3_29.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Yongcan Yu, Lijun Sheng, Ran He, and Jian Liang. STAMP: outlier-aware test-time adaptation with stable memory replay. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, volume 15139 of *Lecture Notes in Computer Science*, pp. 375–392. Springer, 2024. doi: 10.1007/978-3-031-73004-7_22. URL https://doi.org/10.1007/978-3-031-73004-7_22.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15922–15932. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01528. URL https://doi.org/10.1109/CVPR52729.2023.01528.

Wenjie Zhao, Jia Li, Xin Dong, Yapeng Tian, Yu Xiang, and Yunhui Guo. Mitigating the id-ood tradeoff in open-set test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6789–6798, June 2026.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2016. URL http://arxiv.org/abs/1610.02055.

A SUPPLEMENTARY MATERIAL

This supplementary material provides additional theoretical analysis, methodological details, experiments, and reproducibility information supporting the main paper. It includes further analysis of the proposed margin-based separation objective (Sec. A.1), the explicit gradient-isolation derivation (Sec. A.2), the complete DSTA adaptation procedure (Sec. A.3), gradient-interaction analysis across additional datasets (Sec. A.4), and further details on automatic OOD logit-scale calibration (Sec. A.5). We also provide sensitivity analysis of the calibration target (Sec. A.6), evaluation with a smaller backbone (Sec. A.7), and a long-horizon analysis of the separation stream (Sec. A.8). Additional implementation details and baseline hyperparameters (Sec. A.9), HAC and CCT evaluation protocols (Sec. A.10), computational-cost analysis (Sec. A.11), multiple-run statistics with significance testing (Sec. A.12), and expanded HAC results (Sec. A.13) are also reported.

A.1 MARGIN OBJECTIVE ANALYSIS

We first contrast the proposed margin objective with the squared separation objective used by CLIPTTA (Lafon et al., 2025). Given the separation gap Δ, the Outlier Contrastive Exposure (OCE) objective is

$$\mathcal{L}_{\text{OCE}} = -\Delta^2, \qquad \left|\frac{\partial \mathcal{L}_{\text{OCE}}}{\partial \Delta}\right| = 2|\Delta|. \tag{4}$$

Its gradient magnitude therefore grows linearly with $|\Delta|$, so the objective continues to favor larger separation even when the ID and OOD groups are already well separated.

In contrast, for the margin objective in Eq. 3, the derivative with respect to the separation gap is

$$\frac{\partial \mathcal{L}_{\text{sep}}}{\partial \Delta} = \begin{cases} 0, & \Delta \geq m, \\ -\dfrac{m-\Delta}{\beta}, & m-\beta < \Delta < m, \\ -1, & \Delta \leq m-\beta. \end{cases} \tag{5}$$

Hence,

$$\left|\frac{\partial \mathcal{L}_{\text{sep}}}{\partial \Delta}\right| \leq 1. \tag{6}$$

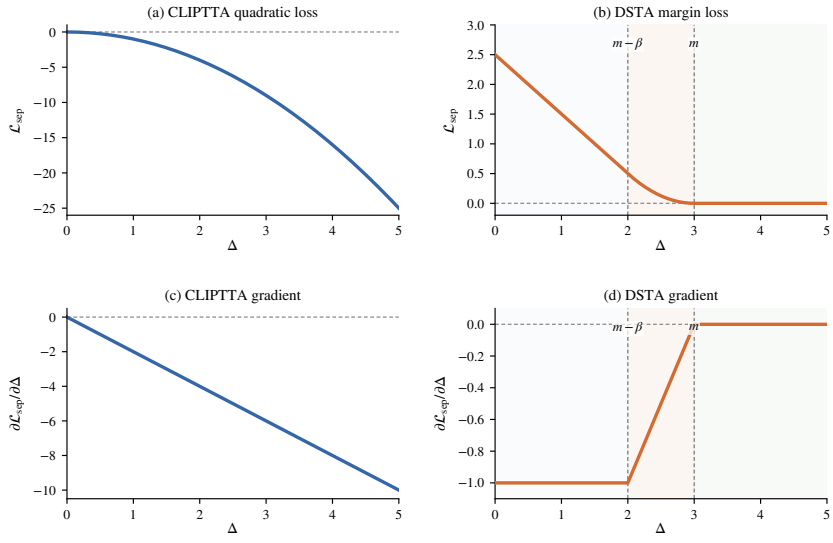

Figure 3: **Comparison of the OCE and margin-based separation objectives.**

Figure 3 illustrates the resulting optimization behavior. When separation is well below the target, the margin objective applies constant pressure to increase the gap. Within the transition region, this pressure decreases smoothly as Δ approaches m, and vanishes once $\Delta \geq m$. In contrast, the gradient magnitude of the quadratic OCE objective continues to grow as $2|\Delta|$.

Having established this bounded-gradient behavior, we now discuss the margin parameters used in our experiments. We set $m = 3$ based on the separation levels observed with the original OCE objective before unstable behavior emerged (Fig. 4), and use the same value across all datasets without dataset-specific tuning. Since s_i is a log-probability score, $\Delta = 3$ corresponds to an approximately $e^3 \simeq 20$-fold ratio between the weighted geometric mean maximum-class probabilities of the soft ID and OOD partitions. We set $\beta = 1$, consistent with the original Smooth L1 formulation, yielding a unit-width transition between the constant-gradient and zero-gradient regimes.

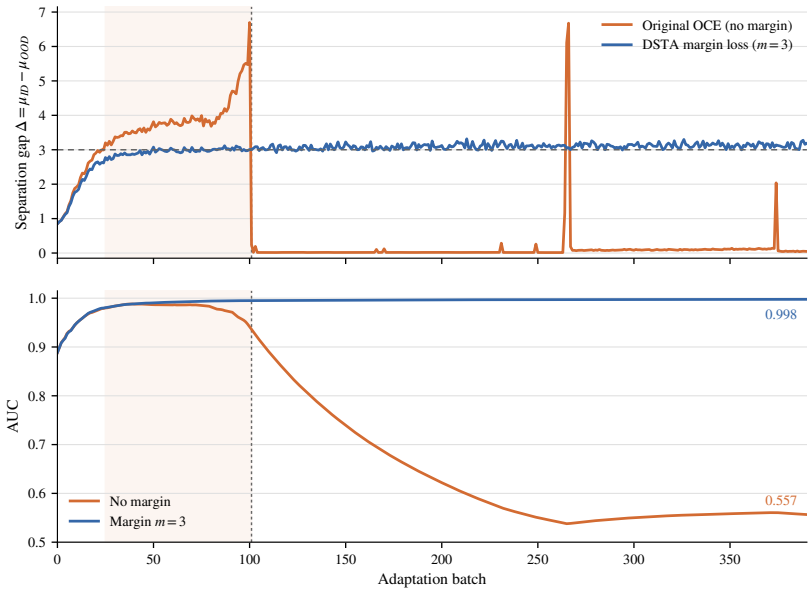

Figure 4: **OCE vs. Margin-based Separation Dynamics. Top:** evolution of Δ under the original OCE objective and the proposed margin objective. Before OCE becomes unstable, the separation gap reaches approximately $\Delta = 3.7$. We therefore set $m = 3$ conservatively, stopping further separation before this regime (horizontal dashed line). The margin objective subsequently keeps the gap close to this target, whereas OCE exhibits unstable behavior. **Bottom:** corresponding AUC, showing that the controlled separation induced by the margin objective supports stronger OOD detection over adaptation.

A.2 Gradient Isolation

In this section, we show that parameter decoupling in DSTA yields exact gradient isolation, with classification and separation gradients supported on disjoint adapter parameters.

Let

$$\phi = [\phi_{\text{cls}}, \phi_{\text{sep}}]$$

denote the joint trainable adapter parameters. Since $\mathcal{L}_{\text{cls}}$ updates only ϕ_{cls} and $\mathcal{L}_{\text{sep}}$ updates only ϕ_{sep}, their gradients over the joint parameter space are

$$\nabla_\phi \mathcal{L}_{\text{cls}} = \begin{bmatrix} \nabla_{\phi_{\text{cls}}} \mathcal{L}_{\text{cls}} \\ 0 \end{bmatrix}, \qquad \nabla_\phi \mathcal{L}_{\text{sep}} = \begin{bmatrix} 0 \\ \nabla_{\phi_{\text{sep}}} \mathcal{L}_{\text{sep}} \end{bmatrix} \tag{7}$$

Their inner product is therefore

$$\langle \nabla_\phi \mathcal{L}_{\text{cls}}, \nabla_\phi \mathcal{L}_{\text{sep}} \rangle = \langle \nabla_{\phi_{\text{cls}}} \mathcal{L}_{\text{cls}}, 0 \rangle + \langle 0, \nabla_{\phi_{\text{sep}}} \mathcal{L}_{\text{sep}} \rangle = 0 \tag{8}$$

Hence, the two objectives have disjoint gradient support over the trainable adapter parameters.

A.3 DSTA Adaptation Procedure

Algorithm 1 summarizes the complete test-time adaptation procedure of DSTA. For each incoming batch, DSTA first activates the separation adapters to compute ID-confidence scores, soft assignments, and the reliable-ID mask, and optimizes $\mathcal{L}_{\text{sep}}$. It then switches to the classification adapters,

re-encodes the same batch, selects the samples indicated by m_{id}, and optimizes $\mathcal{L}_{\text{cls}}$ only on this subset. The two losses are backpropagated through their respective adapter sets, while the shared backbone remains frozen.

Algorithm 1 DSTA test-time adaptation procedure.

Require: CLIP vision encoder f_v, text prototypes C, classif. parameters ϕ_{cls}, sep. parameters ϕ_{sep}

for each test batch x **do**
  Activate ϕ_{sep}
  $z_{\text{sep}} \leftarrow f_v(x; \theta, \phi_{\text{sep}})$
  $s_i \leftarrow \text{MaxLogSoftmax}(z_{\text{sep}}, C)$ {Eq. 1}
  $w_i \leftarrow \sigma((s_i - \alpha)/\gamma)$
  $m_{\text{id}} \leftarrow (\mathbf{1}[w_i > 0.5])_i$
  $\mathcal{L}_{\text{sep}} \leftarrow \mathcal{L}_{\text{sep}}(s_i, w_i)$
  Activate ϕ_{cls}
  $z_{\text{cls}} \leftarrow f_v(x; \theta, \phi_{\text{cls}})$
  $z_{\text{cls}}^{\text{id}} \leftarrow z_{\text{cls}}[m_{\text{id}}]$
  $\mathcal{L}_{\text{cls}} \leftarrow \mathcal{L}_{\text{cls}}(z_{\text{cls}}^{\text{id}}, C)$
  Backpropagate $\mathcal{L}_{\text{cls}}$ only through ϕ_{cls}
  Backpropagate $\mathcal{L}_{\text{sep}}$ only through ϕ_{sep}
end for

A.4 Gradient interaction across other datasets

We further analyze the cosine similarity between the classification and separation gradients across additional benchmarks in Fig. 5. The degree of alignment varies across datasets: CIFAR-10 and CIFAR-100 exhibit individual optimization steps with negative cosine similarity, while HAC and CCT remain positively but only moderately aligned. These results indicate that direct gradient conflict is dataset-dependent, but that the two objectives are not consistently aligned under shared-parameter adaptation. DSTA avoids relying on such compatibility by optimizing classification and separation over disjoint trainable parameter spaces.

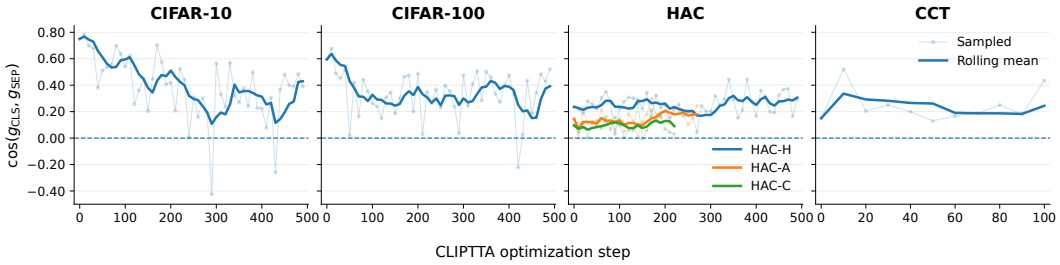

Figure 5: **Gradient interaction across datasets.** Cosine similarity between the classification and separation gradients during shared-parameter CLIPTTA adaptation.

A.5 Automatic OOD Logit-Scale Calibration

We provide additional details on the automatic OOD logit-scale calibration introduced in Sec. 4. Recall that the separation stream computes an ID-confidence score from the similarities between each visual representation and the known-class text prototypes. The logit scale a determines how strongly differences among these similarities are reflected in the resulting confidence scores.

Let $z_{\text{sep},i}$ denote the normalized separation-stream representation of sample i, and let C_c denote the normalized text prototype of known class c. Their image–text similarity is $q_{ic} = z_{\text{sep},i}^{\top} C_c$, and $\mathbf{q}_i = (q_{i1}, \ldots, q_{iC})$ collects the similarities of sample i to all known classes. For a candidate scale a, the corresponding ID-confidence score is

$$s_i(a) = \max_c \log \left[\text{softmax}(a\mathbf{q}_i)\right]_c = \max_c \log \frac{\exp(aq_{ic})}{\sum_{k=1}^{C} \exp(aq_{ik})} \tag{9}$$

Applying the same candidate scale to all samples produces the empirical batch-level score distribution $\mathcal{S}(a) = \{s_i(a)\}_i$.

The scale a changes both the sharpness of the class distributions and, consequently, the spread of $\mathcal{S}(a)$. When a is very small, $\mathrm{softmax}(a\mathbf{q}_i)$ approaches a uniform distribution over the C known classes, so the maximum log-softmax scores tend to concentrate around $-\log C$. Increasing a amplifies differences among the class similarities, allowing samples with different degrees of compatibility with the known classes to produce more differentiated confidence scores. Conversely, for excessively large scales, highly peaked class distributions can cause many scores to saturate near zero. Hence, rather than fixing the raw value of a, we calibrate it according to the dispersion induced in score space.

Specifically, we evaluate a predefined grid $\mathcal{G} = \{a_k\}_{k=1}^{K}$, with $a_k \in [1, 100]$. For each candidate, we compute $\mathcal{S}(a)$ and measure its standard deviation. We then select

$$a^\star = \arg\min_{a \in \mathcal{G}} |\mathrm{Std}(\mathcal{S}(a)) - \sigma_{\mathrm{target}}| \tag{10}$$

Thus, σ_{target} specifies a reference dispersion in the ID-confidence score space, rather than prescribing a particular raw logit scale. The calibration therefore selects the value of a whose induced batch-level score distribution most closely matches this reference.

Importantly, this procedure is fully label-free. It uses only the empirical distribution of confidence scores observed in the first incoming test batch and does not require knowing which samples are ID or OOD. We keep σ_{target} fixed across all datasets, while allowing the selected $a^\star$ to vary according to the similarity statistics of the incoming data. This separates the desired behavior in score space from the dataset-dependent raw scale needed to obtain it.

Calibration is performed only once, on the first test batch. The selected $a^\star$ initializes the OOD logit scale a, which is subsequently treated as a trainable parameter and refined together with ϕ_{sep} through the separation objective $\mathcal{L}_{\mathrm{sep}}$. We use $\sigma_{\mathrm{target}} = 0.5$ for all datasets.

A.6 OOD LOGIT-SCALE CALIBRATION SENSITIVITY

Tab. 5 studies the effect of the calibration target σ_{target} on ImageNet. Very small target values provide insufficient score dispersion and degrade ID/OOD separation, while overly large values lead to excessively large calibrated logit scales and again reduce performance. Performance remains stable across the intermediate range 0.3-0.7, with the strongest results concentrated around 0.4-0.5. This broad operating range indicates that the calibration mechanism is not sensitive to a narrowly tuned value of σ_{target}. We therefore use $\sigma_{\mathrm{target}} = 0.5$ throughout all experiments. Overall, the analysis shows that automatic logit-scale calibration provides reliable ID/OOD separation over a substantial range of target dispersions, avoiding the need to select a dataset-specific logit scale.

Table 5: **Effect of OOD logit-scale calibration on ImageNet.** Sensitivity to the target score standard deviation σ_{target} and the resulting calibrated logit scale α^*.

σ_{target}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
α^*	3.2	6.7	10.4	13.9	17.3	21.5	24.9	93.0
Acc	68.9	68.7	69.1	**69.2**	**69.2**	69.1	69.1	**69.2**
AUC	50.7	97.9	99.5	**99.8**	**99.8**	99.7	99.6	95.7
FPR95	89.2	13.2	1.4	1.0	**0.9**	1.2	1.8	16.4
H	23.6	82.7	86.5	**86.8**	**86.8**	86.7	86.5	81.3

A.7 EVALUATION WITH A SMALLER BACKBONE

Tab. 6 evaluates DSTA with the smaller OpenVision ViT-S/16 backbone to assess whether the behavior observed with ViT-B/16 extends to a different model scale. On clean data, DSTA remains closely competitive with the strongest method overall. Relative to UniEnt, which obtains the best aggregate clean results, DSTA differs by only 0.1 points in ACC, 0.4 points in AUC, 1.4 points in FPR95, and 1.1 points in H-score. At the same time, DSTA improves over CLIPTTA on all four aggregate metrics, increasing ACC by 0.3 points and AUC by 2.1 points, while reducing FPR95 by 4.8 points and improving H-score by 6.3 points. The per-dataset results further show that the proposed design is not tied to a single benchmark: DSTA attains the best ACC and H-score on CIFAR-10 and CIFAR-100, and the best ACC and AUC on ImageNet. The remaining clean settings are more

Table 6: **Open-set TTA results with CLIP ViT-S/16.** Panel (a) reports clean results and Panel (b) the corresponding corrupted settings. Best results are in bold and second-best results are underlined.

	CIFAR-10				CIFAR-100				ImageNet				HAC				CCT				Average			
	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑
(a) Clean																								
OSTTA	88.2	81.9	46.4	71.1	69.9	29.2	94.1	13.8	56.0	49.2	93.1	16.4	60.0	46.8	97.6	6.3	15.8	48.7	94.7	11.0	58.0	51.2	85.2	23.7
SoTTA	61.8	68.5	63.2	51.8	70.9	88.1	33.1	74.3	19.5	44.7	81.2	23.7	50.3	80.6	79.4	32.4	15.5	59.1	92.0	14.5	43.6	68.2	69.8	39.3
STAMP	93.8	99.6	0.3	97.6	74.4	95.6	28.4	79.2	33.2	65.1	78.3	32.8	62.1	70.1	95.4	11.4	**16.5**	53.9	94.2	11.9	56.0	76.9	59.3	46.6
UniEnt	94.4	**100.0**	**0.0**	98.1	77.5	**99.7**	1.1	90.8	64.1	94.8	**18.8**	**78.0**	**70.0**	**82.6**	62.4	**54.3**	**16.5**	59.7	92.3	14.5	**64.5**	**87.4**	**34.9**	**67.1**
CLIPTTA	94.6	**100.0**	**0.0**	98.1	78.3	**99.7**	**0.9**	91.2	65.2	94.7	20.6	77.9	66.2	69.4	94.1	14.4	16.0	**60.9**	**90.0**	**16.8**	64.1	84.9	41.1	59.7
DSTA (ours)	**95.4**	**100.0**	**0.0**	**98.4**	**79.0**	99.5	1.5	**91.3**	**65.6**	**95.4**	25.4	76.7	65.9	79.9	**61.6**	50.4	15.9	60.3	93.1	13.4	64.4	87.0	36.3	66.0
(b) Corrupted																								
OSTTA	63.9	72.7	55.2	58.0	40.2	34.1	92.7	15.7	28.0	48.9	94.2	13.1	38.1	43.3	94.8	12.4	14.5	47.1	95.2	10.0	36.9	49.2	86.4	21.8
SoTTA	38.8	62.9	74.2	37.3	36.3	65.5	76.8	34.9	3.4	38.2	93.6	6.3	41.7	68.0	83.5	29.4	15.6	57.7	92.3	14.2	27.2	58.5	84.1	24.4
STAMP	71.6	94.2	36.2	74.5	44.5	79.2	81.3	33.9	13.0	43.9	96.0	8.6	49.5	70.7	88.1	24.8	14.2	51.5	95.1	10.2	38.6	67.9	79.3	30.4
UniEnt	82.4	**99.2**	2.6	92.4	57.2	95.2	20.2	74.0	29.4	61.4	62.1	39.1	**61.8**	**84.3**	61.4	**55.5**	**15.8**	**59.5**	91.0	15.7	49.3	79.9	47.5	55.3
CLIPTTA	83.4	98.4	7.0	91.2	59.1	94.4	29.8	71.8	**39.7**	**83.6**	58.8	**48.8**	54.9	73.7	84.3	31.0	14.6	56.4	93.5	12.5	**50.3**	**81.3**	54.7	51.1
DSTA (ours)	**84.5**	99.1	**2.0**	**93.4**	**60.0**	**98.5**	**5.6**	**80.2**	33.4	58.2	**50.6**	44.5	54.3	80.4	**59.5**	53.7	14.5	58.9	**89.9**	**16.2**	49.3	79.0	**41.5**	**57.6**

metric-dependent. In particular, on ImageNet the improved classification and AUC do not translate into the best FPR95, while HAC and CCT exhibit stronger trade-offs between classification and OOD detection. Thus, with the smaller backbone, DSTA preserves the main benefit of decoupled adaptation while no longer dominating every aggregate clean metric.

The advantage of DSTA becomes clearer under corruption. Averaged across the five corrupted benchmarks, DSTA achieves the best FPR95 and H-score. Relative to UniEnt, DSTA matches aggregate ACC, reduces FPR95 by 6.0 points, and improves H-score by 2.3 points, with only a 0.9-point decrease in AUC. Compared with CLIPTTA, DSTA reduces FPR95 by 13.2 points and improves H-score by 6.5 points, at the cost of 1.0 point in ACC and 2.3 points in AUC. This aggregate behavior is supported by several individual benchmarks: DSTA obtains the best FPR95 on all five corrupted datasets, the best H-score on CIFAR-10-C, CIFAR-100-C, and CCT-C, and the best results across all four metrics on CIFAR-100-C. On ImageNet-C and HAC-C, the trade-off is more pronounced: DSTA provides substantially stronger rejection, but does not lead all classification or AUC metrics. Taken together, these results show that the characteristic behavior of DSTA transfers to the smaller backbone: its strongest and most consistent advantage lies in improving unknown-sample rejection while retaining competitive ID classification, particularly under covariate shift.

A.8 Long-Horizon Evolution of the Separation Stream

We analyze the separation stream over a longer continuous adaptation trajectory. Fig. 6 shows the evolution of the ID/OOD score distributions and separation representations at 0, 10, 100, 500, and 1000 optimization steps. Although standard TTA processes successive test batches, here we keep the adaptation and visualization samples fixed to isolate the effect of optimization and make the evolution of the separation space directly comparable across checkpoints. The same model is adapted continuously without reinitialization. While DSTA uses 10 steps in the main experiments, longer optimization reveals increasing ID/OOD separation in both score and representation space.

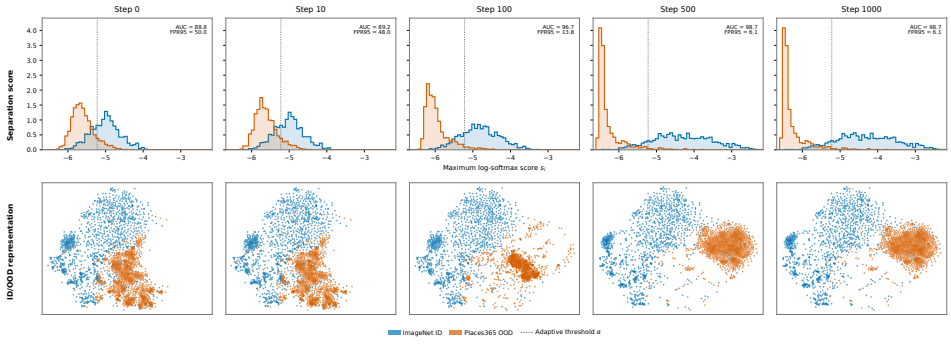

Figure 6: **Long-horizon evolution of the DSTA separation stream.** Top: ID/OOD separation-score distributions. Bottom: joint t-SNE visualization of the corresponding separation representations. Columns show the same continuously adapted model at 0, 10, 100, 500, and 1000 steps.

Importantly, the gains persist under prolonged optimization: after the initial improvement, both the score distributions and the separation representations stabilize, with AUC and FPR95 remaining unchanged between 500 and 1000 steps. This provides additional evidence that the bounded separation objective supports stable long-horizon adaptation rather than continuously driving the two groups apart.

A.9 Further Implementation Details

All experiments were performed on two NVIDIA A6000 GPUs. We optimize using Adam. For DSTA with ViT-B/16, both streams use a learning rate of 5×10^{-5} on CIFAR. On ImageNet/Places365, the classification-stream learning rate is 5×10^{-6}, while the separation-stream learning rate is 10^{-5} on clean data and 2×10^{-5} under corruption. HAC/EPIC and CCT/CCT-C use classification- and separation-stream learning rates of 10^{-6} and 10^{-5}, respectively, with the same rates for clean and corrupted data. The OOD logit-scale learning rate is set equal to the separation-stream learning rate. OpenVision ViT-S/16 follows the same schedule with the stream-specific learning rates halved. For DSTA, LoRA dropout and weight decay are set to 0.25 and 0.2 by default; both are disabled for ImageNet/Places365, where the larger adaptation set is less prone to overfitting. All DSTA loss coefficients are set to 1.

Implementation Details of SOTA. For all competing open-set TTA methods, we adopt the hyperparameter and optimization settings reported in the corresponding papers and/or official repositories, with only the dataset-specific adjustments required by our evaluation protocol. The configurations used in our experiments are summarized below.

- **OSTTA:** We set the confidence-improvement margin to $\delta = 0.4$, with $\beta_{\mathrm{TTA}} = 1$ and $\beta_{\mathrm{reg}} = 0.1$, and optimize with Adam using a learning rate of 10^{-3}.
- **SoTTA:** We use the confidence threshold $\tau = 1/C$, where C is the number of ID classes, and set the memory capacity to 128. We perform one SAM update per batch with perturbation radius $\rho = 0.05$, using Adam as the base optimizer with a learning rate of 10^{-3}.
- **STAMP:** We use an entropy-filtering threshold $H_{\mathrm{thr}} = 0.45 \log C$ and a balanced memory capacity of 128 samples, and optimize with Adam using a learning rate of 10^{-3}.
- **UniEnt:** We set $\lambda_{\mathrm{TTA}} = \lambda_{\mathrm{reg}} = \lambda_{\mathrm{OOD}} = 1$ and optimize with Adam using a learning rate of 10^{-4}.
- **CLIPTTA:** We follow the default CLIPTTA hyperparameterization where applicable, setting $\beta_{\mathrm{TTA}} = \beta_{\mathrm{reg}} = \beta_{\mathrm{OOD}} = 1$ and using Adam. The learning rate is 10^{-4} for CIFAR-10, CIFAR-100, and ImageNet, and 10^{-5} for HAC and CCT. We use $\gamma = 0.005$.

For all competing methods, the ID and OOD batch sizes are both set to 128. Following their standard adaptation setup, only the affine parameters of the visual normalization layers are optimized, while the text encoder remains frozen.

A.10 HAC and CCT Evaluation Protocols

For HAC-EPIC-KITCHENS, we use the H, A, and C domains as ID in separate settings and pair each with EPIC-KITCHENS as OOD. For CCT, images from all 10 training cameras are used as ID, while the single available trans-validation camera is used as OOD. Unlike the other benchmarks, the ID and OOD sets share semantic classes. For the corrupted variants, HAC-C, EPIC-KITCHENS-C, and CCT-C are generated with the same corruption procedure used for ImageNet-C (Hendrycks & Dietterich, 2019). For CCT, we subsample the ID images to 10% (about 1,350 images) so that the number of ID samples is approximately matched to the number of images from the OOD trans-validation camera (about 1,600 images).

A.11 Computational Cost Analysis

We profile the computational cost of CLIPTTA and DSTA under the same ImageNet/Places365 setting used in the main experiments. Both methods use the OpenAI CLIP ViT-B/16 backbone, a batch

of 128 ID and 128 OOD samples, and 10 adaptation steps per incoming batch. We measure adaptation computation only, excluding data loading and disk I/O. GPU execution time is measured on two NVIDIA A6000 GPUs using CUDA events with synchronization outside the measured region.

Tab. 7 shows that DSTA introduces only a small increase in trainable model size, with a 2.1% parameter overhead relative to CLIPTTA. Its two-stream design increases total adaptation time by 29.3%, but this overhead is concentrated primarily in the forward pass, since each incoming batch is evaluated once per objective-specific stream. In contrast, backward computation increases by only 6.6%. This is because both streams share the same frozen backbone and optimize only their respective adapter parameters, avoiding duplication of full backbone optimization.

Despite the additional forward computation, DSTA reduces measured peak memory by 15.1% relative to CLIPTTA. This behavior is consistent with the sequential execution of the two streams, which limits the number of objective-specific activations that must be retained simultaneously. Overall, DSTA trades a moderate increase in adaptation time for objective-specific decoupling while retaining a small parameter overhead and lower peak memory usage.

Table 7: **Computational cost analysis on ImageNet/Places365 with the CLIP ViT-B/16 model.** Parameter counts refer to the CLIP vision encoder only and exclude the CLIP text encoder.

Method	Params (M)	Total ↓	Forward ↓	Backward ↓	Throughput ↑	Peak Mem. ↓
CLIPTTA	63.47	17.96 s	4.86 s	7.82 s	14.40	24.20 GB
DSTA	64.78	23.21 s	9.92 s	8.34 s	11.08	20.55 GB
Relative change	+2.1%	+29.3%	+104.3%	+6.6%	-23.1%	-15.1%

A.12 Multiple-Run Statistics and Significance Analysis

We assess whether the observed improvements are robust to optimization randomness by repeating DSTA and CLIPTTA with three matched random seeds (42, 43, and 44) for each dataset. Tab. 8 reports the mean and standard deviation across runs. We additionally perform two-sided paired t-tests between DSTA and CLIPTTA using runs with the same random seed.

Despite the small number of independent runs, several of the main improvements remain statistically significant (($p_{\mathrm{adj}} < 0.05$)). In particular, DSTA significantly improves both ACC and AUC on CIFAR-10, CIFAR-100 and ImageNet, while its substantial FPR95 reduction on ImageNet also remains significant. The improvement in FPR95 is additionally significant on the CIFAR-100, ImageNet, HAC-H, and HAC-A settings. On the remaining dataset-metric combinations, the direction of the mean differences is generally consistent with the main results, although the available three-run sample does not provide sufficient evidence to establish significance after multiple-comparison correction. Overall, the repeated-run analysis supports that the principal gains of DSTA are reproducible across random seeds rather than being driven by a single favorable run.

Table 8: **Multiple-run statistics of DSTA and CLIPTTA across three matched random seeds per dataset.** Results report mean ± standard deviation. Better results between the two methods are shown in bold. Blue cells indicate a statistically significant difference between DSTA and CLIPTTA under a two-sided paired t-test ($p_{\mathrm{adj}} < 0.05$), with DSTA achieving the better result.

	ACC↑		AUC↑		FPR95↓	
Dataset	DSTA	CLIPTTA	DSTA	CLIPTTA	DSTA	CLIPTTA
CIFAR-10	**95.0** ± 0.12	94.0 ± 0.15	**100.0** ± 0.01	99.9 ± 0.00	**0.0** ± 0.01	0.1 ± 0.03
CIFAR-100	**75.5** ± 0.27	74.2 ± 0.20	**100.0** ± 0.01	99.6 ± 0.03	**0.1** ± 0.01	1.6 ± 0.15
ImageNet	**69.2** ± 0.04	67.9 ± 0.08	**99.8** ± 0.05	97.4 ± 0.19	**0.7** ± 0.25	10.9 ± 0.48
HAC-H	**84.5** ± 0.33	84.2 ± 0.30	**82.7** ± 0.79	80.2 ± 0.30	**52.1** ± 0.71	89.3 ± 0.92
HAC-A	**88.2** ± 0.06	88.0 ± 0.12	**83.2** ± 0.84	80.9 ± 0.22	**59.1** ± 0.91	78.0 ± 2.06
HAC-C	**65.6** ± 0.84	65.5 ± 0.90	**56.5** ± 1.35	56.4 ± 0.27	**91.4** ± 1.42	96.6 ± 0.03
CCT	35.5 ± 1.16	**35.6** ± 1.26	**67.8** ± 0.63	64.1 ± 0.55	**78.6** ± 1.51	82.5 ± 1.06

A.13 EXPANDED HAC RESULTS

The main results report HAC as the metric-wise average over the H/EK, A/EK, and C/EK settings. For completeness, Tables 9 and 10 report the corresponding results for each HAC domain separately.

The expanded HAC results show that DSTA's gains are consistent across domains rather than being driven by a single HAC split. With CLIP ViT-B/16 (Table 9), DSTA improves both FPR95 and H-score over CLIPTTA on all three clean settings, raising the average H-score from 28.2 to 50.3 while reducing average FPR95 from 86.8 to 67.7. Under corruption, DSTA again improves both metrics over CLIPTTA on every split and achieves the best average FPR95 (49.4) and H-score (64.5). The largest gain appears on C/EK, where DSTA reaches an H-score of 72.6 compared with 48.8 for CLIPTTA.

The same pattern largely transfers to the smaller OpenVision ViT-S/16 backbone (Table 10). On clean data, DSTA improves FPR95 and H-score over CLIPTTA on all three splits, attaining the best H-score on H/EK and A/EK and the best average FPR95 (61.6). Under corruption, DSTA preserves this consistent advantage over CLIPTTA across all three domains, achieving the best average FPR95 (59.5) and the second-best average H-score (53.7), close to UniEnt at 55.5.

Overall, these split-level results show that DSTA's improved ID/OOD separation is consistent across heterogeneous HAC domains and backbone sizes, with particularly strong unknown rejection under corruption while maintaining competitive classification.

Table 9: **Expanded HAC results with CLIP ViT-B/16.** Panel (a) reports the individual H/EK, A/EK, and C/EK settings on clean data, while Panel (b) reports their corrupted counterparts. The final block reports their HAC average. Best results are in bold and second-best results are underlined.

	H/EK				A/EK				C/EK				Average			
	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑
(a) Clean																
CLIP	83.2	77.6	93.1	17.6	86.9	79.1	81.8	37.9	64.2	55.1	97.7	6.3	78.1	70.6	90.9	20.6
OSTTA	73.1	34.6	97.8	6.2	81.9	45.7	93.4	16.1	63.4	33.2	98.9	3.3	72.8	37.8	96.7	8.5
SoTTA	77.5	81.0	79.6	40.4	81.6	83.6	68.2	53.9	57.7	58.2	94.9	13.0	72.3	74.3	80.9	35.8
STAMP	79.9	75.1	94.7	14.0	86.1	77.0	86.1	31.1	60.5	54.7	98.1	5.3	75.5	68.9	93.0	16.8
UniEnt	81.7	**90.4**	**50.2**	**69.1**	85.3	**89.3**	**56.0**	**65.7**	**70.1**	**67.3**	**88.9**	**25.1**	79.0	**82.3**	**65.1**	**53.3**
CLIPTTA	84.2	81.0	85.8	31.7	87.9	80.5	78.4	42.8	66.3	56.8	96.2	10.2	79.5	72.8	86.8	28.2
DSTA (ours)	**84.6**	81.6	53.2	66.0	**88.2**	82.0	60.2	61.7	66.5	57.7	89.6	23.3	**79.7**	73.8	67.7	50.3
(b) Corrupted																
CLIP	67.0	74.1	88.6	25.9	71.2	77.0	76.7	42.8	56.0	69.4	82.3	33.8	64.8	73.5	82.5	34.2
OSTTA	49.9	30.7	96.3	9.3	60.8	35.9	94.1	14.1	42.5	35.2	94.9	12.0	51.1	33.9	95.1	11.8
SoTTA	52.9	55.3	88.7	23.9	65.4	73.0	77.4	41.0	51.5	64.6	85.5	28.9	56.6	64.3	83.9	31.3
STAMP	63.9	71.7	91.4	20.6	70.2	74.9	79.0	39.9	52.2	67.8	82.7	32.7	62.1	71.5	84.4	31.1
UniEnt	**70.6**	**86.6**	56.8	**61.4**	**74.5**	**89.5**	**44.4**	**70.5**	64.3	80.7	61.5	55.6	**69.8**	**85.6**	54.2	62.5
CLIPTTA	58.7	71.0	81.4	35.3	70.2	78.5	82.0	36.4	**73.2**	80.3	71.7	48.8	67.4	76.6	78.4	40.1
DSTA (ours)	58.8	80.2	**53.5**	58.9	70.5	85.3	55.4	62.1	**73.2**	**89.4**	**39.3**	**72.6**	67.5	85.0	**49.4**	**64.5**

Table 10: **Expanded HAC results with ViT-S/16.** Panel (a) reports the individual H/EK, A/EK, and C/EK settings on clean data, while Panel (b) reports their corrupted counterparts. The final block reports their HAC average. Best results are in bold and second-best results are underlined.

	H/EK				A/EK				C/EK				Average			
	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑	ACC↑	AUC↑	FPR95↓	H↑
(a) Clean																
OSTTA	64.7	35.8	98.4	4.4	64.3	60.2	94.5	14.0	51.0	44.5	99.8	0.5	60.0	46.8	97.6	6.3
SoTTA	54.7	78.6	82.9	33.5	53.0	**90.1**	58.3	55.6	43.4	**73.2**	97.0	8.0	50.3	80.6	79.4	32.4
STAMP	71.8	77.5	96.2	10.4	65.8	74.7	90.6	22.2	48.7	58.2	99.5	1.6	62.1	70.1	95.4	11.4
UniEnt	75.0	**92.2**	40.4	73.2	**72.4**	83.9	62.6	57.2	**62.6**	71.6	**84.0**	**32.5**	**70.0**	**82.6**	62.4	**54.3**
CLIPTTA	**75.4**	77.5	94.5	14.5	70.5	73.4	88.6	26.0	52.6	57.4	99.1	2.6	66.2	69.4	94.1	14.4
DSTA (ours)	75.1	91.8	**40.3**	**73.3**	70.1	89.2	**48.9**	**66.6**	52.4	58.8	95.6	11.4	65.9	79.9	**61.6**	50.4
(b) Corrupted																
OSTTA	37.8	36.3	95.4	11.1	44.4	49.6	92.2	17.5	32.0	43.9	96.7	8.5	38.1	43.3	94.8	12.4
SoTTA	40.1	56.6	87.8	24.0	47.7	77.8	73.0	42.4	37.4	69.7	89.7	21.7	41.7	68.0	83.5	29.4
STAMP	54.7	72.2	89.8	23.0	54.2	73.4	82.3	33.9	39.7	66.3	92.3	17.6	49.5	70.7	88.1	24.8
UniEnt	**65.5**	**86.3**	58.9	**58.6**	**64.7**	**86.0**	56.7	59.8	**55.3**	**80.5**	**68.5**	**48.1**	**61.8**	**84.3**	61.4	**55.5**
CLIPTTA	60.8	76.4	84.8	31.5	58.8	76.3	78.6	39.1	45.1	68.3	89.7	22.4	54.9	73.7	84.3	31.0
DSTA (ours)	60.1	86.2	**56.9**	58.3	58.3	82.9	**49.0**	**61.5**	44.4	72.0	72.6	41.2	54.3	80.4	**59.5**	53.7