

CLIPTTA: Robust Contrastive Vision-Language Test-Time Adaptation - Appendix

The appendix is organized as follows. In Appendix A, we present a detailed theoretical analysis of the gradients of the $\mathcal{L}_{\text{CLIPTTA}}$ loss. In Appendix B, we provide additional insights into how $\mathcal{L}_{\text{CLIPTTA}}$ helps mitigate collapse and pseudo-labeling errors. Finally, in Appendix C, we elaborate on the experimental protocol and include additional experimental results.

A Theoretical Analysis

A.1 Gradient Analysis

In this section, we provide a detailed analysis of the gradients of the TENT loss $\mathcal{L}_{\text{TENT}}$, CLIP’s contrastive loss $\mathcal{L}_{\text{cont.}}$ (*i.e.* using hard pseudo-captions), our soft contrastive loss $\mathcal{L}_{\text{s-cont.}}$, the regularization loss \mathcal{L}_{reg} and the CLIPTTA loss $\mathcal{L}_{\text{CLIPTTA}}$. Furthermore, we show how, benefiting from the information of other predictions in the batch, both contrastive losses allow to avoid collapse. Finally, when combined with the regularization loss, CLIPTTA allows mitigating the effect of pseudo label errors.

Let’s consider a batch of examples $\mathbf{x}_1, \dots, \mathbf{x}_N$, and let’s write N_k , the number of predictions assigned to the k^{th} class. To simplify the computations, we place ourselves in the case where only the parameters of the visual encoder are updated.

Gradient of $\mathcal{L}_{\text{TENT}}$. We recall that the TENT loss writes as follows:

$$\mathcal{L}_{\text{TENT}} = - \sum_{k=1}^C q_{ik} \log q_{ik}$$

with q_{ik} the probability of image \mathbf{x}_i being classified as class k (see Eq. (1)). The gradient of $\mathcal{L}_{\text{TENT}}$ w.r.t. \mathbf{z}_i is:

$$\begin{aligned} \nabla_{\mathbf{z}_i} \mathcal{L}_{\text{TENT}} &= - \sum_{k=1}^C \nabla_{\mathbf{z}_i} q_{ik} \log q_{ik} = - \sum_{k=1}^C (1 + \log q_{ik}) \nabla_{\mathbf{z}_i} q_{ik} \\ &= - \sum_{k=1}^C (1 + \log q_{ik}) q_{ik} \sum_{c=1}^C q_{ic} (\mathbf{z}_t^k - \mathbf{z}_t^c) \quad (10) \\ &= - \sum_{k=1}^C \left[\sum_{c=1}^C \log \frac{q_{ik}}{q_{ic}} q_{ic} \right] q_{ik} \mathbf{z}_t^k \end{aligned}$$

From Eq. (10) we can see that the gradient will always push \mathbf{z}_i in the direction of the predicted class \hat{k} because in that case we have $\log \frac{q_{ik}}{q_{ic}} > 0, \forall c \neq \hat{k}$. And there is no mechanism allowing to reduce the magnitude of the gradient towards the predicted class even when we are approaching a situation of collapse.

Gradient of $\mathcal{L}_{\text{cont.}}$ Using the notation introduced in the main paper, let $\hat{\mathbf{t}}_i$ represent the pseudo-caption associated with the example \mathbf{x}_i in the batch, and let $p(\hat{\mathbf{t}}_j | \mathbf{x}_i)$ denote the probability of \mathbf{x}_i matching $\hat{\mathbf{t}}_j$ within the batch. Specifically, we have:

$$p(\hat{\mathbf{t}}_j | \mathbf{x}_i) = \frac{e^{\mathbf{z}_i^\top \hat{\mathbf{z}}_t^j}}{\sum_{l=1}^N e^{\mathbf{z}_i^\top \hat{\mathbf{z}}_t^l}}$$

The unsymmetrized version of CLIP’s contrastive loss writes:

$$\mathcal{L}_{\text{cont.}} = \sum_{i=1}^N -\log p(\hat{\mathbf{t}}_i | \mathbf{x}_i) = \sum_{i=1}^N -\mathbf{z}_i^\top \hat{\mathbf{z}}_t^i + \log \left(\sum_{j=1}^N e^{\mathbf{z}_i^\top \hat{\mathbf{z}}_t^j} \right) = \sum_{i=1}^N -\mathbf{z}_i^\top \hat{\mathbf{z}}_t^i + \log \left(\sum_{k=1}^C N_k e^{\mathbf{z}_i^\top \mathbf{z}_t^k} \right)$$

where $\hat{\mathbf{z}}_t^j$ is the embedding of the pseudo caption associated with image \mathbf{x}_j and \mathbf{z}_t^k is the embedding of class k . Let's compute the gradient of $\mathcal{L}_{\text{cont.}}$ w.r.t. \mathbf{z}_i :

$$\begin{aligned}
\nabla_{\mathbf{z}_i} \mathcal{L}_{\text{cont.}} &= -\hat{\mathbf{z}}_t^i + \frac{\nabla_{\mathbf{z}_i} \sum_{k=1}^C N_k e^{\mathbf{z}_i^\top \mathbf{z}_t^k}}{\sum_{c=1}^C N_c e^{\mathbf{z}_i^\top \mathbf{z}_t^c}} = -\hat{\mathbf{z}}_t^i + \sum_{k=1}^C \frac{N_k e^{\mathbf{z}_i^\top \mathbf{z}_t^k}}{\sum_{c=1}^C N_c e^{\mathbf{z}_i^\top \mathbf{z}_t^c}} \mathbf{z}_t^k \\
&= -\hat{\mathbf{z}}_t^i + \sum_{k=1}^C \frac{N_k e^{\mathbf{z}_i^\top \mathbf{z}_t^k}}{\sum_{c=1}^C N_c e^{\mathbf{z}_i^\top \mathbf{z}_t^c}} \frac{\sum_{c=1}^C e^{\mathbf{z}_i^\top \mathbf{z}_t^c}}{\sum_{c=1}^C e^{\mathbf{z}_i^\top \mathbf{z}_t^c}} \mathbf{z}_t^k \\
&= -\hat{\mathbf{z}}_t^i + \sum_{k=1}^C \frac{N_k q_{ik}}{\sum_{c=1}^C N_c q_{ic}} \mathbf{z}_t^k \\
&= -\hat{\mathbf{z}}_t^i + \sum_{k=1}^C w_{k,i} \mathbf{z}_t^k
\end{aligned} \tag{11}$$

with $w_{k,i} = \frac{N_k q_{ik}}{\sum_{c=1}^C N_c q_{ic}}$.

From Eq. (11), we observe that CLIP's contrastive loss consistently drives the visual embedding \mathbf{z}_i toward the embedding of its predicted class $\hat{\mathbf{z}}_t^i$, as $w_{k,i} \leq 1$. However, the gradient's magnitude is influenced by the proportion of predictions assigned to the same class within the batch. Specifically, as the system approaches a collapse scenario (i.e., $w_{k,i} \rightarrow 1$), the gradient of $\mathcal{L}_{\text{cont.}}$ diminishes and eventually vanishes:

$$\|\nabla_{\mathbf{z}_i} \mathcal{L}_{\text{cont.}}\| \xrightarrow{w_{k,i} \rightarrow 1} 0 \tag{12}$$

Gradient of $\mathcal{L}_{\text{s-cont}}$ The unsymmetrized version of our $\mathcal{L}_{\text{s-cont}}$ loss writes:

$$\mathcal{L}_{\text{s-cont}} = \sum_{i=1}^N - \sum_{j=1}^N p(\hat{\mathbf{t}}_j | \mathbf{x}_i) \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i)$$

Let's compute the gradient of $\mathcal{L}_{\text{s-cont}}$ w.r.t. \mathbf{z}_i :

$$\begin{aligned}
\nabla_{\mathbf{z}_i} \mathcal{L}_{\text{s-cont}} &= - \sum_{j=1}^N \nabla_{\mathbf{z}_i} [p(\hat{\mathbf{t}}_j | \mathbf{x}_i) \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i)] \\
&= - \sum_{j=1}^N p(\hat{\mathbf{t}}_j | \mathbf{x}_i) \nabla_{\mathbf{z}_i} \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i) + \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i) \nabla_{\mathbf{z}_i} p(\hat{\mathbf{t}}_j | \mathbf{x}_i).
\end{aligned}$$

Using the fact that $\nabla p = p \nabla \log p$, we have:

$$\begin{aligned}
\nabla_{\mathbf{z}_i} \mathcal{L}_{\text{s-cont}} &= - \sum_{j=1}^N p(\hat{\mathbf{t}}_j | \mathbf{x}_i) \nabla_{\mathbf{z}_i} \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i) + \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i) p(\hat{\mathbf{t}}_j | \mathbf{x}_i) \nabla_{\mathbf{z}_i} \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i) \\
&= - \sum_{j=1}^N [1 + \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i)] p(\hat{\mathbf{t}}_j | \mathbf{x}_i) \nabla_{\mathbf{z}_i} \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i)
\end{aligned}$$

Now we can use the fact that $\nabla_{\mathbf{z}_i} - \log p(\hat{\mathbf{t}}_j | \mathbf{x}_i) = -\hat{\mathbf{z}}_t^j + \sum_{k=1}^C w_{k,i} \mathbf{z}_t^k$ based on the computation $\mathcal{L}_{\text{cont.}}$ in Eq. (11). Therefore, we have:

$$\nabla_{z_i} \mathcal{L}_{\text{s-cont}} = \sum_{j=1}^N \beta_{i,j} [-\hat{z}_t^j + \sum_{k=1}^C w_{k,i} z_t^k] \quad (13)$$

861 with $\beta_{i,j} = p(\hat{t}_j | \mathbf{x}_i)(1 + \log p(\hat{t}_j | \mathbf{x}_i))$.

862 The gradient of $\mathcal{L}_{\text{s-cont}}$ does not solely push the visual embedding toward the predicted class. Instead,
 863 it incorporates other predictions within the batch to guide the gradient direction, thereby mitigating the
 864 risk of pseudo-labeling errors. However, similar to CLIP’s contrastive loss, the gradient diminishes as
 865 we approach a collapse scenario. In the case of collapse, where all examples in the batch are predicted
 866 to belong to the same class c , the following conditions hold: $w_c(\mathbf{x}_i) = 1$ and $w_{k,i} = 0, \forall k \neq c$, and
 867 $\hat{z}_t^j = z_t^c \forall j$. Consequently, the term $[-\hat{z}_t^j + \sum_{k=1}^C w_{k,i} z_t^k]$ cancels out, leading to a null gradient.

868 **Binary classification case.** We derive Eq. (7) in the main paper, starting from Eq. (13), and
 869 assuming that the classification task comprises two classes $K = \{a, b\}$, with $N = N_a + N_b$ as the
 870 total batch size. To build on the intuition of the working mechanisms of our soft contrastive loss,
 871 we adopt the case where class a is dominant in the batch (*i.e.*, $N_a \gg N_b$). First, we expand on the
 872 second sum term inside Eq. (13), as follows:

$$\sum_{k=1}^C w_{k,i} z_t^k = w_{a,i} z_t^a + w_{b,i} z_t^b = \frac{N_a q_{ia}}{N_a q_{ia} + N_b q_{ib}} z_t^a + \frac{N_b q_{ib}}{N_a q_{ia} + N_b q_{ib}} z_t^b = \underbrace{\frac{N_a q_{ia} z_t^a + N_b q_{ib} z_t^b}{N_a q_{ia} + N_b q_{ib}}}_Q \quad (14)$$

873 We notice that we can partition the main sum term in Eq. (13) into two sums that account for the N_a
 874 samples predicted as class a , and the N_b samples predicted as class b :

$$\begin{aligned} \nabla_{z_i} \mathcal{L}_{\text{s-cont}} &= \sum_{j=1}^N \beta_{i,j} [-\hat{z}_t^j + Q] = N_a \beta_{ia} [-z_t^a + Q] + N_b \beta_{ib} [-z_t^b + Q] \\ &= N_a \beta_{ia} [-z_t^a + \frac{N_a q_{ia}}{N_a q_{ia} + N_b q_{ib}} z_t^a + \frac{N_b q_{ib}}{N_a q_{ia} + N_b q_{ib}} z_t^b] \\ &\quad + N_b \beta_{ib} [-z_t^b + \frac{N_a q_{ia}}{N_a q_{ia} + N_b q_{ib}} z_t^a + \frac{N_b q_{ib}}{N_a q_{ia} + N_b q_{ib}} z_t^b] \\ &= N_a \beta_{ia} [\frac{-N_b q_{ib}}{N_a q_{ia} + N_b q_{ib}} z_t^a + \frac{N_b q_{ib}}{N_a q_{ia} + N_b q_{ib}} z_t^b] \\ &\quad + N_b \beta_{ib} [\frac{-N_a q_{ia}}{N_a q_{ia} + N_b q_{ib}} z_t^b + \frac{N_a q_{ia}}{N_a q_{ia} + N_b q_{ib}} z_t^a] \\ &= \beta_{ia} q_{ib} \frac{N_a N_b}{N_a q_{ia} + N_b q_{ib}} (z_t^b - z_t^a) - \beta_{ib} q_{ia} \frac{N_a N_b}{N_a q_{ia} + N_b q_{ib}} (z_t^b - z_t^a) \\ &= [\beta_{i,a} q_{ib} - \beta_{i,b} q_{ia}] \frac{N_a N_b}{N_a q_{ia} + N_b q_{ib}} (z_t^b - z_t^a) \end{aligned} \quad (15)$$

875 As pointed out, the increasing dominance of class a ($N_b \rightarrow 0$) reduces the gradient to 0, vanishing
 876 the negative effect of class collapse.

877 **Gradient of \mathcal{L}_{reg} .** The regularization loss \mathcal{L}_{reg} writes as:

$$\mathcal{L}_{\text{reg}} = - \sum_{c=1}^C \bar{q}_c \log \bar{q}_c. \quad (16)$$

878 where \bar{q}_c correspond to the average predicted probability for class c inside the batch.

879 Let's compute the gradient of \mathcal{L}_{reg} w.r.t. z_i :

$$\nabla_{z_i} \mathcal{L}_{reg} = \sum_{k=1}^C \nabla_{z_i} \bar{p}_k \log \bar{p}_k = \sum_{k=1}^C (1 + \log \bar{p}_k) \nabla_{z_i} \bar{p}_k$$

880 Therefore, we only need to compute $\nabla_{z_i} \bar{p}_k$:

$$\nabla_{z_i} \bar{p}_k = \nabla_{z_i} \frac{1}{N} \sum_{i=1}^N q_{ik} = \frac{1}{N} \nabla_{z_i} q_{ik} = \frac{1}{N} \nabla_{z_i} \frac{e^{z_i^\top z_t^k}}{\sum_{j=1}^C e^{z_i^\top z_t^j}} = \frac{1}{N} q_{ik} \sum_{j=1}^C q_{ij} [z_t^k - z_t^j]$$

881 Then we have:

$$\begin{aligned} \nabla_{z_i} \mathcal{L}_{reg} &= \frac{1}{N} \sum_{k=1}^C (1 + \log \bar{q}_k) q_{ik} \sum_{j=1}^C q_{ij} (z_t^k - z_t^j) \\ &= \frac{1}{N} \sum_{k=1}^C [(1 + \log \bar{q}_k) q_{ik} \sum_{j \neq k} q_{ij} - q_{ik} \sum_{j \neq k} q_{ij} (1 + \log \bar{q}_j)] z_t^k \\ &= \frac{1}{N} \sum_{k=1}^C [\sum_{j=1}^C q_{ij} \log \frac{\bar{q}_k}{\bar{q}_j}] q_{ik} z_t^k \end{aligned} \quad (17)$$

882 From Eq. (17), we observe that the gradient is influenced by the ratios $\log \frac{\bar{q}_k}{\bar{q}_j}$, driving it towards
883 the classes that are underrepresented in the batch predictions. The use of the regularization loss in
884 conjunction with our soft contrastive loss creates a powerful combined effect, enabling the effective
885 relabeling of misclassified examples, as discussed in Appendix B.

886 **Gradient of $\mathcal{L}_{CLIPTTA}$.** We recall from the main paper that the final CLIPTTA loss combines both
887 \mathcal{L}_{s-cont} and \mathcal{L}_{reg} , thus benefiting both from an enhanced adaptation loss as well a mechanism to
888 combat pseudo-labeling errors (we omit the effect of the memory for simplicity):

$$\mathcal{L}_{CLIPTTA} = \mathcal{L}_{s-cont} + \lambda_{reg} \mathcal{L}_{reg}, \quad (18)$$

889 Therefore the gradient of $\mathcal{L}_{CLIPTTA}$ writes:

$$\nabla_{z_i} \mathcal{L}_{CLIPTTA} = \sum_{j=1}^N \beta_{i,j} [-\hat{z}_t^j + \sum_{k=1}^C w_{k,i} z_t^k] + \lambda_{reg} \frac{1}{N} \sum_{k=1}^C [\sum_{j=1}^C q_{ij} \log \frac{\bar{q}_k}{\bar{q}_j}] q_{ik} z_t^k. \quad (19)$$

890 Depending on the composition of the batch, we can see that $\mathcal{L}_{CLIPTTA}$ will strongly benefit from the
891 contribution of the soft-contrastive loss to provide accurate adaptation, or will be able to correct
892 misclassified examples due to the positive interaction of the combined corrective terms in $\nabla_{z_i} \mathcal{L}_{s-cont}$
893 and $\nabla_{z_i} \mathcal{L}_{reg}$.

894 A.2 Analysis of OCE Loss

895 As discussed in the main paper, our outlier contrastive exposure (OCE) in Eq. (9) of the main paper is
896 a special case of the intra-class variance minimization:

$$\sigma^2 = p_{id} \frac{\sum_i^N w_i (s_i - \mu_{id})^2}{\sum_{i=1}^N w_i} + p_{ood} \frac{\sum_i^N (1 - w_i) (s_i - \mu_{ood})^2}{\sum_{i=1}^N (1 - w_i)} \quad (20)$$

with $p_{\text{id}} = \frac{1}{N} \sum_i w_i$ and $p_{\text{ood}} = \frac{1}{N} \sum_i (1 - w_i)$. This is further condensed into the loss function Eq. (21):

$$\sigma_{\text{intra}}^2 = p_{\text{id}} \mu_{\text{id}}^2 - p_{\text{ood}} \mu_{\text{ood}}^2 \quad (21)$$

Here, samples from the same distribution might tend to collapse into a single point. An alternative formulation is inter-class variance maximization, as shown in Eq. 22:

$$\sigma_{\text{inter}}^2 = p_{\text{id}} p_{\text{ood}} (\mu_{\text{id}} - \mu_{\text{ood}})^2 \quad (22)$$

The impact of the proportions p_{id} and p_{ood} is twofold. First, when neither ID nor OOD samples are detected, the respective proportion nullifies, and the inter-class variance reaches its minimum. On the contrary, an equilibrium can be reached with both $p_{\text{id}} = p_{\text{ood}} = 0.5$, which displays the implicit assumption of equally distributed scores between ID and OOD. We argue that this constraint limits the flexibility of the OOD detection at test time; as the nature of incoming samples is unknown, allowing for a non-uniform distribution in the detection can help filter out less useful samples. Secondly, the product of these probabilities would reduce the scale of the loss, especially compared to the other components of our CLIPTTA framework, which limits its impact on the adaptation of the model. Hence, a fully contrastive metric can attain the same detection objective by diminishing the latter negative effects:

$$\sigma_{\text{inter}}^2 = (\mu_{\text{id}} - \mu_{\text{ood}})^2 \quad (23)$$

B Discussion on CLIPTTA’s robustness

We further study the properties of CLIPTTA, to expand the insights on the working mechanisms that assist in its success. Initially, the accuracy across batches (see Fig. 1 in the main paper) serves as a straightforward depiction of (a) the general preeminence of CLIPTTA over other methods, particularly entropy-based techniques, and (b) the collapse effect in methods such as TENT. To elaborate on the underlying advantages of our method, we examine the adaptation process more closely, first in a controlled toy example, then using CIFAR-10-C across all of its corruptions.

Mitigating pseudo-label errors. In Fig. 5, we present a controlled toy example demonstrating how CLIPTTA effectively mitigates misclassifications. This example features a batch of six samples in a three-class classification problem. It focuses on the gradient orientations of the TENT, CLIPTTA, and regularized CLIPTTA losses for a single misclassified and ambiguous sample. The sample in question exhibits high probabilities for both the predicted and correct labels, indicating low confidence. The gradient of the CLIPTTA loss is directed toward the correct label, working to minimize the difference between the top two probabilities—a behavior further amplified by the regularized CLIPTTA loss. In contrast, TENT prioritizes increasing the highest probability, thereby reinforcing the incorrect prediction.

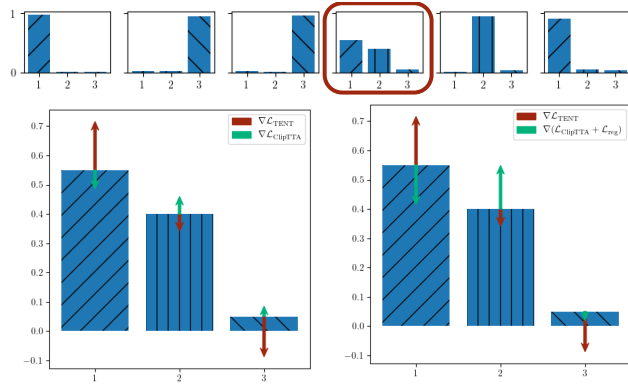


Figure 5: **Gradient Behavior: TENT vs. CLIPTTA** Illustration of gradient directions for TENT, CLIPTTA, and regularized CLIPTTA losses on a misclassified sample (circled in red). While TENT (red arrows) reinforces the incorrect prediction to reduce entropy, CLIPTTA and its regularized version (green arrows) aim to minimize top-2 probability differences, guiding the correction.

A similar behavior is observed with real images. In Fig. 8, we show misclassified ImageNet samples where the CLIPTTA loss pushes gradients toward the ground truth (GT) class rather than the predicted

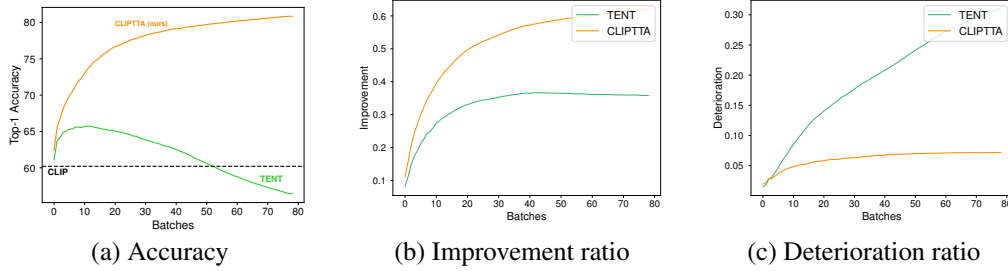


Figure 6: **Improvement & Deterioration ratios on CIFAR-10-C.** (a) While TENT’s accuracy collapses, CLIPTTA shows consistent improvement. (b) The improvement ratio quantifies the proportion of misclassified examples correctly relabeled after adaptation. (c) The deterioration ratio captures the proportion of correctly classified examples that become misclassified post-adaptation.

class. Notably, the difference between the gradient toward the GT label and the gradient toward the predicted label is consistently positive for TENT (favoring the incorrect prediction) and negative for CLIPTTA (counteracting the incorrect prediction).

We provide quantitative insights on the CIFAR-10-C dataset in Fig. 6. In Fig. 6-a, we observe the collapse of TENT, while CLIPTTA maintains robust performance. To further analyze this, we quantify two key metrics: the “improvement ratio” shown in Fig. 6-b, which represents the proportion of misclassified examples that are correctly classified after adaptation, and the “deterioration ratio” shown in Fig. 6-c, which denotes the proportion of correctly classified examples that become misclassified after adaptation. CLIPTTA outperforms TENT by achieving a higher improvement ratio and a lower deterioration ratio.

C Experimental details

We provide further information about the experimental setup that was conducted in the main paper. This includes the specifics of the experimental protocol, the baselines and benchmarks that were considered, as well as an extension of the empirical results.

C.1 Detailed experimental protocol

In our experiments, we follow the widely explored *non-episodic* TTA setting [11, 17, 28], in which the model is adapted continually to batches of data, without recovering its original weights. This poses a challenge, as adaptation risks of severely degrading the model, which can aggravate as adaptation goes longer. As some of the considered baselines were originally conceived for an *episodic* setting (e.g. CLIPArTT [6]), some conditioning was applied in order to amplify their performance in this scenario.

C.2 Details on baselines and datasets

Benchmarks. We provide more detailed information about the datasets that compose the different benchmarks used through the main paper. For all the experiments, images of different sizes were reshaped for compatibility with CLIP (i.e., to 224×224).

Natural images. We employed CIFAR-10 and CIFAR-100 [31], both containing 10,000 images of size 28×28 , and spanning 10 and 100 classes, respectively. We use Imagenet [32] as a larger-scale dataset, with 1000 classes and 50,000 images in total.

Corruptions. Transformed variants of the previous benchmarks are built by applying 15 different corruptions such as *gaussian noise*, *fog*, or *pixelate*. This results in CIFAR-10-C and CIFAR-100-C [33] and Imagenet-C. Each corruption is utilized in its highest severity level (e.g., *level 5*), yielding the most complex version of each dataset. The number of images in each corrupted set and their size correspond to the previous benchmark, which results in 45 different datasets to evaluate in total.

Dataset	Classes	Size	Category
Aircraft	100	3,333	Transportation
Caltech101	100	2,465	Objects
Cars	196	8,041	Transportation
DTD	47	1,692	Textures
EuroSat	10	8,100	Satellite
Flowers102	102	2,463	Flora
Food101	101	30,300	Food
Pets	37	3,669	Fauna
SUN397	397	19,850	Scenes
UCF101	101	3,783	Actions

Table 5: **Detailed information of the fine-grained classification benchmark.**

Dataset	Domain	Size	Dataset	Domain	Classes	Size
PACS	Photo	1,670	Imagenet	Natural	1,000	50,000
	Cartoon	2,344	Imagenet-V2	Natural	1,000	10,000
	Sketch	3,929	Imagenet-S	Sketch	1,000	50,000
	Art painting	2,048	Imagenet-R	Art	200	30,000
OfficeHome	Art	965	Imagenet-A	Adversarial	7,500	7,500
	Clipart	2,535				
	Product	2,470				
	Real world	1,495				

(a) PACS and OfficeHome

(b) Imagenet-Domains

Table 6: **Detailed dataset statistics.** (a) PACS and OfficeHome. (b) Imagenet-Domains.

Fine-grained classification datasets are a popular choice in zero-shot classification with CLIP, as they span a wide semantic variety in their classes. We utilize Imagenet as well 10 other datasets covering: Aircraft [34], Caltech101 [35], Cars [36], DTD [37], EuroSat [38], Flowers102 [39], Food101 [40], Pets [41], SUN397 [42], and UCF101 [43]. The specific details of each dataset are condensed in Table 5.

Domain generalization. This is a set of datasets popularly use in the context of Domain Adaptation. We use Visda-C [44], which includes 12 common classes and contains two main sets: a set of 152,397 3D renderings and a set of 55,388 of images cropped from MS COCO [45]. We also incorporate PACS [46], with seven classes, and OfficeHome [47] with 65 classes, which include images in four different styles, as summarized in Table 6a). Finally, we include the challenging Imagenet-Domains benchmark, involving four variants of Imagenet: Imagenet-V2 [48], Imagenet-R [49], Imagenet-S [50], Imagenet-A [51], each of which is detailed in Table 6b).

Out-of-distribution datasets. In our open-set TTA setup, each ID dataset in the natural and corrupted image benchmarks is paired with a corresponding OOD dataset. The classification task is performed only on ID samples, while OOD samples are solely used for detection (*i.e.*, recognizing and rejecting unknowns). Thus, OOD class labels are not meaningful in this context. Following prior work [28, 27], we use SVHN [52] (26,032 street view digit images) as the OOD set for CIFAR-10 and CIFAR-100, and Places365 [53] (1.8M scene images) for ImageNet. In the corrupted setting (*i.e.*, CIFAR-10/100-C and ImageNet-C), we use SVHN-C and Places365-C as OOD sources, matched by corruption type (*e.g.*, JPEG compression) and set to maximum severity.

Baselines. We group baselines into three categories based on their adaptation strategy. The first group includes entropy-based methods for standard classifiers such as TENT [11], ETA [12], SAR [13], RoTTA [14], OSTTA [17], SoTTA [26], STAMP [27], and UniEnt [28]. These methods typically operate by minimizing the conditional entropy of the model’s predictions and require adaptations to work with CLIP’s vision-language outputs. The second group comprises CLIP-specific methods such as CLIPArTT [6] and WATT [7], which modify the loss or prompt structure to better

leverage CLIP’s multimodal nature. The third group includes alternative CLIP-based adaptation approaches: TPT [3], which performs prompt tuning via entropy minimization, and TDA [4], which operates without gradients using a memory-based episodic scheme. All baselines are implemented following their respective publications. For CLIP-based methods, minimal changes were needed to integrate into our framework. For non-CLIP methods, we use CLIP’s image-to-text similarities (as defined in Eq.1, Sec.3) as classification logits. Entropy-based baselines directly apply their loss to these logits. Hyperparameter details are provided below when applicable.

- ETA: a similarity threshold of $\epsilon = 1$ and an entropy threshold $\alpha = 0.4$ are used. These are kept for all cases.
- SAR: an entropy threshold $\alpha = 0.4$ and an exponential moving average (EMA) weight $m = 0.2$ are used for all cases. The SAM optimizer is employed.
- RoTTA: we use a timeliness weight $\lambda_t = 1$ and an uncertainty weight $\lambda_u = 1$, a memory capacity equivalent to the batch size. These are kept for all cases.
- TDA: we use the same values used for Imagenet in the original paper. We employ $\alpha_{\text{pos}} = 2.0$, $\beta_{\text{pos}} = 2.0$, $\alpha_{\text{neg}} = 0.117$, $\beta_{\text{neg}} = 1.0$, entropy thresholds $H_o = \{0.2, 0.5\}$, entropy masks $M_o = \{0.03, 1.0\}$, and positive and negative shot capacities of 2 and 3, respectively.
- CLIPArTT: we take $K = 3$ most probable classes in all datasets, except for $K = 5$ in VisDA-C, which uses a learning rate of 1×10^{-5} .
- WATT: we use two adaptation iterations per text prompt, and two meta-repetitions are used. A learning rate of 1×10^{-5} is used for VisDA-C.
- SoTTA: we use the confidence threshold $\tau = 1/|\mathcal{C}|$, with \mathcal{C} the number of classes. The memory capacity is equal to the batch size. The SAM optimizer is employed.
- UniEnt: we use $\lambda_{\text{reg}} = 1$ and $\lambda_{\text{ood}} = 1$.

C.3 Extended experimental results

Adapting the text encoder. CLIPTTA is evaluated across a diverse set of datasets by adapting not only the visual encoder but also the text encoder, as shown in Table 7. Updating the text encoder proves beneficial in many cases, particularly for semantically complex datasets where CLIP’s pre-trained embeddings may lack sufficient specialization. This is evident in datasets focused on fine-grained classification, such as SUN397 and OxfordPets, where incorporating text encoder updates yields notable improvements. However, updating the text encoder can sometimes have detrimental effects, especially on datasets containing general or well-represented concepts, such as EuroSat. Despite being visually challenging, the broad and commonly encountered class labels in such datasets may already be adequately represented in CLIP’s original text embeddings. In these cases, further adaptation of the text encoder may disrupt this alignment, leading to performance degradation. This behavior underscores the importance of selectively adapting the text encoder based on the semantic complexity of the dataset.

Dataset	CLIPTTA (Vision only)	CLIPTTA (Vision + Text)
CIFAR-10	95.0	93.5 (-1.5)
CIFAR-100	74.9	75.0 (+0.1)
ImageNet	69.1	69.6 (+0.5)
ImageNet-V2	62.7	63.1 (+0.4)
ImageNet-A	54.0	54.2 (+0.2)
ImageNet-R	80.1	79.9 (-0.2)
ImageNet-S	50.8	51.2 (+0.4)
Aircraft	26.5	26.9 (+0.4)
Caltech101	94.2	94.4 (+0.2)
Cars	66.7	67.1 (+0.4)
DTD	46.5	48.1 (+1.6)
EuroSat	80.3	72.9 (-7.4)
Flowers102	71.3	71.7 (+0.4)
Food101	86.7	86.8 (+0.1)
OxfordPets	91.6	92.4 (+0.8)
SUN397	65.2	67.5 (+2.5)
UCF101	69.3	70.3 (+1.0)
Median	69.2	70.3 (+1.1)

Table 7: **Impact of updating the text encoder.**

Moreover, the results highlight the trade-off between generalization and specialization when jointly adapting both encoders. While semantically complex datasets benefit from increased specialization, datasets with simpler or well-represented class concepts risk losing the robust generalization capabilities inherent to CLIP’s pre-trained representations. This suggests that a targeted or dataset-specific strategy for adapting the text encoder may be more effective in leveraging its potential.

Open-set TTA on corrupted datasets. Table 9 reports results in the challenging open-set setting under corruption shifts. This scenario is challenging because models must adapt to noisy in-distribution samples while maintaining robustness to unseen OOD classes. As previously observed, TENT is highly unstable in these settings, suffering from severe model collapse that is exacerbated by corrupted inputs. Its accuracy drops to 2.1% on ImageNet-C and 10.6% on CIFAR-100-C, with poor OOD detection (FPR95 above 95%), confirming its sensitivity to pseudo-label noise.

In contrast, CLIPTTA with the OCE loss maintains high performance across all benchmarks, achieving the best overall results on both accuracy and OOD detection. On average, on the corrupted datasets, it improves over UniEnt by +5.8 points in accuracy and reduces FPR95 by nearly 20 points. These gains demonstrate the benefit of aligning the adaptation objective with CLIP’s pre-training loss while integrating an explicit OOD detection signal. The results confirm that CLIPTTA is well-suited for open-set test-time adaptation, even under strong distribution shifts such as corruptions.

	CIFAR-10			CIFAR-100			ImageNet			Average		
	ACC↑	AUC↑	FPR95↓	ACC↑	AUC↑	FPR95↓	ACC↑	AUC↑	FPR95↓	ACC↑	AUC↑	FPR95↓
CLIP	89.3	98.5	5.2	68.1	86.8	83.5	66.7	90.1	43.8	74.7	91.8	44.2
TENT [11]	93.0	42.3	89.3	69.1	36.2	94.8	12.4	49.9	89.4	58.2	42.8	91.2
OSTTA [17]	90.9	60.5	72.8	70.9	43.3	93.8	66.9	84.9	59.2	76.2	62.9	75.3
SoTTA [26]	89.5	98.5	4.9	68.9	88.5	76.3	66.7	89.3	47.1	75.0	92.1	42.8
STAMP [27]	89.9	98.6	5.5	67.5	87.7	80.0	29.7	63.0	80.2	62.4	83.1	55.2
UniEnt [28]	<u>94.2</u>	99.9	0.0	<u>72.7</u>	<u>97.8</u>	<u>8.7</u>	65.2	95.4	17.1	<u>77.3</u>	<u>97.7</u>	<u>8.6</u>
CLIPTTA + OCE (ours)	94.6	<u>99.8</u>	<u>0.4</u>	74.9	98.4	7.6	67.6	97.7	9.7	79.0	98.6	5.9

Table 8: **Open-set TTA results.** Top-1 accuracy with ViT-B/16 backbone on the open-set setting.

	CIFAR-10-C			CIFAR-100-C			Imagenet-C			Average		
	ACC↑	AUC↑	FPR95↓	ACC↑	AUC↑	FPR95↓	ACC↑	AUC↑	FPR95↓	ACC↑	AUC↑	FPR95↓
CLIP	60.2	88.0	58.0	35.2	67.0	93.8	24.6	68.6	89.2	40.0	74.5	80.3
TENT [11]	26.9	50.7	91.2	10.6	43.1	95.0	2.1	48.0	95.0	13.2	47.3	93.7
OSTTA [17]	62.7	53.4	85.2	34.5	36.3	92.6	<u>31.2</u>	<u>75.1</u>	79.8	42.8	54.9	85.9
STAMP [27]	60.4	88.0	57.1	34.5	66.8	93.8	9.0	53.8	92.9	34.6	69.5	81.3
UniEnt [28]	<u>78.7</u>	<u>98.6</u>	5.7	<u>48.9</u>	<u>91.0</u>	<u>31.0</u>	23.6	44.8	90.8	<u>50.4</u>	<u>78.1</u>	<u>42.5</u>
CLIPTTA + OCE (ours)	79.1	98.7	<u>6.1</u>	50.4	96.7	19.2	39.0	89.0	43.2	56.2	94.8	22.8

Table 9: **Open-set TTA results on Corrupted Datasets.** Top-1 accuracy with ViT-B/16 backbone on the open-set setting.

Domain shifts benchmarks. We provide the extended results of the different domain shifts (Table 1), including Imagenet-Domains (Table 10), VisDA-C (Table 11), OfficeHome (Table 12), and PACS (Table 13). CLIPTTA achieves the best performance across all of these datasets on average, demonstrating great flexibility across domains. Our method also obtains highly competitive results independently in each sub-dataset.

	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Average
CLIP	66.7	47.8	60.8	74.0	47.8	59.4
TPT (NeurIPS ’22)	69.0	<u>54.8</u>	63.5	77.1	47.9	62.4
TDA (CVPR ’24)	<u>69.5</u>	60.1	64.7	80.2	<u>50.5</u>	65.0
TENT (ICLR ’21)	66.5	51.3	60.2	79.4	43.7	60.2
ETA (ICML ’22)	67.4	49.2	60.9	75.3	46.8	59.9
SAR (ICLR ’23)	66.7	51.5	60.5	79.6	44.6	60.6
RoTTA (CVPR ’23)	68.4	51.2	62.5	78.1	47.8	61.6
CLIPArTT (WACV ’25)	67.6	50.7	61.2	76.2	47.9	60.7
WATT (NeurIPS ’24)	69.0	51.1	62.5	78.1	48.2	61.8
CLIPTTA (ours)	69.6	54.0	62.7	80.2	50.8	<u>63.4</u>

Table 10: **Detailed results on the Imagenet-Domains benchmark.**

	Synthetic 3D	MS COCO	Average
CLIP	87.2	86.7	87.0
TPT [3]	85.5	84.5	85.0
TDA [4]	86.6	86.5	86.5
TENT [11]	93.2	85.3	<u>89.3</u>
ETA [12]	91.1	85.4	88.3
SAR [13]	88.1	87.5	87.8
RoTTA [14]	80.6	86.7	83.7
CLIPArTT [6]	82.2	86.0	84.1
WATT [7]	88.4	<u>87.0</u>	87.7
CLIPTTA (ours)	<u>92.2</u>	86.9	89.6

Table 11: Detailed results on the two domains of the Visda-C dataset.

	Art	Clipart	Product	Real	Average
CLIP	83.2	68.0	89.1	89.8	82.5
TPT [3]	82.5	66.3	88.5	89.2	81.7
TDA [4]	83.2	68.8	89.8	90.4	83.0
TENT [11]	84.1	68.8	90.0	90.5	83.4
ETA [12]	84.3	70.8	<u>90.4</u>	<u>90.7</u>	<u>84.1</u>
SAR [13]	84.4	70.9	89.6	90.3	83.8
RoTTA [14]	82.9	68.0	89.1	89.8	82.5
CLIPArTT [6]	82.6	68.4	87.6	89.6	82.0
WATT [7]	83.8	69.0	90.0	90.5	83.4
CLIPTTA (ours)	<u>84.2</u>	<u>70.7</u>	91.0	91.0	84.2

Table 12: Detailed results on the four domains of the OfficeHome (OH) dataset.

	Photo	Art	Cartoon	Sketch	Average
CLIP	99.9	97.4	99.1	88.1	96.1
TPT [3]	99.5	95.3	93.9	87.2	94.0
TDA [4]	99.9	97.5	98.9	88.1	96.1
TENT [11]	99.8	98.0	<u>99.2</u>	89.1	96.6
ETA [12]	99.8	97.9	99.3	89.8	<u>96.7</u>
SAR [13]	99.9	97.5	99.1	88.2	96.2
RoTTA [14]	99.9	93.8	98.8	88.1	95.8
CLIPArTT [6]	99.5	96.9	98.3	90.4	96.2
WATT [7]	99.9	<u>97.6</u>	<u>99.2</u>	88.4	96.2
CLIPTTA (ours)	99.9	98.0	99.3	92.0	97.5

Table 13: Detailed results on the four domains of the PACS dataset.

Semantic datasets. We report closed-set adaptation results on both coarse- and fine-grained classification tasks in Tables 14 and 15. On coarse-grained benchmarks (CIFAR-10 and CIFAR-100), CLIPTTA achieves the highest accuracy on both datasets, with a strong average of 85.2%, outperforming all TENT-based and CLIP-based methods, including CLIPArTT and WATT. Notably, it improves over TENT by +0.2 points on CIFAR-10 and +2.4 points on CIFAR-100, and remains significantly ahead of zero-shot CLIP (+6.5 points on average). On fine-grained datasets, CLIPTTA consistently ranks among the top methods, achieving the best average accuracy across the 11 datasets (69.8%). Despite its simplicity, it performs favorably compared to more complex CLIP-specific methods such as TPT, TDA, and CLIPArTT, which rely on prompt tuning or heuristic loss modifications. CLIPTTA performs particularly well on datasets like EuroSAT (+22.3 over CLIPArTT) and OxfordPets (+4.5 over TDA) while maintaining competitive results on the others. These findings highlight the ro-

	CIFAR-10	CIFAR-100	Average
CLIP	89.3	68.1	78.7
TPT [3]	89.8	67.4	78.6
TDA [4]	91.4	69.8	80.6
TENT [11]	94.9	72.9	83.9
ETA [12]	94.8	<u>73.7</u>	<u>84.3</u>
SAR [13]	92.1	73.2	82.7
RoTTA [14]	89.4	68.5	79.0
CLIPArTT [6]	88.4	73.2	80.8
WATT [7]	92.5	70.8	81.7
CLIPTTA (ours)	95.1	75.3	85.2

Table 14: **Closed-set TTA on coarse-grained datasets.** Top-1 accuracy with ViT-B/16 backbone on coarse-grained datasets (CIFAR-10 and CIFAR-100).

	<i>ImageNet</i>	<i>Aircraft</i>	<i>Caltech101</i>	<i>Cars</i>	<i>DTD</i>	<i>EuroSAT</i>	<i>Flowers102</i>	<i>Food101</i>	<i>OxfordPets</i>	<i>SUN397</i>	<i>UCF101</i>	<i>Average</i>
CLIP	66.7	24.8	92.2	65.5	44.1	48.3	70.7	84.8	88.4	62.3	64.7	64.7
TPT [3]	69.0	<u>24.8</u>	94.2	<u>66.9</u>	47.8	42.4	69.0	84.7	87.8	65.5	68.0	65.6
TDA [4]	<u>69.5</u>	23.9	94.2	67.3	<u>47.4</u>	<u>58.0</u>	71.4	86.1	88.6	67.6	70.7	<u>67.7</u>
TENT [11]	66.5	15.5	93.8	63.0	43.1	58.4	<u>71.3</u>	86.5	89.5	63.1	68.0	65.3
ETA [12]	67.4	24.8	93.0	65.2	44.4	47.5	71.4	85.9	89.2	63.6	66.6	65.4
SAR [13]	66.7	21.9	93.9	64.0	43.9	50.2	70.9	<u>86.5</u>	89.6	63.3	67.7	65.3
RoTTA [14]	68.4	22.3	94.0	58.1	45.2	24.2	70.5	81.6	87.0	64.9	66.8	62.1
CLIPArTT [6]	67.5	24.0	92.7	64.0	43.4	46.7	67.0	84.2	87.1	64.2	67.0	64.4
WATT [7]	69.0	23.6	<u>94.1</u>	65.8	44.7	40.0	71.4	<u>86.2</u>	88.7	<u>66.3</u>	68.2	65.3
CLIPTTA (ours)	69.6	26.5	94.2	66.7	46.5	80.3	<u>71.3</u>	86.7	91.6	65.2	<u>69.3</u>	69.8

Table 15: **Closed-set TTA on fine-grained datasets.** Top-1 accuracy comparison of CLIPTTA against other TTA methods on a suite of 11 fine-grained datasets.

1083 bustness of our adaptation objective across both coarse- and fine-grained tasks without requiring
1084 task-specific tuning or architectural modifications.

1085 **Statistical significance.** We con-
1086 duct additional runs of CLIPTTA to as-
1087 sess its sensitivity to random initializa-
1088 tion, reporting the mean accuracy and
1089 95% confidence interval in Tab. 16. The
1090 results indicate that CLIPTTA is
1091 highly stable, with very low variance across independent runs. The tight confidence intervals (e.g.,
1092 ± 0.01 on ImageNet) confirm the reliability and reproducibility of the observed performance gains,
1093 further supporting the robustness of the method across different datasets.

	CIFAR-10	CIFAR-100	Imagenet
CLIPTTA	94.9 ± 0.03	75.3 ± 0.07	69.1 ± 0.01

Table 16: Accuracy of CLIPTTA averaged over three random initializations (mean \pm 95% CI).

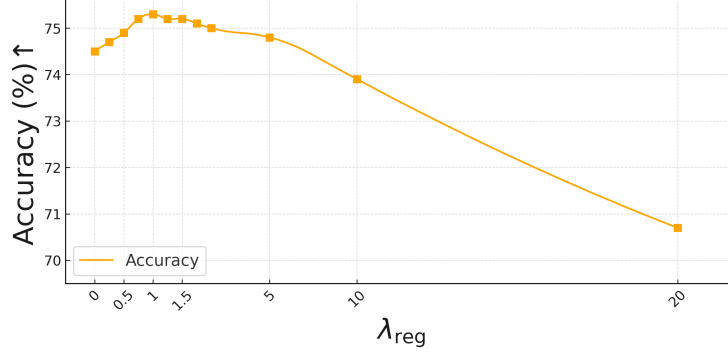


Figure 7: **Impact of λ_{reg} on CIFAR-100.** Effect of λ_{reg} on the closed-set accuracy of CLIPTTA when evaluated on CIFAR-100.

C.4 Hyperparameter analysis.

In this section, we evaluate the sensitivity of CLIPTTA to its key hyperparameters: the regularization weight λ_{reg} , the OOD loss weight λ_{oce} , and the adaptation batch size. Results show that CLIPTTA remains robust across a wide range of values, requiring minimal tuning for strong performance.

Effect of λ_{reg} . Figure 7 shows the impact of the regularization weight λ_{reg} on CIFAR-100 accuracy. We observe that CLIPTTA is remarkably stable for values ranging from 0.5 to 2.0, with accuracy consistently above 75% in this range. Performance peaks around $\lambda_{reg} = 1$, which we use as the default. Beyond that, accuracy gradually declines, indicating that overly strong regularization may suppress beneficial updates. Overall, this confirms that CLIPTTA does not require precise tuning of λ_{reg} to perform well and that a wide range of values yields near-optimal performance.

Effect of λ_{oce} . Table 17 reports the impact of λ_{oce} on ImageNet in the open-set setting. While accuracy stays stable for small values, OOD detection improves substantially: AUROC increases from 93.5% (no OCE) to 97.7% at $\lambda_{oce} = 1$, and FPR95 drops by 16 points. Performance remains robust in the range [0.25–2], confirming the stability of the OCE loss.

λ_{oce}	0	0.25	0.5	1	2	5	10	20	100
Acc	67.6	67.6	67.6	67.6	67.5	67.3	66.4	64.5	56.6
AUC	93.5	97.5	97.6	97.7	97.8	98.0	98.4	98.8	99.2
FPR	25.7	10.1	9.8	9.7	8.8	7.8	6.3	4.7	2.3

Table 17: **Impact of λ_{oce} on Imagenet.** Effect of λ_{oce} on accuracy and open-set metrics AUROC (AUC) and false positive rate (FPR).

Effect of batch size. Table 18 presents accuracy on CIFAR-10 for batch sizes ranging from 1 to 512. Accuracy increases with batch size and saturates around 64 samples, showing that CLIPTTA benefits from richer batch-level statistics. Remarkably, even in the extreme case of a single image per batch, CLIPTTA remains competitive (93.4% accuracy). This is made possible by the confident memory, which stores reliable past predictions and enables the use of our soft contrastive loss even when no other images are available in the current batch. As a result, CLIPTTA is well-suited for deployment in streaming where batch sizes may be small.

Batch size	1	2	8	32	64	128	256	512
Accuracy	93.4	94.7	94.7	94.8	95.0	95.1	95.1	95.2

Table 18: **Accuracy on different batch sizes on the CIFAR-10 dataset.** Although CLIPTTA benefits from larger batches, it remains competitive even in the extreme case of 1 image adaptation.



$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -1.3e$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$



$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -1.5e^{-2}$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$



$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -3.0e$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$



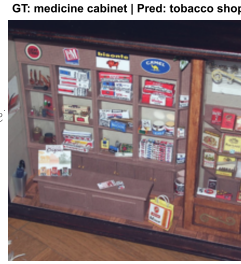
$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -6.3e^{-3}$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$



$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -5.3e$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$



$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -6.5e^{-3}$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$



$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -3.9e$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$



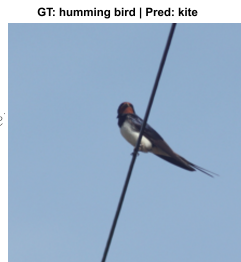
$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -4.3e^{-3}$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$



$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -1.6e$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$



$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}}) = -2.6e^{-3}$$

$$\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}}) > 0$$

Figure 8: **CLIPTTA allows to correct misclassified examples on Imagenet** We show examples whose individual gradients of our loss points toward the ground truth label while the gradient of TENT always points towards the predicted class, which accentuates the pseudo-labeling error. The difference between the gradients towards GT and predictions is $\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{ours}})$ for CLIPTTA, whereas for TENT it corresponds to $\Delta_{\text{GT-Pred}}(\nabla \mathcal{L}_{\text{TENT}})$.