Segmentation Assisted Incremental Test Time Adaptation in an Open World

Manogna Sreenivas
manognas@iisc.ac.in
Soma Biswas
somabiswas@iisc.ac.in

Department of Electrical Engineering,
Indian Institute of Science,
Bengaluru, India

Abstract

In dynamic environments, unfamiliar objects and distribution shifts are often encountered which challenge the generalization abilities of the deployed trained models. This work addresses Incremental Test Time Adaptation (ITTA) of Vision-Language Models (VLMs), tackling scenarios where unseen classes and unseen domains continuously appear during testing. Unlike traditional Test Time Adaptation approaches where the test samples come only from a predefined set of classes, our framework allows models to adapt simultaneously to both covariate and label shifts, actively incorporating new classes as they emerge. Towards this goal, we establish a new benchmark for ITTA, integrating single-image TTA methods for VLMs with active labeling to query an oracle for samples potentially representing unseen classes during test time. We propose a segmentation assisted active labeling module, termed SegAssist, which is training-free and repurposes the VLM's segmentation capabilities to refine active sample selection, prioritizing samples likely to belong to unseen classes. Extensive experiments on several benchmark datasets demonstrate the potential of SegAssist to enhance the performance of VLMs in real-world scenarios, where continuous adaptation to emerging data is essential.

1 Introduction

Consider a model deployed in real world, navigating a dynamic environment where it encounters familiar objects in unfamiliar settings and new objects it has never seen before. For instance, a model trained to recognize urban entities like cars, trucks, and pedestrians might suddenly come across unfamiliar elements, such as bicycles or animals, when it transitions to a rural landscape. In these situations, the model must not only adapt to the changed appearance of known objects, affected by varying factors like lighting and weather, but also correctly identify unfamiliar objects as new classes instead of misclassifying them. Moreover, it must seamlessly integrate this new information to remain accurate and relevant in real-time. This is the underlying motivation of *Incremental Test Time Adaptation (ITTA)*, where a deployed model must continuously adapt to changes in data distribution and incorporate emerging classes in an online manner. Traditional TTA methods [2, 25, 28, 30] aim to recognize known classes in unseen domains, enabling models to adapt with batches of test data. Recently, VLMs such as CLIP [20], have shown excellent zero-shot generalization abilities due to their pre-training on web-scale image-text data, making them promising candidates for the challenging single-image TTA scenario [12, 23, 24].

In this work, we propose a more realistic Incremental TTA setting, since data in the open-world can come not only from unseen distributions, but from unseen classes as well. Our protocol reimagines TTA as an evolving learning process, suited for real-world applications where environments are constantly changing. This is done by actively querying an oracle for labels of uncertain samples to discover and incorporate new classes, thereby continuously expanding its knowledge in an online setting. Towards this goal, we propose a novel active sample selection module termed SegAssist, which leverages pixel-wise segmentation to determine whether an uncertain sample should be chosen for active labeling. This ensures that we utilize the limited active labeling budget judiciously, primarily for labeling samples of new classes. *SegAssist is a training-free plug-in module which can be seamlessly integrated with existing single-sample TTA frameworks* to adapt them to the ITTA scenario. Our key contributions are summarized as follows:

- We address the real-world challenging task of Incremental Test Time Adaptation (ITTA) in an open-world using Vision-Language Models (VLMs), enabling it to adapt to both covariate and label shift during test time.
- We establish strong baselines using different active labeling strategies and state-of-the-art CLIP-based TTA methods, paving the way for future research in this direction.
- We propose a novel training-free active labeling module, named SegAssist, which dynamically queries and incorporates new classes, supporting class recognition in a continuously expanding set of classes in an online setting. SegAssist can be seamlessly plugged in with existing TTA methods to equip them for ITTA.
- We propose a novel metric termed Incremental Class Detection Delay (ICDD) to quantify an ITTA method in terms of the timeliness of detecting new classes.
- We demonstrate the effectiveness of our method through extensive experiments on several benchmark datasets, showcasing its robustness in handling both covariate and label shifts during test time.

2 Related Works

Here, we describe some of the related works in literature.
Test Time Adaptation: TTA methods aim to adapt a model trained on one domain to test data from another domain without requiring access to the original training data. Early TTA approaches [1, 3, 17, 22, 25, 28] focus on adapting models to handle covariate shifts, such as changes in lighting, weather conditions and other common corruptions. These methods typically assume access to a batch of test samples at a time for continuous adaptation. More recently, single-sample TTA based on pre-trained vision-language models like CLIP [20] has been explored by approaches such as TPT [23], TDA [12], and DPE [33]. However, these methods address only covariate shifts and are not equipped to handle label shifts during testing, as in the realistic ITTA scenario which is of our interest.
Active Labeling (AL): AL methods aim to efficiently query informative samples to reduce the need for extensive labeling, making it particularly valuable in resource-constrained

scenarios [27]. Traditional AL methods [13, 19, 32] operate in an offline setting, typically requiring access to large, fixed datasets, and are applied during model training. Recently, AL has been explored in the TTA context, as in ATTA [5], which queries samples during test time. However, [5] assumes that test samples arrive in batches and belong to a fixed set of classes, limiting their applicability in single image ITTA scenario. In contrast to these works which query uncertain samples from the fixed set of classes, in this work, we propose to leverage active labeling to query uncertain samples during single-image TTA to enable real-time incorporation of new classes as they appear. While many existing AL strategies rely on abundant data or batch processing [5, 13, 19, 32] to compute selection metrics, we explore selection criteria that can be calculated on a per-sample basis, suitable for single image TTA.

Open World and Incremental Learning: Learning to recognize new classes in an open-world is a challenging problem, which has been studied from several perspectives. Out of Distribution (OOD) detection methods [6, 7, 29] aim to distinguish between seen and unseen class samples, typically formulated as a $(C_s + 1)$-way classification problem, where C_s represents the seen classes and a $+1$ class is reserved for the unknown classes. Traditional OOD approaches rely on access to abundant data and are generally applied during training [4, 6, 11, 26, 29], although recent studies have explored OOD detection in a batch-wise TTA setting [14, 15]. In contrast, incremental learning [10, 16, 21] address the challenge of adding new classes over time without forgetting previously learned classes. However, these methods assume access to training data for new classes and typically require retraining to incorporate new information, making them unsuitable for real-time TTA scenarios.

3 Incremental Test Time Adaptation

Training	TTA	Covariate shift	Label shift	Single image	Classification task	Methods
✓	✓	✓	✗	✗	C_s	TENT[28], CoTTA[30], ROID[17]
✗	✓	✓	✗	✓	C_s	TPT[23], TDA[12], DPE[33]
✗	✓	✓	✓	✗	$C_s + 1$	Open world TTA [14, 15]
✓	✗	✗	✓	✗	C_s+C_u	Class Incremental Learning [10, 16, 21]
✗	✓	✓	✓	✓	C_s+C_u	Proposed ITTA framework

Table 1: Comparison of the proposed ITTA framework with existing research directions.

Motivation. The need for addressing Incremental Test Time Adaptation (ITTA) stems from the growing demand for models that can handle dynamically changing, real-world environments, where assumptions of a fixed distribution between training and test data rarely hold. Existing approaches have attempted to address isolated aspects of these challenges, yet they often fall short in real-world scenarios such as autonomous systems, surveillance, etc. which demand that models adapt to changes in both the nature of input data (covariate shift) and the introduction of entirely new object categories (label shift). We first describe the existing closely related research directions (Table 1), motivating the need for the proposed ITTA framework.

- **TTA with only Covariate Shift:** Traditional TTA methods like TENT [28], CoTTA [30], ROID [17] focus on adapting models trained on clean data to handle changes in data distribution during test-time without changing the underlying classes. *These methods cannot accommodate dynamically emerging new classes.*

- **Single-Image TTA with Covariate Shift:** With the advent of pretrained VLMs, methods like TPT [23], TDA [12], DPE [33] address single-image TTA, which is closer to a real-time scenario, but with fixed label space. *Thus, they also cannot handle test samples from unseen classes outside the predefined classes of interest* C_s.

- **Open-World TTA with Covariate and Label Shift:** To handle unknown classes, open-world TTA methods [14, 15] propose to do (C_s+1)-way classification, where C_s represents the seen classes and an additional class is reserved for all unknown classes. *These methods do not support incremental learning of incoming new classes.*

- **Class Incremental Learning (CIL):** Traditional CIL techniques [16, 21] aim to incorporate new classes by assuming access to training data for these new classes enabling continual learning. *Hence, they are ill-suited to dynamic TTA settings where only one unlabeled test sample is available at a time and immediate adaptation is desired.*

These works address isolated aspects of real world scenarios, but they fall short in scenarios that combine multiple, complex conditions like covariate shift, label shift, and incremental learning. Our work combines these considerations by establishing a novel *Incremental Test Time Adaptation (ITTA)* protocol reflecting the online nature of real-world deployment. ITTA requires the model to not only adapt to covariate shifts, but also incrementally learn new classes without relying on additional training data. It enables the model to: (1) Continuously adapt to covariate shifts. (2) Recognize and incorporate new classes as they appear. (3) Operate on a single-image basis without access to batches or any labeled data. (4) Maintain performance across seen classes and a growing set of dynamically incorporated unseen classes. To address the need for continuous class expansion, we leverage active learning to query samples where the model exhibits uncertainty. When a queried sample belongs to a previously unseen class, we dynamically add this class to the list of seen classes, thereby enabling ongoing classification across both previously seen and emerging new categories. This framework enables efficient adaptation in highly variable, real-world environments.

3.1 Problem Formulation

In this work, we address Incremental Test-Time Adaptation (ITTA) of VLMs in an open-world, where both covariate and label shifts can occur. During test-time, the model encounters a stream of test samples $\{x_t\}_{t=1,2,\ldots}$, where each sample x_t may either belong to the seen classes $C_s = \{c_1, c_2, \ldots c_S\}$ or a previously unseen class. In the context of VLM, seen classes indicate the set of classes expected to be encountered in that scenario and for which labels are provided for classification. The task is to predict a label y_t for each incoming test sample x_t while also adapting to distributional shifts and dynamically incorporating new classes as they emerge. This is done by integrating active labeling with TTA (Figure. 1).

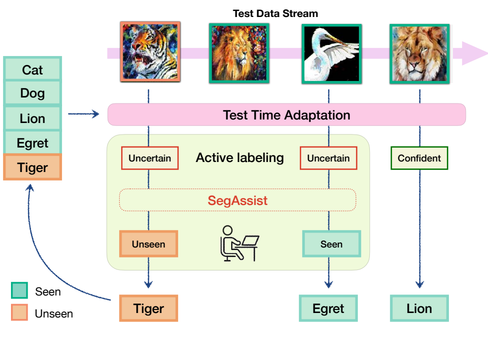

Figure 1: ITTA setup handles both covariate and label shift via oracle labeling setup.

We use CLIP [20] consisting of an image encoder ϕ_I and text encoder ϕ_T. Given the initial set of classes $C^0 = C_s$, the text encoder ϕ_T is used to obtain the text classifiers by embedding text prompts of the form “A photo of a $\{c_i\}$" for each class $c_i \in C^0$. For an incoming sample x_t at time t, the model adapts itself and then predicts the label $y_t \in C^t$, where C^t is the set of seen classes till time t. However, when the model encounters an uncertain sample, possibly a sample that does not belong to the set of seen classes $\mathcal{C}^t$, an active learning query is triggered. The goal is to label this uncertain sample and determine if it belongs to a new class. If the queried sample x_t is confirmed to belong to a new, previously unseen class u, the model updates the set of known classes by adding the new class, i.e., $\mathcal{C}^{t+1} \leftarrow \mathcal{C}^t \cup \{u\}$ as illustrated in Fig. 1. Further, the model continues to classify subsequent samples from the expanded set of classes $\mathcal{C}^{t+1}$. The adaptation and prediction continues at each time step.

Budget for ITTA. Prior active labeling works define the annotation budget as a percentage of the dataset size or a fixed number of samples that can be queried for labeling. Here, as the test samples are streamed in an online manner, we define a replenishing budget $\mathcal{B}_t$ more suited for an online scenario where we do not know the number of test samples apriori. $\mathcal{B}_t$ is set based on a budget rate r (analogous to the percentage of data in offline active labeling) which is replenished after seeing every N samples. This budget controls the number of samples that can be queried for labeling. We set the budget rate r to 1% and N to 1000, replenishing the budget by 10 queries every 1000 test samples seen.

Towards the goal of ITTA, we establish a strong benchmark, choosing VLM based single-image TTA methods such as ZS-Eval [20], TDA [12], DPE [33]. We integrate these methods with active learning strategies to perform ITTA. We now describe our proposed method SegAssist, a simple plug-in module for active labeling, designed to favour selecting unseen class samples during test time.

4 Proposed Framework

In Incremental Test-Time Adaptation (ITTA), selecting informative samples for active labeling is critical in budget-constrained scenarios where querying each uncertain sample may not be feasible. If a test sample’s confidence score is below a predefined threshold, the oracle is queried for active labeling. However, with limited budget, it becomes crucial to refine this pool further, to select samples that are likely to belong to unseen classes. Towards this goal, we analyse the CLIP scores and propose a novel module termed *SegAssist* to further refine the samples obtained using existing active labeling strategies like confidence, entropy etc.

Can CLIP scores identify New Classes?
Analyzing the Maximum Softmax Probability (MSP) scores obtained by matching CLIP’s global image features with text features, from Fig. 2, we observe that the confidence scores for known and unknown class samples exhibit substantial overlap. This makes it challenging to reliably distinguish unseen from seen classes based on scores like MSP alone, potentially leading to inefficient use of the available budget $\mathcal{B}_t$ as seen class samples may be selected more frequently than unseen class samples. Here, we explore whether fine-grained, local image features can be leveraged to improve the ability to discriminate between seen and unseen classes for better active sample selection.

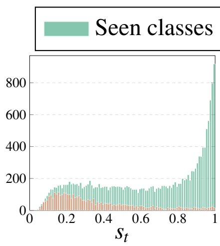

Figure 2: MSP scores for Painting

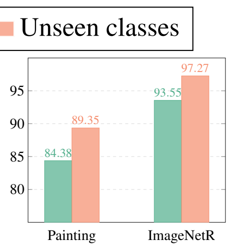

Figure 3: % of background images.

To investigate this hypothesis, we perform pixel-wise segmentation for all uncertain samples, aiming to identify distinctive patterns between samples from seen vs. unseen classes. Using CLIP, we perform dense predictions on the image [34] by matching the patch features with the top-K predictions and a "background" class. For seen class samples, we expect the model to correctly classify the pixels into one of the top-K classes predicted, while samples from unseen classes should primarily register as "background" across most pixels. While this approach may not be perfectly accurate, as foreground pixels from unseen classes could map to related seen classes within the top-5 predictions, it can still aid in distinguishing seen and unseen class samples. Our analysis from Fig. 3 reveals that, among the uncertain samples, the number of background images (majority of pixels classified as background) is more for unseen class samples compared to uncertain seen class samples.

Based on this insight, we propose SegAssist to effectively select samples for active labeling. Only uncertain samples with over 95% of pixels classified as "background" are selected. Thus, SegAssist aims to refine the sample selection process, enhancing the budget utilization by prioritizing those test samples that are most likely to represent new, unseen classes. *SegAssist is a simple and effective plug-in module, requiring no additional model training. With no computation overhead, it can be seamlessly integrated with any existing TTA and active labeling method to adapt them to the realistic ITTA setting.*

4.1 SegAssist for Active Sample Selection

The first step of active sample selection is to identify if an incoming sample $x_t \in \mathcal{X}$ is uncertain. We can utilize any standard measure like low confidence, high entropy, etc. to identify uncertain samples. By default, we use the Maximum Softmax Probability (MSP) calculated as $s_t = \max_c p_c(x_t)$ to identify uncertain samples. Here, $p_c(x_t)$ is the softmax score predicted for class c. We consider x_t as *uncertain* if: $s_t < \tau$, where τ is the uncertainty threshold. Further, we refine this selection through the proposed SegAssist module, which consists of two key steps: (i) Generating pixel-wise segmentation maps, (ii) Selecting samples for active learning based on background pixel proportion, as shown in Fig. 4.

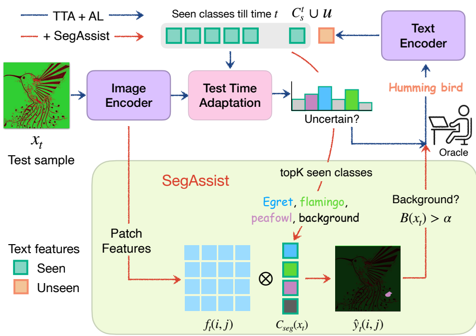

Figure 4: SegAssist: A plug-in module which can be integrated with existing TTA and active labeling methods. It prioritizes unseen class samples by performing pixel wise segmentation of uncertain samples, selecting it for active labeling only if the pixels are predominantly classified as "background".

(i) Pixel-Wise Segmentation of Uncertain Samples. For each uncertain sample x_t, let $topK(x_t)$ denote the top-K class names as predicted by the model. We define the set $C_{\text{seg}}(x_t) = topK(x_t) \cup \{$"background"$\}$ as the classes of interest to perform pixel-wise segmentation. Though global image feature, which is the embedding of the [CLS] token is used for classification in CLIP, we also have access to the patch features, which can be leveraged to get dense predictions. Following [31, 34], we obtain the patch features using the value vectors from the last self attention layer followed by the final projection layer. We $f_t(i,j)$ denote the feature of patch (i,j) of image x_t. Its prediction is obtained as

$$\hat{y}_t(i,j) = \arg\max_{c \in C_{\text{seg}}(x_t)} \langle f_t(i,j), t_c \rangle \quad (1)$$

where t_c is the text embedding of class $c \in C_{seg}(x_t)$. These predictions at the patch level are upsampled using bilinear interpolation to obtain the segmentation map $S(x_t)$.

(ii) Background Ratio for Active Sample Selection. To assess if an uncertain sample x_t is more likely to belong to an unseen class, we compute the proportion of pixels in $S(x_t)$ classified as "background". We denote $|S(x_t)|$ as the total number of pixels in $S(x_t)$. The background ratio $B(x_t)$ is then computed as:

$$B(x_t) = \frac{1}{|S(x_t)|} \sum_{i=1}^{|S(x_t)|} \mathbb{I}[S_i(x_t) = \text{"background"}], \tag{2}$$

where $\mathbb{I}[\cdot]$ is an indicator function, returning 1 if $S_i(x_t) =$ background and 0 otherwise.

We hypothesize that a high proportion of background pixels indicate that x_t belongs to an unknown class, as samples from known classes are expected to contain foreground predictions that align with seen classes C_s. Thus, we filter uncertain samples for active labeling based on their background ratio. Specifically, we select x_t for querying only if $B(x_t) > \alpha$ where α is the background threshold, set to 0.95. This ensures that samples with a significant proportion of pixels classified as "background" are prioritized for active labeling.

In essence, *SegAssist refines the selection process in ITTA, leveraging pixel-wise segmentation maps to prioritize samples that are genuinely novel, making the budget utilization and adaptation process more effective* (Fig. 4). By leveraging local features, SegAssist goes beyond traditional active labeling approaches to improve sample selection in open-world settings, where incorporating new classes dynamically is integral to the performance of the model. It can be integrated with any VLM based TTA framework, with practically no computation overhead. *As the patch features are already computed, the only additional computation is the comparison of patch features with C_{seg}, thus making it a very light weight module.*

4.2 Incremental Class Detection Delay (ICDD)

In an open-world, where new classes appear dynamically with time, a key challenge for ITTA approaches is detecting these emerging classes as they are introduced. Towards this objective, we design *Incremental Class Detection Delay (ICDD)*, a metric to quantify the performance of an ITTA system in identifying new classes in a timely manner. It compares the rate of new class introductions in the *ground truth sequence* with the rate of class detections by the ITTA framework (bottom plot in Fig. 5), visualized as a normalized Area Under the Curve (AUC) plot. For each normalized time step t, we track two cumulative functions:

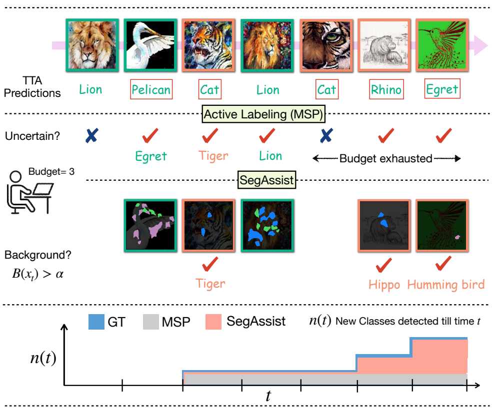

Figure 5: Given a budget of 3 samples, active selection based on MSP selects 3 samples, of which two are seen and only one is unseen.

(1) $n_{\text{gt}}(t)$: the cumulative count of classes newly introduced till time t.

(2) $n_{\text{det}}(t)$: the cumulative count of classes detected by the ITTA system till time t.

Both $n_{\text{gt}}(t)$ and $n_{\text{det}}(t)$ are normalized between 0 and 1 with respect to the total number of new classes. Similarly, time t is normalized to the interval $[0,1]$, representing the fraction

	Method	ImageNet-R		ImageNet-A		DN-Clipart		DN-Painting	
		HM↑	ICDD ↓	HM↑	ICDD ↓	HM↑	ICDD ↓	HM↑	ICDD ↓
ZS-Eval	Random	49.57	25.90	23.60	45.33	22.04	75.30	32.54	68.41
	Entropy	59.89	21.99	21.11	43.83	48.25	61.88	63.34	43.89
	MSP	62.92	17.21	**24.59**	**41.47**	47.65	62.04	70.59	39.14
	SegAssist	**64.05**	**16.01**	23.94	41.89	**49.08**	**61.20**	**71.21**	**38.56**
TDA	Random	53.33	25.90	16.58	44.33	22.54	75.30	32.66	68.41
	Entropy	64.06	19.52	18.78	44.95	58.79	49.10	67.27	46.08
	MSP	70.38	13.25	19.27	43.93	58.99	50.68	68.83	**39.31**
	SegAssist	**70.48**	**13.07**	20.02	**43.93**	**60.25**	**48.96**	**69.20**	41.27
DPE	Random	54.90	25.90	28.52	46.21	23.36	75.30	33.49	68.41
	Entropy	0.19	56.05	4.44	57.88	15.50	80.27	00.00	81.81
	MSP	68.09	19.64	24.62	47.18	65.35	49.00	71.45	34.70
	SegAssist	**69.39**	**16.38**	**33.42**	**45.98**	**65.49**	**49.00**	**72.40**	**33.12**

Table 2: ITTA Results: ImageNet-R, ImageNet-A, DN-Clipart, DN-Painting

of the total test stream duration. This yields two cumulative curves spanning the area from (0,0) to (1,1), where $n_{\text{det}}(t)$ lags behind $n_{\text{gt}}(t)$ when there is a delay in the detection of an unseen class. The ICDD score is defined as the difference in AUC between these two curves

$$\text{ICDD} = \text{AUC}(n_{\text{gt}}(t)) - \text{AUC}(n_{\text{det}}(t)) \tag{3}$$

ICDD measures the delay in detecting new classes relative to their actual introduction in the test stream. A lower ICDD score suggests that the ITTA system is highly responsive, detecting new classes almost immediately as they appear. Conversely, a higher score indicates a greater delay in identifying unseen classes, thereby leading to a decrease in the performance. We discuss the complementary nature of ICDD and accuracy metrics in the Supplementary.

5 Experimental Evaluation

Here, we describe the baseline methods, datasets and implementation details, experiments and analysis performed to evaluate the effectiveness of the proposed framework.

5.1 Baseline methods

Baseline VLM-based Single Image TTA Methods. In our work, we plug-in SegAssist with Zero-Shot Evaluation (ZS-Eval) and state-of-the-art TTA approaches TDA [12], DPE [33] to demonstrate its effectiveness.

1) ZSEval [20]: A test image x_t is classified by matching its image feature $f_t = \phi_I(x_t)$ with text features $\phi_T(c)$ as $\hat{y}_t = \arg\max_{c \in C^t} \langle \phi_I(x_t), \phi_T(c) \rangle$, where C^t refers to the classes seen till time t in ITTA scenario.

2) TDA [12] (CVPR'24): They employ a training-free dynamic adapter to perform TTA of VLMs through a lightweight key-value cache to efficiently refine pseudo labels.

3) DPE [33] (NeurIPS'24): They perform TTA by effectively accumulating task-specific knowledge from both textual and visual modalities through a Dual Prototype Evolving (DPE) approach. They also optimize learnable residuals to align prototypes from both modalities.

Baseline Active Labeling Methods. To enable a model deployed in real world to discover new classes, we propose to use Active Labeling (AL) to identify test samples to be queried by an oracle. Most recent AL techniques require abundant data [5, 13, 19, 32] to identify active samples, which may not be suitable for single-sample ITTA scenario which is of our interest. Consequently, we employ AL methods that enable querying on a per-sample basis, which we now describe.
1. Random: This is the simplest AL baseline, in which a test sample is randomly selected for querying with probability r. Here, r is the predefined budget rate defining the budget B_t as described in Section 3.1.
2. Maximum Softmax Probability (MSP): This method selects samples based on the MSP score $s_t = \max_{c \in C^t} p_c$ predicted by the model. Here, p_c is the predicted probability for class c. Samples are queried if their MSP score falls below a threshold, indicating low confidence.
3. Entropy: The uncertainty of model predictions can be measured using the entropy measure $H(x_t) = -\sum_{c \in C^t} p_c \log p_c$ and is frequently used to identify samples for active labeling. Test samples with entropy above a predefined threshold are selected for querying.

5.2 Datasets and Implementation Details.

We use the following large-scale benchmark datasets for evaluation. **ImageNet-R** [8] is a realistic style transfer dataset encompassing interpretations of 200 ImageNet classes. **ImageNet-A** [9] is an adversarial style transfer dataset, containing of 7,500 images belonging to 200 categories. We perform experiments on **Clipart** and **Painting** domains from the **DomainNet** [18] dataset and use CLIP ViT-B16 [20]. We split the total number of classes in each dataset into seen and unseen classes, in the ratio of 4:1 to create the test stream. We use a budget rate r as 1% and the budget $\mathcal{B}_t$ is replenished with 10 samples for every 1000 test samples. We set the active labeling threshold for MSP and Entropy as 0.2 and 0.5 and respectively. For SegAssist, we consider the top 5 classes predicted by the model and "background" to perform pixelwise segmentation. We use CLIP ViT-B16 [20] model for all our experiments. For baseline TTA methods, we use the same hyperparameters used in their works [12, 33]. We run all experiments on an NVIDIA A6000 graphics card.

5.3 Evaluation Metrics

We evaluate the performance of different methods based on the following metrics:
(i) Harmonic Mean (HM) of accuracy on seen and unseen classes: We first calculate the accuracy on seen classes and unseen classes. To equally weigh the performance of both seen and unseen classes, we calculate the harmonic mean of these two accuracy values.
(ii) Incremental Class Detection Delay (ICDD): We report the proposed ICDD metric (in % area) specifically designed for open-world scenarios. ICDD measures the delay in discovering new classes, as described in Section 4.2.

5.4 Experimental Results

Comparison with baseline TTA and AL methods. In Table 2, we observe that the performance of SegAssist is better than prior baselines in most cases (across TTA and AL baselines), only at times second to MSP. Although SegAssist tries to judiciously use the budget for unseen classes coming later in time, unseen classes semantically related to seen classes

Budget	Method	DN-Painting	ImageNet-R
0.5%	ZSEval*	62.00 / 50.32	52.12 / 26.21
	+SegAssist	**63.13 / 50.28**	**52.58 / 25.54**
1.0%	ZSEval*	70.59 / 39.14	62.92 / 17.21
	+SegAssist	**71.21 / 38.56**	**64.05 / 16.01**
1.5%	ZSEval*	73.82 / 32.56	68.02 / 13.55
	+SegAssist	**75.98 / 29.93**	**68.98 / 12.02**

Table 3: Varying budget rates r (HM $\uparrow$ / ICDD $\downarrow$).

Thresh.	Method	DN-Painting	ImageNet-R
0.1	ZSEval*	59.92 / 49.64	52.14 / 29.39
	+SegAssist	**60.82 / 48.35**	**52.14 / 29.39**
0.2	ZSEval*	70.59 / 39.14	62.92 / 17.21
	+SegAssist	**71.21 / 38.56**	**64.05 / 16.01**
0.3	ZSEval*	66.83 / 45.48	62.97 / 13.47
	+SegAssist	**69.18 / 43.12**	**63.91 / 13.25**

Table 4: Varying threshold (HM $\uparrow$ / ICDD $\downarrow$).

Method	0.1	0.25	0.5	0.75	1.0
ZSEval*	62.30	70.59	64.38	60.21	60.50
+SegAssist	**70.38**	**71.21**	**65.12**	**60.55**	**61.10**
TDA*	64.34	68.83	64.96	61.17	60.28
+SegAssist	**64.74**	**69.20**	**66.78**	**63.21**	**62.35**
DPE*	63.32	71.45	64.00	60.82	60.31
+SegAssist	**66.24**	**72.40**	**64.56**	**61.34**	**61.20**

Table 5: Varying ratio of unseen to seen (HM$\uparrow$).

can have background ratio $B(x_t) < \alpha$. Here, MSP would select the test sample for active labeling while SegAssist would not, leading to delayed class detection in some cases.

5.5 Analysis

Varying Budget rate. We vary the budget rate as 0.5%, 1.0% and 1.5% to study the effect of the budget rate (Table 3) on the performance of our method. We observe that the performance of our method consistently improves with respect to the baseline AL method MSP (represented by *) across all budget rates. A higher budget rate naturally allows the model to query more samples, thereby enabling it to discover new classes faster, improving the overall performance.

Varying uncertainty threshold. From Table 4, we observe that, across all thresholds chosen, SegAssist consistently outperforms MSP. This is because SegAssist uses pixel-wise segmentation to identify samples which predominantly have background pixels, relatively increasing the probability of selecting unseen class sample for querying.

Varying ratio of unknown to known classes. We evaluate our module across test scenarios with varying unknown-to-known class ratios: 0.1, 0.25, 0.5, 0.75, and 1.0 on DN-Painting. From Table 5, we observe SegAssist consistently outperforms MSP across all ratios.

In addition, we demonstrate the plugin nature of SegAssist, the effect of parameter K and also analyse the complementary nature of ICDD and accuracy in Supplementary material.

6 Conclusion

Towards the goal of enabling VLMs to continuously adapt in real-world scenarios often characterized by both covariate and label shifts, we introduce and setup a benchmark for Incremental Test-Time Adaptation (ITTA). We leverage VLMs with active labeling to achieve this. We propose *SegAssist*, a simple plug-in module which can be seamlessly integrated with current baseline TTA methods, to prioritize active labeling of unseen class samples over seen classes. Through extensive experiments, our results demonstrate that the proposed ITTA framework, enhanced by SegAssist, advances the adaptability of VLMs in dynamic, real-world settings. We hope this work will pave the way for further research in adapting VLM-based systems in an open-world.

Future research directions. In this work, actively labeled samples are only used to incorporate new classes. These samples could further enhance model performance through finetuning or to guide diverse sample selection for active labeling. Developing advanced active labeling methods and specialized ITTA algorithms is an important direction for future research.

Acknowledgements. The first author was supported by Prime Minister's Research Fellowship awarded by the Ministry of Education, Govt. of India.

References

[1] Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. A simple signal for domain shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023.

[2] Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. Santa: Source anchoring network and target alignment for continual test time adaptation. *Transactions on Machine Learning Research*, 2023.

[3] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[4] Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations*, volume 1, page 5, 2022.

[5] Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. *arXiv preprint arXiv:2404.05094*, 2024.

[6] Haodong, Dong Lu, Shuo Gong, Jason Wang, and Kristen Moore Xue, Lina Yao. Learning with mixture of prototypes for out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[8] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out1-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[9] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[10] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019.

[11] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021.

[12] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.

[13] K Krishna and M Narasimha Murty. Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3):433–439, 1999.

[14] Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[15] Yushu Li, Xun Xu, Yongyi Su, and Kui Jia. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[16] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[17] Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proceedings of the IEEE/CVF International Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[18] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019.

[19] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8505–8514, 2021.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 2021.

[21] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[22] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Neural Information Processing Systems*, 2020.

[23] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Neural Information Processing Systems*, 2022.

[24] Manogna Sreenivas and Soma Biswas. Efficient open set single image test time adaptation of vision language models. *Transactions on Machine Learning Research*, 2025.

[25] Manogna Sreenivas, Goirik Chakrabarty, and Soma Biswas. pstarc: Pseudo source guided target clustering for fully test-time adaptation. In *Proceedings of the IEEE/CVF International Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[26] Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. *arXiv preprint arXiv:2303.02966*, 2023.

[27] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014.

[28] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

[29] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[30] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[31] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. *ECCV*, 2024.

[32] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8708–8716, 2022.

[33] Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. *arXiv preprint arXiv:2410.12790*, 2024.

[34] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022.