TEST-TIME GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-Time Adaptation (TTA) and Generalized Category Discovery (GCD) are traditionally treated as disjoint problems: the former adapts models to domain shift assuming all test classes are known, while the latter discovers novel categories assuming labeled training data for known classes. However, real-world deployment rarely fits either setting. Motivated by this gap, we introduce **Test-Time Generalized Category Discovery (TT-GCD)**, a unified and more realistic scenario where a vision-language model must adapt to distribution shifts, classify known categories using only textual supervision, and discover novel categories, all during test time and without access to labeled data. To address this challenging scenario, we propose **PACT**(**P**rototype **A**ssignment for **C**ategory discovery at **T**est time), a fully unsupervised framework that casts known-class recognition and novel-class discovery via prototype assignment. PACT first re-aligns shifted visual features with the text-derived class representations of the VLM using confident zero-shot predictions. Known and novel categories are then both represented by prototypes in the visual embedding space, estimated from the unlabeled test stream, and each test image is assigned to the category whose prototype is most similar to its visual feature. Extensive experiments across corruption and domain-shift benchmarks demonstrate that PACT outperforms adapted state-of-the-art TTA and GCD methods, effectively bridging the gap between adaptation and discovery.

1 INTRODUCTION

Vision-language models (VLMs) such as CLIP (Radford et al., 2021) have emerged as a powerful paradigm for visual recognition, enabling generalization across a wide range of downstream vision tasks (Mishra et al., 2026; Dong et al., 2025; Wang et al., 2024; Caselli et al., 2026; Schrodi et al., 2025). Yet real-world deployment poses a compound challenge as the visual distribution may shift while previously unseen categories are simultaneously encountered. Consider an autonomous driving system trained on urban streets in New York (see Fig. 1[*Left*]) and deployed to mountain highways of Banff, Alberta (see Fig. 1[*Middle*]). Familiar object categories such as sedans, pedestrians, and traffic signs now appear under unfamiliar conditions of snow glare, fog, and unpaved road surfaces. Simultaneously, the system encounters entirely new object categories it was never trained on, including wildlife such as moose, deer, and bears on the roadway and recreational vehicles absent from its urban training set. The deployed model must adapt to these domain shifts, maintain reliable recognition of known categories and also group unknown objects into coherent categories, all at inference time, without labeled data or access to the original training set.

Existing paradigms (Fig. 1[*Right*]) address only parts of this problem. Test-Time Adaptation (TTA) methods Wang et al. (2021); Mishra et al. (2026) handle distribution shifts by updating model parameters on unlabeled test data, but they operate under a *closed label set* assumption, leaving novel-class samples to be silently misclassified into known categories. Generalized Category Discovery (GCD) approaches Vaze et al. (2022); Wen et al. (2023); Chiaroni et al. (2023); Wang et al. (2024) discover novel categories alongside known ones, but require *labeled training data* for the known classes and assume *no distribution shift* between training and inference. HiLo Wang et al. (2025) extends GCD to handle domain shifts, yet it still requires *training with cross-domain labeled data*, a resource unavailable in deployment. Additionally, Open-set TTA Gao et al. (2024); Lee et al. (2023) detects out-of-distribution samples during adaptation but *rejects* them rather than organizing them into semantically coherent clusters. Thus, these settings fail to jointly address and perform well in distribution shift, known-class recognition and novel-category discovery simultaneously under fully unsupervised test-time conditions (refer Fig. 2).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

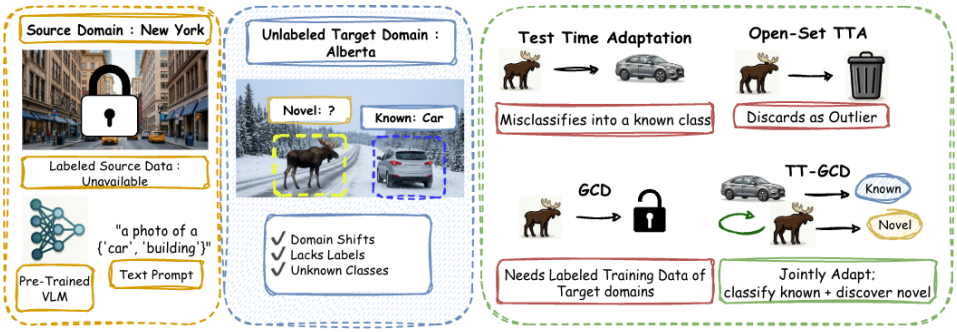


Figure 1: **The TT-GCD setting.** A VLM trained on New York streets, with no labeled source data [*Left*], is deployed on an unlabeled Alberta stream [*Middle*] that combines domain shift (snow), known classes (`car`) and novel classes (`moose`). [*Right*] TTA misclassifies the novel class as a known class, Open-Set TTA discards it as an outlier, GCD needs labeled target-domain data, and TT-GCD adapts to the shift, classifies the known and discovers the novel.

VLMs provide a natural foundation for bridging this gap. Their vision-language alignment enables known categories to be recognized directly from textual descriptions alone, eliminating the labeled source data required by conventional GCD. However, an off-the-shelf VLM remains vulnerable to distribution shift and provides no direct mechanism for discovering categories absent from its text prompts. We therefore formulate **Test-Time Generalized Category Discovery (TT-GCD)**, where a pretrained VLM is given only the text descriptions of known categories and an unlabeled test stream containing both known and novel categories under distribution shift. The goal is to simultaneously (1) adapt to distribution shifts, (2) classify known categories using textual supervision, and (3) discover and cluster novel categories, all at test time without any labeled data, source dataset access, or prior knowledge of how many novel categories exist.

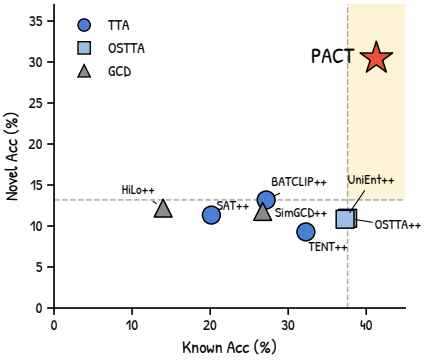


Figure 2: Mean performance (**Known vs. novel**) over CIFAR-10-C, CIFAR-100-C and ImageNet-C. Dashed lines mark the best baseline on each axis.

To address TT-GCD, we propose **PACT** (**P**rototype **A**ssignment for **C**ategory discovery at **T**est time), a fully unsupervised framework with two complementary stages. First, PACT re-aligns shifted visual representations with the fixed text space by adapting only the LayerNorm affine parameters of the image encoder, guided by confidence-weighted zero-shot predictions. Second, to avoid the modality mismatch between text representations for known classes and visual representations for novel ones (Liang et al., 2022), PACT represents both as prototypes in the adapted visual space. Known and novel categories are then handled through a unified nearest-prototype assignment, with prototypes updated online as the test stream arrives. This yields a simple streaming framework in which encoder re-alignment is performed once, while subsequent adaptation operates only on category prototypes.

- We introduce **Test-Time Generalized Category Discovery (TT-GCD)**, a setting in which a pretrained VLM must perform domain adaptation, known-class recognition from text, and novel-category discovery entirely at inference time without labeled data.
- We propose **PACT**, a fully unsupervised framework that couples lightweight feature re-alignment with online prototype estimation, enabling a unified nearest-prototype rule to jointly recognize known categories and discover novel ones.
- We establish a unified TT-GCD evaluation protocol and adapt representative TTA, open-set TTA, and GCD methods as competitive baselines from relevant literature. Extensive experiments on five benchmarks spanning synthetic corruptions and natural domain shifts demonstrate consistent improvements over these baselines.

2 Related Works

Test-Time Adaptation (TTA). Early methods recalibrate normalization statistics (Nado et al., 2021; Schneider et al., 2020) or minimize prediction entropy through normalization-layer updates (Wang et al., 2021), with later work adding sample filtering (Niu et al., 2022) and sharpness-aware optimization (Niu et al., 2023). With the advent of VLMs, prompt-based (Shu et al., 2022; Feng et al., 2023) and cache-based (Karmanov et al., 2024; Zhang et al., 2024) approaches emerged, alongside transport and alignment-based methods such as SAT (Mishra et al., 2026) and BATCLIP (Maharana et al., 2025). All share one assumption, that the test stream contains only known categories. Novel samples are silently absorbed into known classes, a failure that closed-set evaluation cannot see.

Open-Set TTA. A second line of work detects out-of-distribution samples during adaptation, by entropy separation (Gao et al., 2024), confidence filtering (Lee et al., 2023) or, for VLMs, a threshold on image-text similarity (Sreenivas and Biswas, 2025), and rejects them at inference. AEO (Dong et al., 2025) extends the entropy route to multimodal inputs. In every case the unknown is treated as noise. OWTTT (Li et al., 2023) is closest to our mechanism as it grows a prototype pool for out-of-distribution samples, but reports them as a single rejection class. It needs a source-trained model with source feature statistics. TT-GCD instead requires that novel samples be organized into categories, with no textual supervision.

Generalized Category Discovery (GCD) (Vaze et al., 2022) jointly classifies known and discovers novel categories from unlabeled data. Methods such as SimGCD (Wen et al., 2023), UNO (Fini et al., 2021), Meta-GCD (Wu et al., 2023), PIM (Chiaroni et al., 2023) and PromptCAL (Zhang et al., 2023) combine supervised learning on labeled known-class data with self-labeling, self-distillation or mutual-information objectives. VLM-based extensions (Ouldnoughi et al., 2023; Wang et al., 2024; Caselli et al., 2026) use CLIP's embedding space for known-novel interaction. All assume that labeled and unlabeled data share a domain, except HiLo (Wang et al., 2025), which handles domain shift but needs to fine-tune the backbone on labeled source data. Every GCD method therefore depends on labeled known-class data, without which known-class boundaries degrade and clustering fails under shift.

The modality gap in VLMs. Despite contrastive alignment, CLIP's image and text embeddings occupy two separate, narrow regions of the shared space, a phenomenon known as the modality gap (Liang et al., 2022). It arises from the cone effect at initialization, survives contrastive training (Liang et al., 2022) and concentrates in a few embedding dimensions (Schrodi et al., 2025). Closing it post hoc can degrade downstream performance (Liang et al., 2022; Schrodi et al., 2025), and closing it by retraining (Eslami and de Melo, 2025) is not available at test time, which is why PACT assigns among image-space prototypes (Sec. 4.2).

Test-Time Category Discovery. On-the-fly Category Discovery (Du et al., 2023) and PHE (Zheng et al., 2024) label streaming samples of known and unknown classes with hash-based descriptors, but both train on labeled known-class data and assume no distribution shift. TTD (Lyu et al., 2025) discovers classes without training, hashing the features of a DINO backbone (Caron et al., 2021) fine-tuned on labeled known classes, but it targets class shift alone and its purely visual hashing has no counterpart for the text anchors a VLM provides.

3 Test-Time Generalized Category Discovery (TT-GCD)

In this work, we introduce the task of **Test-Time Generalized Category Discovery (TT-GCD)**, which aligns with the practical demands of real-world deployment. In this setting, we use only a pretrained VLM and an unlabeled test set, to assign semantic labels to *all* test samples covering both known and novel categories while simultaneously adapting to distribution shift.

Formally, we assume access to a pretrained VLM (e.g., CLIP (Radford et al., 2021)) consisting of an image encoder $f_I : \mathcal{X} \to \mathbb{R}^d$ and a text encoder $f_T : \mathcal{T} \to \mathbb{R}^d$. At inference, the model receives an unlabeled test set $\mathcal{D}_{\text{test}} = \{\mathbf{x}_n\}_{n=1}^N$, where each sample is observed once. The test stream contains samples from known categories $\mathcal{C}_{\text{known}}$ and novel categories $\mathcal{C}_{\text{novel}}$, with $\mathcal{C}_{\text{known}} \cap \mathcal{C}_{\text{novel}} = \emptyset$ and is also subjected to distribution shift.

We denote the ℓ_2-normalized visual embedding of test image $\mathbf{x}_n$ by $\mathbf{v}_n = f_I(\mathbf{x}_n) \in \mathbb{R}^d$, and the ℓ_2-normalized text embedding of known class k by $\mathbf{z}_k = f_T(t_k) \in \mathbb{R}^d$, where t_k is a text prompt for

class k, e.g., "A PHOTO OF A [CLASS$_k$]". Following standard CLIP-based TTA settings (Shu et al., 2022; Maharana et al., 2025; Mishra et al., 2026), the names of the $K = |\mathcal{C}_{\text{known}}|$ known categories are available at inference, providing semantic grounding through their text embeddings. In contrast, novel categories have no textual descriptions and must be discovered directly from the unlabeled test stream. The objective of TT-GCD is to assign every test sample either to one of the K known categories or to one of the $L = |\mathcal{C}_{\text{novel}}|$ novel categories, while adapting to the distribution shift. Also, to better match real-world deployment, L is unknown at test time and must be inferred from the stream. For fair comparison with baseline methods, the main protocol supplies it to every method for comparability, and Sec. 5.3 reports results when it is estimated from the stream. TT-GCD therefore requires two capabilities: (1) *known-category classification*, using the available class descriptions, and (2) *novel-category discovery*, by grouping novel samples into semantically coherent categories. Both must be performed under distribution shift, using only the pretrained VLM and the unlabeled test stream. As summarized in Table 1, TT-GCD unifies domain adaptation, known-category recognition and novel-category discovery within a fully unsupervised test-time setting.

Table 1: Comparison of TT-GCD with related test-time and category-discovery settings. TT-GCD uniquely combines distribution shift, novel categories, novel-category discovery, and fully unlabeled test-time adaptation.

Setting	Domain Shift	Novel Classes	Novel Discovery	No Labeled Train Data
TTA (Wang et al., 2021; Shu et al., 2022; Maharana et al., 2025; Mishra et al., 2026)	✓	✗	✗	✓
Open-Set TTA (Lee et al., 2023; Gao et al., 2024)	✓	✓	✗	✓
GCD (Vaze et al., 2022; Chiaroni et al., 2023; Wang et al., 2024)	✗	✓	✓	✗
GCD w/ Shift (Wang et al., 2025)	✓	✓	✓	✗
TT-GCD (Ours)	✓	✓	✓	✓

4 PROTOTYPE ASSIGNMENT FOR CATEGORY DISCOVERY AT TEST TIME (PACT)

PACT decouples TT-GCD into two problems, each solved by minimizing an explicit objective. **(1) Feature Re-Alignment.** At test time, distribution shift primarily affects the visual input, while the text encoder continues to receive the same known class-name prompts. As a result, the shifted visual embeddings may become misaligned with their corresponding text embeddings. We first re-align the visual representations with the fixed text features by adapting only the LayerNorm affine parameters of the image encoder, providing a lightweight mechanism for test-time adaptation (Schneider et al., 2020; Wang et al., 2021; Maharana et al., 2025) (Sec. 4.1). **(2) Unified Category Discovery.** In the re-aligned space, known-class recognition and novel-category discovery are formulated jointly through prototype assignment on the unit hypersphere, with both known and novel categories represented by prototypes in the visual embedding space (Sec. 4.2).

4.1 FEATURE RE-ALIGNMENT

Initial buffer. Let θ_I denote the LayerNorm affine parameters of the image encoder with pretrained values θ_I^0. We write the corresponding visual embedding as $\mathbf{v}_n^{\theta_I}$, while the text embeddings $\{\mathbf{z}_k\}_{k=1}^K$ remain fixed. The zero-shot probability over the K known classes is

$$p_n^{\theta_I}(k) = \frac{\exp\left(\mathbf{v}_n^{\theta_I\top}\mathbf{z}_k/\tau\right)}{\sum_{j=1}^K \exp\left(\mathbf{v}_n^{\theta_I\top}\mathbf{z}_j/\tau\right)}, \tag{1}$$

where τ is the temperature. Before adaptation, we collect the first W test samples in an initial buffer $\mathcal{B}$. For each $\mathbf{x}_n \in \mathcal{B}$, the pretrained image encoder θ_I^0 together with the fixed text embeddings provides a pseudo-label $\hat{y}_n$ and confidence weight ω_n:

$$\hat{y}_n = \arg\max_k p_n^{\theta_I^0}(k), \qquad \omega_n = p_n^{(1)} - p_n^{(2)} \in [0,1], \tag{2}$$

where $p_n^{(1)}$ and $p_n^{(2)}$ denote the largest and second-largest zero-shot probabilities under θ_I^0, respectively. The resulting $\hat{y}_n$ and ω_n are computed once and remain fixed throughout re-alignment.

Re-alignment objective. We re-align the shifted visual representations with the fixed text space by optimizing only the LayerNorm affine parameters θ_I of the image encoder. Specifically,

$$\theta_I^\star = \arg\min_{\theta_I} -\frac{1}{\sum_{n\in\mathcal{B}}\omega_n}\sum_{n\in\mathcal{B}}\omega_n \log p_n^{\theta_I}(\hat{y}_n). \tag{3}$$

Weighting samples by ω_n emphasizes predictions that are more reliable under the initial model. Expanding the cross-entropy term gives,

$$-\log p_n^{\theta_I}(\hat{y}_n) = \underbrace{-\frac{1}{\tau}{\mathbf{v}_n^{\theta_I}}^\top \mathbf{z}_{\hat{y}_n}}_{\text{alignment}} + \underbrace{\log\sum_{j=1}^{K}\exp\left({\mathbf{v}_n^{\theta_I}}^\top \mathbf{z}_j/\tau\right)}_{\text{normalization}}. \tag{4}$$

The first term pulls each visual embedding toward the text embedding of its assigned class, while the second accounts for the competing known classes. Together, they increase the relative alignment with the assigned class, re-aligning the shifted visual representations with the initial text space. Importantly, while θ_I is optimized, the text embeddings $\{\mathbf{z}_k\}$, pseudo-labels $\hat{y}_n$, and confidence weights ω_n remain fixed. The fixed text embeddings provide semantic anchors, while confidence-weighted pseudo-labels reduce the influence of unreliable predictions from the evolving visual encoder mitigating confirmation bias (Arazo et al., 2020).

4.2 Unified Category Discovery

After re-alignment, we freeze $\theta_I^\star$ and denote the resulting adapted visual embeddings by $\mathbf{v}_n^\star$. All subsequent category assignments are performed in this adapted space.

Unified feature space. Although CLIP aligns image and text representations, a modality gap remains between the two embedding spaces, leading to systematically different image–text and image–image similarities (Liang et al., 2022). Directly comparing text embeddings for known classes against visual prototypes for novel classes can therefore bias category assignment toward one modality. We avoid this by representing both known and novel categories as visual prototypes in the unified space.

Since the adapted visual features are ℓ_2-normalized, they lie on the unit hypersphere. We model them as a mixture of von Mises–Fisher (vMF) components with shared concentration, one per category (Banerjee et al., 2005; Mardia and Jupp, 2000). Under hard assignments and uniform priors, maximum-likelihood estimation of this model is equivalent to spherical k-means (Banerjee et al., 2005). This motivates our prototype construction: known prototypes are estimated from confidence-weighted zero-shot assignments, while novel prototypes are initialized and subsequently refined from their assigned visual samples.

Known prototypes. For each known class k, we aggregate the adapted embeddings in the initial buffer of those samples that receive predicted class $\hat{y}_n = k$, weighted by their zero-shot confidence ω_n, and project the resulting vector onto the unit sphere:

$$\bar{\boldsymbol{\mu}}_k = \sum_{\substack{n\in\mathcal{B}\\ \hat{y}_n=k}} \omega_n\, \mathbf{v}_n^\star, \qquad \boldsymbol{\mu}_k = \frac{\bar{\boldsymbol{\mu}}_k}{\left\|\bar{\boldsymbol{\mu}}_k\right\|_2}, \qquad e_k = \sum_{\substack{n\in\mathcal{B}\\ \hat{y}_n=k}} \omega_n, \tag{5}$$

Thus, $\boldsymbol{\mu}_k$ forms the confidence-weighted estimated prototypes of class k and e_k the total confidence of that predicted class. To avoid unreliable estimates supported by few or low-confidence samples, we retain only classes with total confidence higher than a given threshold $e_{\min}$. These retained known prototypes form the active known set $\mathcal{A} = \{\boldsymbol{\mu}_k : e_k \geq e_{\min}\}$.

Novel prototypes. The initial buffer may also contain samples from novel categories, for which neither class names nor text embeddings are available. Unlike known classes, their prototypes therefore cannot be estimated from zero-shot assignments and must instead be discovered from the geometry of the adapted visual features. We treat the active known prototypes $\mathcal{A}$ as pre-existing centers and sequentially initialize L additional prototypes $\{\boldsymbol{\nu}_l\}_{l=1}^{L}$ using spherical k-means++ (Arthur

and Vassilvitskii, 2007). Before selecting the l-th novel prototype, let $\mathcal{C}_l = \mathcal{A} \cup \{\boldsymbol{\nu}_r\}_{r<l}$ denote the set of prototypes available so far. For each buffered feature $\mathbf{v}_n^\star$, we compute its squared distance to the closest existing prototype as

$$D_l^2(n) = \min_{c \in \mathcal{C}_l} \|\mathbf{v}_n^\star - c\|_2^2 = 2\Big(1 - \max_{c \in \mathcal{C}_l} \mathbf{v}_n^{\star\top} c\Big), \tag{6}$$

Thus, features already well represented by an active known prototype or a previously selected novel prototype receive low sampling probability, whereas features in uncovered regions receive higher probability. Repeating this process yields L novel prototypes that complement the active known prototypes and provide initial coverage of the buffered visual feature space.

Class assignment. After initialization, the unified prototype set is

$$\mathcal{P} = \mathcal{A} \cup \{\boldsymbol{\nu}_l\}_{l=1}^{L}, \tag{7}$$

where known and novel categories are represented in the same unified visual embedding space. Under the vMF model with shared concentration and uniform priors, MAP based inference reduces to nearest-prototype assignment by cosine similarity (Banerjee et al., 2005):

$$\ell(\mathbf{v}^\star) = \arg\max_{c \in \mathcal{P}} \mathbf{v}^{\star\top} c. \tag{8}$$

An assignment to $\boldsymbol{\mu}_k$ predicts known class k, whereas an assignment to $\boldsymbol{\nu}_l$ identifies discovered novel category l.

Streaming adaptation. Re-alignment is performed only once on the initial buffer $\mathcal{B}$. Thereafter, $\theta_I^\star$ remains frozen and adaptation proceeds by updating only the category prototypes as new test batches arrive. For each incoming batch $\mathcal{B}_t$, all assignments are computed before updating the prototypes. Known categories remain semantically anchored by the adapted zero-shot classifier: let $\hat{y}_n^\star$ and $\omega_n^\star$ denote the prediction and confidence obtained from Eq. 1 using $\theta_I^\star$. Novel categories, for which no textual supervision is available, instead follow the unified prototype assignment in Eq. 8. We therefore define $\mathcal{K}_k^t = \{n \in \mathcal{B}_t : \hat{y}_n^\star = k\}$ and $\mathcal{I}_l^t = \{n \in \mathcal{B}_t : \ell(\mathbf{v}_n^\star) = \boldsymbol{\nu}_l\}$ as the samples supporting known class k and novel category l, respectively. For each known class, we compute the confidence-weighted mean direction and update its prototype

$$\boldsymbol{\mu}_k \leftarrow \frac{(1-\eta)\,\boldsymbol{\mu}_k + \eta\,\tilde{\boldsymbol{\mu}}_k^t}{\|(1-\eta)\,\boldsymbol{\mu}_k + \eta\,\tilde{\boldsymbol{\mu}}_k^t\|_2}, \qquad \tilde{\boldsymbol{\mu}}_k^t = \frac{\sum_{n \in \mathcal{K}_k^t} \omega_n^\star\,\mathbf{v}_n^\star}{\|\sum_{n \in \mathcal{K}_k^t} \omega_n^\star\,\mathbf{v}_n^\star\|_2}, \tag{9}$$

Similarly, novel prototypes are updated as

$$\boldsymbol{\nu}_l \leftarrow \frac{(1-\lambda)\,\boldsymbol{\nu}_l + \lambda\,\tilde{\boldsymbol{\nu}}_l^t}{\|(1-\lambda)\,\boldsymbol{\nu}_l + \lambda\,\tilde{\boldsymbol{\nu}}_l^t\|_2}, \qquad \tilde{\boldsymbol{\nu}}_l^{\,t} = \frac{\sum_{n \in \mathcal{I}_l^t} \mathbf{v}_n^\star}{\|\sum_{n \in \mathcal{I}_l^t} \mathbf{v}_n^\star\|_2}, \tag{10}$$

where $\eta, \lambda \in (0, 1]$ control the update rates for known and novel prototypes, respectively. If no samples support a prototype in $\mathcal{B}_t$, it remains unchanged. Finally, confidence for each known class accumulates over the stream:

$$e_k \leftarrow e_k + \sum_{n \in \mathcal{K}_k^t} \omega_n^\star, \tag{11}$$

and its prototype enters the active set $\mathcal{A}$ once $e_k \geq e_{\min}$.

5 Experiments

5.1 Setup

Datasets. We evaluate PACT on five benchmarks spanning synthetic corruption and natural domain shift. We use the standard TTA benchmarks CIFAR-10-C, CIFAR-100-C and ImageNet-C (Hendrycks et al., 2021). We also evaluate on CUB-C (Wang et al., 2025), which applies seven corruptions to the fine-grained CUB dataset (Wah et al., 2011) and DomainNet (Peng et al., 2019), comprising 345 classes across six domains. For each dataset, a proportion ρ of classes is designated as known and the

Table 2: **TT-GCD on corruption benchmarks** (severity 5, mean over 15 corruption types). Clustering accuracy (%), mean over 3 seeds with standard deviation. Best in **bold**, second best underlined.

	CIFAR-10-C			CIFAR-100-C			ImageNet-C		
Method	All	Known	Novel	All	Known	Novel	All	Known	Novel
Test Time Adaptation									
TENT++	$33.41_{\pm.10}$	$60.06_{\pm.12}$	$6.75_{\pm.25}$	$15.78_{\pm.01}$	$17.16_{\pm.02}$	$14.41_{\pm.04}$	$13.14_{\pm.07}$	$19.58_{\pm.12}$	$6.69_{\pm.03}$
BATCLIP++	$29.24_{\pm.08}$	$46.88_{\pm.01}$	$11.59_{\pm.16}$	$15.57_{\pm.01}$	$15.31_{\pm.07}$	$15.83_{\pm.09}$	$15.74_{\pm.06}$	$18.83_{\pm.05}$	$12.64_{\pm.07}$
SAT++	$23.43_{\pm.07}$	$34.48_{\pm.31}$	$12.39_{\pm.29}$	$13.25_{\pm.07}$	$13.69_{\pm.18}$	$12.81_{\pm.12}$	$10.55_{\pm.03}$	$12.31_{\pm.04}$	$8.80_{\pm.03}$
Open-Set Test Time Adaptation									
OSTTA++	$39.24_{\pm.11}$	$\underline{70.43}_{\pm.31}$	$8.06_{\pm.12}$	$19.29_{\pm.14}$	$20.97_{\pm.19}$	$17.61_{\pm.08}$	$14.30_{\pm.06}$	$\underline{21.44}_{\pm.11}$	$7.16_{\pm.10}$
UniEnt++	$\underline{39.48}_{\pm.13}$	$\mathbf{71.18}_{\pm.55}$	$7.77_{\pm.74}$	$\underline{19.79}_{\pm.09}$	$\underline{21.70}_{\pm.17}$	$\underline{17.88}_{\pm.04}$	$12.91_{\pm.05}$	$19.00_{\pm.16}$	$6.83_{\pm.07}$
Generalized Category Discovery									
SimGCD++	$28.49_{\pm1.02}$	$47.50_{\pm1.66}$	$9.48_{\pm1.21}$	$15.20_{\pm.37}$	$14.77_{\pm.31}$	$15.64_{\pm.54}$	$14.07_{\pm.28}$	$17.93_{\pm.39}$	$10.21_{\pm.17}$
HiLo++	$16.63_{\pm.16}$	$16.85_{\pm1.42}$	$\underline{16.41}_{\pm1.67}$	$5.69_{\pm.06}$	$4.89_{\pm.15}$	$6.49_{\pm.21}$	$\underline{16.93}_{\pm.60}$	$20.19_{\pm.71}$	$\underline{13.66}_{\pm.13}$
PACT (*Ours*)	$\mathbf{56.31}_{\pm.58}$	$65.39_{\pm.60}$	$\mathbf{47.24}_{\pm1.09}$	$\mathbf{29.11}_{\pm.11}$	$\mathbf{33.67}_{\pm.18}$	$\mathbf{24.55}_{\pm.28}$	$\mathbf{22.20}_{\pm.12}$	$\mathbf{24.84}_{\pm.27}$	$\mathbf{19.56}_{\pm.03}$

remaining $(1-\rho)$ as novel, we use $\rho = 0.5$, i.e., 50% known and 50% novel classes. Class splits and seeds are detailed in App. B.

Baselines. As no prior method directly addresses TT-GCD, we adapt seven representative baselines to this setting. For TTA methods, TENT++ (Wang et al., 2021), BATCLIP++ (Maharana et al., 2025), SAT++ (Mishra et al., 2026), OSTTA++ (Lee et al., 2023), and UniEnt++ (Gao et al., 2024), we first separate known and novel samples using a two-component Gaussian mixture over their maximum similarity to the known text embeddings. Adaptation follows each original method on the predicted known subset, updating only LayerNorm parameters. Known samples use the adapted model predictions, while novel samples are grouped into L clusters using k-means. For GCD methods, SimGCD++ (Wen et al., 2023) and HiLo++ (Wang et al., 2025), we replace their required labeled samples with zero-shot pseudo-labeled known samples and keep the backbone fixed. Their learned prototypes are used for both known and novel categories. Full adaptation details are in App. A.1.

Evaluation protocol. We report clustering accuracy (%) following the standard practice of prior GCD literature (Wen et al., 2023) on known classes (Known), novel classes (Novel) and all classes (`All`), read from a single Hungarian assignment over the full stream. Results are averaged over all corruption types or domains of each benchmark. Tables 2 and 3 report the mean and standard deviation over three seeds. More details in App. B.

Implementation details. We use CLIP ViT-B/16 as the backbone unless stated otherwise, with the prompt `a photo of a {}.` and batch size 128. Re-alignment is performed once per stream with a learning rate of 10^{-2}, while known and novel prototypes are updated with rates $\eta = 0.05$ and $\lambda = 0.10$, respectively. All hyperparameters for PACT and the baselines are selected on four held-out CIFAR-100-C corruptions and kept fixed across benchmarks(App. C).

5.2 Main Results

Corrupted datasets. The results on the three corruption benchmarks are reported in Table 2. PACT obtains the best `All` accuracy on the three, **outperforming the second best method by 16.8** points on CIFAR-10-C, **9.3** on CIFAR-100-C and **5.3** on ImageNet-C. Furthermore, it yields the best accuracy in eight of the nine `All`, `Known` and `Novel` columns. A closer look reveals a consistent pattern, i.e., no baseline is competitive on both halves of the label space. For example, on CIFAR-10-C the three highest `Known` accuracies, UniEnt++ (71.18), OSTTA++ (70.43) and TENT++ (60.06), coincide with the three weakest `Novel` results (7.77, 8.06 and 6.75). HiLo++ shows the reverse, with the best baseline `Novel` on CIFAR-10-C (16.41) and the worst `All` on CIFAR-100-C (5.69), as its robustness relies on cross-domain labeled training, which TT-GCD does not provide (App. A.1).

Natural shift (Table 3). In this scenario, we evaluate PACT changes in visual style shifts across domains. We observe that **PACT still remains the best method on all six domains** in DomainNet,

Table 3: **TT-GCD under natural shift.** DomainNet, per domain, $K=172$. Clustering accuracy (`All`, %), mean over 3 seeds with the standard deviation. Best in **bold**, second best underlined. Known and Novel per domain are reported in App. D.1.

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	**Average**
TENT++	$28.33_{\pm.19}$	$17.42_{\pm.09}$	$26.95_{\pm.04}$	$0.52_{\pm.00}$	$\underline{31.69}_{\pm.03}$	$23.58_{\pm.14}$	21.41
BATCLIP++	$\underline{29.09}_{\pm.15}$	$16.64_{\pm.14}$	$\underline{28.10}_{\pm.31}$	$\underline{6.04}_{\pm.09}$	$30.63_{\pm.39}$	$\underline{24.85}_{\pm.40}$	$\underline{22.56}$
SAT++	$14.17_{\pm.28}$	$14.28_{\pm.41}$	$9.81_{\pm.48}$	$2.20_{\pm.01}$	$7.93_{\pm.14}$	$10.90_{\pm.09}$	9.88
OSTTA++	$26.82_{\pm.34}$	$19.19_{\pm.18}$	$25.69_{\pm.40}$	$5.71_{\pm.21}$	$27.21_{\pm.67}$	$23.92_{\pm.12}$	21.42
UniEnt++	$28.01_{\pm.09}$	$\underline{19.70}_{\pm.14}$	$26.81_{\pm.17}$	$5.50_{\pm.25}$	$30.24_{\pm.34}$	$24.80_{\pm.37}$	22.51
SimGCD++	$17.10_{\pm15.65}$	$10.53_{\pm4.56}$	$15.35_{\pm16.43}$	$5.23_{\pm.02}$	$17.97_{\pm14.91}$	$8.43_{\pm2.19}$	12.44
HiLo++	$6.13_{\pm.38}$	$5.63_{\pm.51}$	$6.76_{\pm.18}$	$2.66_{\pm.26}$	$5.94_{\pm.20}$	$5.11_{\pm.08}$	5.37
PACT (*Ours*)	$\mathbf{57.22}_{\pm1.90}$	$\mathbf{34.50}_{\pm.75}$	$\mathbf{54.49}_{\pm.93}$	$\mathbf{20.74}_{\pm.45}$	$\mathbf{65.04}_{\pm.83}$	$\mathbf{49.36}_{\pm.59}$	**46.89**

with a mean accuracy of **46.89** ***versus*** **22.56** for the strongest baseline, and per-domain margins ranging from **14.7** (Quickdraw) to **33.4** (Real). Among the baselines, it is noteworthy to mention that TENT++ collapses to 0.52 on Quickdraw, a domain that is particularly challenging for all baselines, where the best baseline yields 6.04 of accuracy. In contrast, PACT remains robust and achieves 20.74, representing a substantial margin over every other approach.

Fine-grained categories. Finally, Table 4 reports the results on CUB-C with three backbones. The accuracy of every method drops considerably with respect to the coarse-grained benchmarks. Yet, **PACT yields the best `All` accuracy with the three backbones** (18.20, 23.59 and 45.48). Moving from CLIP B/16 to BioCLIP-2 improves all methods, e.g., TENT++ gains 22.4 points, which indicates that a large part of the fine-grained gap lies in the representation. However, the gain of TENT++ comes entirely from `Known`, whereas the `Novel` accuracy achieved by PACT more than doubles, from 15.09 to 34.32 (App. D.2).

Table 4: **Fine-grained shift on CUB-C.** `All` (%), severity 5.

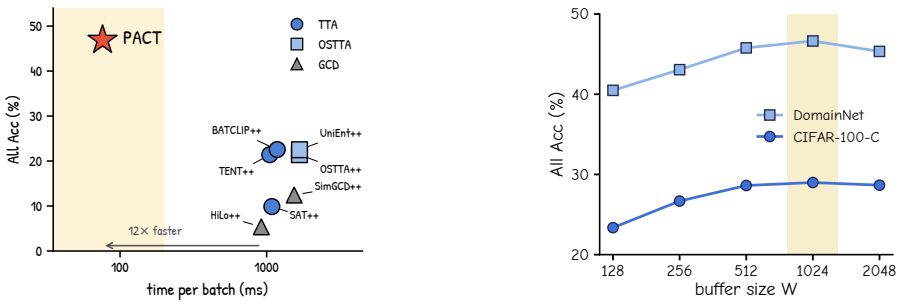
Method	B/16	L/14	BioCLIP-2
TENT++	9.54	15.01	$\underline{31.97}$
BATCLIP++	11.57	15.52	28.63
SAT++	9.11	10.42	26.22
OSTTA++	8.86	14.00	28.82
UniEnt++	7.57	13.96	20.58
SimGCD++	$\underline{13.74}$	$\underline{17.85}$	24.76
HiLo++	5.83	7.08	9.57
PACT (*Ours*)	**18.20**	**23.59**	**45.48**

5.3 Ablation Studies

Contribution of each stage. Table 5 jointly ablates the two key components of PACT: feature re-alignment(Sec. 4.1) and visual prototypes for known classes in unified category discovery(Sec.4.2). Without re-alignment, the encoder remains at θ_I^0 and without visual prototypes, known classes are represented by their text embeddings. Both components contribute to substantial performance gains. With visual prototypes fixed, re-alignment improves `All` accuracy by 5.91 points on CIFAR-100-C and 4.51 on DomainNet. Conversely, with re-alignment enabled, replacing text embeddings with visual prototypes improves `All` by 6.85 and 12.53 points, respectively. Their strongest effects are complementary: visual prototypes improve `Known` by 11.70–14.55 points on CIFAR-100-C and 20.80–27.28 on DomainNet, supporting the modality-gap motivation in Sec. 4.2. Re-alignment consistently improves `Novel` by 5.84–9.34 points despite using only known-class pseudo-labels, indicating that adapting the shared visual encoder also benefits novel categories.

Table 5: **Component ablation.** Clustering accuracy (%)

		CIFAR-100-C			**DomainNet**		
Re-align	Known proto	All	Known	Novel	All	Known	Novel
–	Text	15.11	13.83	16.39	25.22	24.45	25.91
–	Image	23.08	28.38	17.77	42.13	51.73	32.62
✓	Text	22.14	22.04	22.23	34.11	32.90	35.25
✓	Image	**28.99**	**33.74**	**24.25**	**46.64**	**53.70**	**39.68**



(a) Streaming efficiency (b) Buffer size W

Figure 3: (a) Time per batch of 128 (log scale) against mean `All` on DomainNet. (b) `All` (%) as a function of the buffer size W.

Streaming efficiency. Once re-alignment is complete, PACT freezes the encoder and updates only the prototypes, requiring no backward pass during streaming. As shown in Fig. 3(a), PACT processes a batch of 128 in 76 ms, compared with 919–1675 ms for online adaptation baselines, yielding a 12–22× speedup. The one-time re-alignment costs 30 s per stream (App. C.2).

Buffer size. Fig. 3(b) varies the buffer size W. `All` increases monotonically up to $W{=}1024$ on both benchmarks, from 23.36 to 28.99 on CIFAR-100-C and from 40.48 to 46.64 on DomainNet, and drops at $W{=}2048$ by 0.3 and 1.3. The buffer samples receive their known-class prediction from θ_0 (Eq. 1), so a larger buffer improves the re-alignment and the prototype estimates while placing a larger fraction of the stream under the pretrained model.

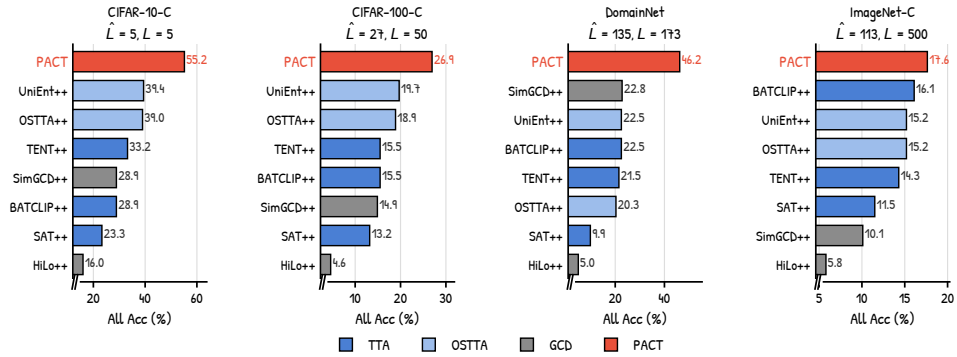


Figure 4: **TT-GCD under estimated L.** `All` accuracy (%) when the novel budget is estimated once per stream and shared by every method (App. D.3).

Unknown number of novel classes. We follow the estimation procedure of GCD (Vaze et al., 2022) to estimate the number of novel classes. k-means is run over the zero-shot features for a range of cluster counts, the count maximizing the silhouette score is found by Brent's method, and the novel budget is $\hat{L} = \hat{k} - K$ with $K = |\mathcal{C}_{\text{known}}|$ given by the setting. Fig. 4 reports the resulting accuracies. PACT remains first on the four benchmarks, **by 15.8 points on CIFAR-10-C, 7.3 on CIFAR-100-C and 23.4 on DomainNet**, and by 1.6 on ImageNet-C.

6 Conclusion

We introduced TT-GCD, where a vision-language model must adapt to a shifted, unlabeled data stream where it must recognise known classes as well as discover the novel ones, without any source data or labels. Existing test-time adaptation and category discovery methods fail to address this problem. We introduced PACT which solves this with a one-shot re-alignment of the image-encoder LayerNorm parameters on a warmup buffer and a streaming assignment among visual prototypes. PACT consistently outperforms seven adapted baselines across five benchmark datasets to effectively handle the task of TT-GCD.

7 AI USE STATEMENT

We used large language models (LLMs) to check the grammar and to help format tables and style figures. LLMs were not used to review the literature or design the method or to generate any mathematical proofs. The authors reviewed and verified all LLM-assisted content and take full responsibility for the paper.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2020.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-898716-24-5.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6: 1345–1382, 2005.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Lorenzo Caselli, Marco Mistretta, Simone Magistri, and Andrew D. Bagdanov. SpectralGCD: Spectral concept selection and cross-modal representation learning for generalized category discovery. In *International Conference on Learning Representations (ICLR)*, 2026.

Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1729–1739, 2023.

Hao Dong, Eleni Chatzi, and Olga Fink. Towards robust multimodal open-set test-time adaptation via adaptive entropy-aware optimization. In *International Conference on Learning Representations (ICLR)*, 2025.

Ruoyi Du, Dongliang Chang, Kongming Liang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. On-the-fly category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.17639.

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

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23975–23984, 2024. doi: 10.1109/CVPR52733.2024.02263.

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

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 (CVPR)*, pages 14162–14171, June 2024.

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 (ICCV)*, pages 16380–16389, 2023.

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 (ICCV)*, 2023.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 17612–17625, 2022.

Fan Lyu, Tianle Liu, Zhang Zhang, Fuyuan Hu, and Liang Wang. Test-time discovery via hashing memory. *arXiv preprint arXiv:2503.10699*, 2025. URL https://arxiv.org/abs/2503.10699.

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

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, 2000.

Shambhavi Mishra, Julio Silva-Rodriguez, Ismail Ben Ayed, Marco Pedersoli, and Jose Dolz. Semantic anchor transport: Robust test-time adaptation for vision-language models, 2026. URL https://arxiv.org/abs/2411.17002.

Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift, 2021. URL https://arxiv.org/abs/2006.10963.

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 Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16888–16905. PMLR, 17–23 Jul 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 *International Conference on Learning Representations*, 2023.

Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. Clip-gcd: Simple language guided generalized category discovery, 2023. URL https://arxiv.org/abs/2305.10420.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation, 2019. URL https://arxiv.org/abs/1812.01754.

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 (ICML)*, pages 8748–8763, 2021.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2404.07983.

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. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14274–14289, 2022.

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

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7492–7501, 2022.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

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

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, and Ming-Ming Cheng. Get: Unlocking the multi-modal potential of clip for generalized category discovery. *arXiv preprint arXiv:2403.09974*, 2024.

Hongjun Wang, Sagar Vaze, and Kai Han. Hilo: A learning framework for generalized category discovery robust to domain shifts. In *ICLR*, 2025.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16590–16600, 2023.

Yanan Wu, Zhixiang Chi, Yang Wang, and Songhe Feng. Metagcd: Learning to continually learn in generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2308.11063.

Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *CVPR*, 2023.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. *Advances in Neural Information Processing Systems*, 37:67795–67825, 2024.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Prototypical hash encoding for on-the-fly fine-grained category discovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2410.19213.