RELIABILITY-GATED TEST-TIME ADAPTATION FOR OPEN-WORLD VISION-LANGUAGE RECOGNITION

Anonymous Authors

ABSTRACT

Pretrained vision-language models provide strong zero-shot recognition but remain brittle when deployment streams combine evolving domain shifts with categories absent from the task label set. Existing test-time adaptation methods typically update on confident predictions in a closed label space, making open-world streams especially vulnerable to confirmation bias and irreversible drift. We present REG-TTA, a reliability-gated framework for online adaptation of frozen vision-language models. REG-TTA scores each incoming sample using prediction consistency across stochastic augmentations, image-text semantic agreement, and local feature density in a dual-timescale prototype memory. Only reliable known-class instances update a lightweight prompt adapter, while uncertain or potentially unknown inputs are excluded from entropy minimization and organized into temporary semantic prototypes. We analyze the gradient induced by unknown-class samples and show that closed-set entropy minimization introduces a systematic pseudo-class bias whenever the model is overconfident on open-set inputs. The released supplementary package contains an implementation, public-dataset stream builders, aggregation scripts, and plotting utilities for ImageNet-C, ImageNet-R, ImageNet-A, ImageNet-V2, ObjectNet, and DomainNet evaluation.

1 INTRODUCTION

Vision-language models (VLMs) trained on image-text pairs, such as CLIP and related contrastive systems, make it possible to classify images by matching visual embeddings to natural-language class prompts (Radford et al., 2021; Jia et al., 2021; Li et al., 2022). This zero-shot interface is attractive for deployment because the label vocabulary can be changed without retraining. Real deployment streams, however, rarely remain stationary: images may be corrupted, rendered in non-photographic styles, captured from unusual viewpoints, or drawn from categories that were not enumerated when the system was configured (Hendrycks and Dietterich, 2019; Hendrycks et al., 2021b;a; Recht et al., 2019; Barbu et al., 2019; Peng et al., 2019).

Test-time adaptation (TTA) attempts to close this gap by updating the model during inference without access to the source data (Sun et al., 2020; Wang et al., 2021). Entropy minimization and prompt tuning can improve adaptation under benign covariate shift (Shu et al., 2022; Niu et al., 2022; 2023; Wang et al., 2022). Yet the standard closed-set assumption is hazardous in open-world streams: an unknown object must still be assigned to one of the known labels, and high confidence can be a symptom of spurious alignment rather than reliability. Updating on such samples can move prompts toward unsupported pseudo-classes, degrading both known-class accuracy and unknown rejection.

REG-TTA addresses this failure mode by placing a reliability gate before every parameter update. The gate uses three complementary signals. First, stochastic-view consistency checks whether predictions remain stable under augmentations. Second, cross-modal semantic agreement checks whether the image embedding is close to the text prototype of its predicted class. Third, prototype density checks whether the sample lies near recent reliable known-class features rather than in a sparse region of the stream. A dual-timescale memory separates persistent class semantics from short-term domain statistics, allowing adaptation to rapid shifts while retaining a conservative anchor for recurring domains.

This paper makes three contributions. (i) We formulate open-world TTA for VLMs in which both domain identity and unknown-class arrivals vary over time. (ii) We propose a selective adaptation objective with a reliability gate and dual-timescale memory. (iii) We provide a gradient-bias analysis

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.

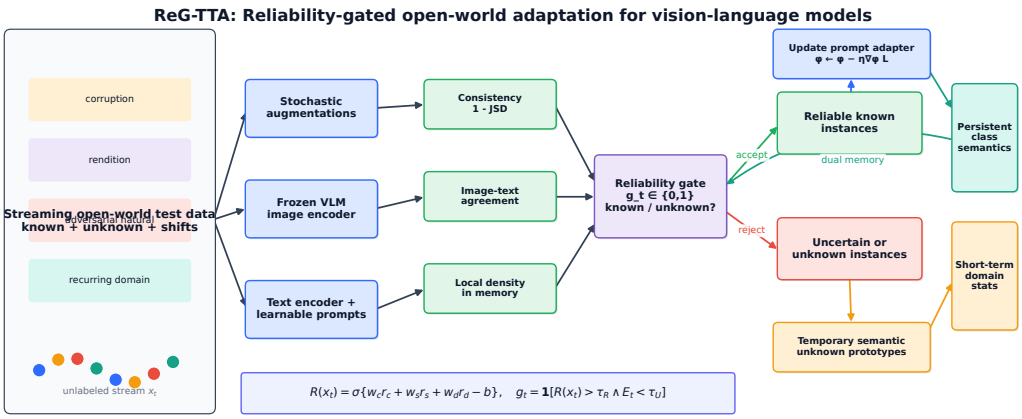

Figure 1: **Overview of REG-TTA.** The method processes an unlabeled open-world stream with a frozen VLM, computes reliability from consistency, semantic agreement, and memory density, and updates a prompt adapter only for reliable known-class instances. Rejected samples are preserved as temporary semantic prototypes rather than used for closed-set entropy minimization.

showing why unknown samples should not be treated as ordinary entropy-minimization updates. Figure 1 summarizes the framework.

2 RELATED WORK

Vision-language recognition and prompt adaptation. Contrastive image-text pretraining enables classification by comparing image embeddings with text embeddings of class prompts (Radford et al., 2021; Jia et al., 2021). Prompt learning methods tune the textual context around class names to adapt VLMs to downstream tasks (Zhou et al., 2022b;a; Zhang et al., 2022). Test-time prompt tuning extends this idea by optimizing prompts on augmented views of each test image (Shu et al., 2022). REG-TTA keeps the VLM backbone frozen and updates only a small prompt adapter, but differs by treating the open-world decision as a prerequisite for any update.

Test-time adaptation. Fully test-time adaptation has been studied through self-supervised objectives, entropy minimization, reliable sample selection, teacher-student stabilization, and parameter-free alternatives (Sun et al., 2020; Wang et al., 2021; Niu et al., 2022; Wang et al., 2022; Boudiaf et al., 2022; Chen et al., 2022; Niu et al., 2023). These approaches primarily assume that every test sample belongs to the known label space. REG-TTA follows the selective-update philosophy of reliable TTA, but its gate explicitly combines VLM semantics and prototype density to avoid learning from unknown-class inputs.

Open-set and out-of-distribution recognition. Open-set recognition and OOD detection study the rejection of samples outside the known label set (Scheirer et al., 2013; Hendrycks and Gimpel, 2017; Liang et al., 2018; Lee et al., 2018; Liu et al., 2020; Fort et al., 2021; Vaze et al., 2022). VLM-specific OOD detectors use image-text representations to measure semantic mismatch (Ming et al., 2022). REG-TTA is complementary: rejection is not only an evaluation target but also a control signal for whether online learning should occur.

3 PROBLEM SETUP

Let a frozen VLM provide an image encoder f_θ and text encoder g_θ. The known label vocabulary is $\mathcal{Y} = \{1, \ldots, K\}$ with class names c_k. At time t, the system observes an unlabeled image x_t from a stream distribution P_t that may drift over time and may place mass on unknown labels $\mathcal{U}_t \cap \mathcal{Y} = \emptyset$. The learner can update only lightweight prompt parameters ϕ and memory state $\mathcal{M}_t$; source training data are unavailable.

For prompt template $q_\phi(c_k)$, normalized image feature $z_t = f_\theta(x_t)/\left\|f_\theta(x_t)\right\|_2$, and normalized text feature $u_{k,\phi} = g_\theta(q_\phi(c_k))/\left\|g_\theta(q_\phi(c_k))\right\|_2$, closed-set logits and probabilities are

$$\ell_{t,k} = \tau^{-1} z_t^\top u_{k,\phi}, \qquad p_{t,k} = \mathrm{softmax}(\ell_t)_k. \tag{1}$$

The open-world goal is to maintain accurate known-class predictions while assigning high unknownness scores to samples outside $\mathcal{Y}$, and to avoid increasing future loss through updates on unreliable samples.

4 METHOD

4.1 RELIABILITY-GATED ADAPTATION

For each incoming image we sample M stochastic augmentations $a_m(x_t)$ and compute probabilities $p_t^{(m)}$. Let $\bar{p}_t = M^{-1}\sum_m p_t^{(m)}$ and $\hat{y}_t = \arg\max_k \bar{p}_{t,k}$. Reliability is computed from three scores in $[0,1]$:

$$r_c(x_t) = 1 - \frac{1}{\log K}\,\mathrm{JSD}\left(p_t^{(1)}, \ldots, p_t^{(M)}\right), \tag{2}$$

$$r_s(x_t) = \frac{1 + z_t^\top u_{\hat{y}_t,\phi}}{2}, \tag{3}$$

$$r_d(x_t) = \max_{k\in\mathcal{Y}} \exp\left(-\frac{\left\|z_t - \mu_{k,t}^s\right\|_2^2}{2(\sigma_{k,t}^s)^2 + \epsilon}\right), \tag{4}$$

where $\mu_{k,t}^s$ and $\sigma_{k,t}^s$ denote short-term class-prototype mean and scale. The reliability probability and binary update gate are

$$R_t = \sigma(w_c r_c + w_s r_s + w_d r_d - b), \qquad g_t = \mathbf{1}\{R_t \geq \tau_R\}\mathbf{1}\{E_t \leq \tau_U\}, \tag{5}$$

with energy $E_t = -\tau \log\sum_k \exp(\ell_{t,k}/\tau)$. A sample is used for adaptation only when both reliability and unknownness tests are satisfied.

The prompt adapter is optimized using a gated consistency-entropy objective,

$$\mathcal{L}_t(\phi) = g_t\left[H(\bar{p}_t) + \lambda_c \frac{1}{M}\sum_{m=1}^{M}\mathrm{KL}(\bar{p}_t \| p_t^{(m)})\right] + \lambda_0\left\|\phi - \phi_0\right\|_2^2, \tag{6}$$

where ϕ_0 is the initial prompt state. Unknown or unreliable samples do not backpropagate through Eq. (6); their embeddings are retained only in a temporary semantic-prototype buffer.

4.2 DUAL-TIMESCALE PROTOTYPE MEMORY

Memory stores a persistent class-level prototype $(\mu_{k,t}^p, n_{k,t}^p)$ and a short-term domain prototype $(\mu_{k,t}^s, \sigma_{k,t}^s, n_{k,t}^s)$ for each known class. For a reliable sample with predicted class $\hat{y}_t = k$, the updates are

$$\mu_{k,t+1}^p = (1-\alpha_p)\mu_{k,t}^p + \alpha_p z_t, \qquad \alpha_p \ll \alpha_s, \tag{7}$$

$$\mu_{k,t+1}^s = (1-\alpha_s)\mu_{k,t}^s + \alpha_s z_t, \tag{8}$$

$$(\sigma_{k,t+1}^s)^2 = (1-\alpha_s)(\sigma_{k,t}^s)^2 + \alpha_s\left\|z_t - \mu_{k,t+1}^s\right\|_2^2. \tag{9}$$

Rejected samples are clustered with a cosine-threshold rule into temporary prototypes ν_j that support unknownness calibration but never define known labels. The persistent branch changes slowly and anchors recurring domains, while the short-term branch adapts quickly to local appearance statistics.

4.3 WHY UNKNOWNS BIAS ENTROPY MINIMIZATION

The following proposition formalizes the risk of closed-set entropy minimization on unknown inputs.

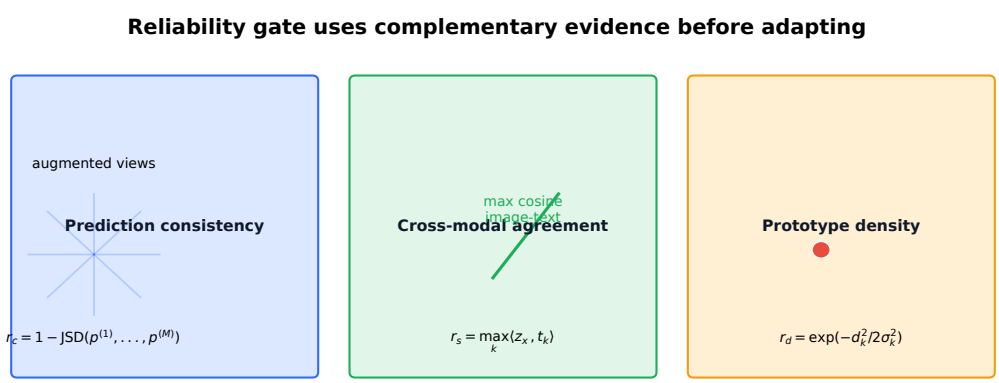

Figure 2: **Reliability evidence.** Consistency tests invariance over augmented views, semantic agreement measures image-text alignment with the predicted class, and density compares the input to online prototypes. The gate combines all three before allowing any update.

Algorithm 1 REG-TTA for an unlabeled open-world stream

Require: frozen encoders f_θ, g_θ, class names $\{c_k\}_{k=1}^K$, prompt adapter ϕ, memory $\mathcal{M}_0$

for $t = 1, 2, \ldots$ **do**
  receive image x_t and construct stochastic views $\{a_m(x_t)\}_{m=1}^M$
  compute $p_t^{(m)}$ with current prompt adapter and $\bar{p}_t = M^{-1} \sum_m p_t^{(m)}$
  compute $r_c(x_t)$, $r_s(x_t)$, $r_d(x_t)$, energy E_t, and gate g_t
  if $g_t = 1$ **then**
    update ϕ by one or more gradient steps on Eq. (6)
    update persistent and short-term known-class memories
  else
    assign z_t to a temporary unknown prototype or create a new prototype
  end if
  output known prediction $\arg\max_k \bar{p}_{t,k}$ and unknownness score $1 - R_t$
end for

Proposition 1 (Open-set gradient bias)**.** *Let $x \sim P_U$ be drawn from an unknown class and let $p_\phi(y \mid x)$ be the induced closed-set distribution over $\mathcal{Y}$. If p_ϕ is not uniform almost surely, then the entropy-minimization gradient $\nabla_\phi H(p_\phi(\cdot \mid x))$ has a nonzero component that increases the logit margin of some known pseudo-class for a positive-measure subset of unknown samples. If a gate accepts unknown samples with probability ρ_U, the expected unknown-induced gradient norm is reduced by at least a factor $1 - \rho_U$ under independent gating.*

The proof in Appendix A expands $\nabla_\phi H(p)$ through the logit Jacobian and shows that high-confidence unknowns behave like biased pseudo-label updates. This motivates selective adaptation: reducing false acceptance of unknowns reduces the leading bias term without preventing reliable known-class updates.

5 EVALUATION PROTOCOL

The bundled code defines continuous streams from the public benchmarks in Table 1. Each stream interleaves known classes with held-out unknown classes and changes domains over time. ImageNet-derived datasets stress corruption, natural adversarial examples, renditions, and re-sampling; ObjectNet stresses viewpoint, background, and rotation; DomainNet stresses multi-domain transitions. Figure 3 summarizes coverage across stress factors.

Baselines. The code includes zero-shot CLIP, TENT, CoTTA-style teacher adaptation, EATA/SAR-style reliable entropy minimization, TPT, energy-based OOD scoring, maximum-softmax-probability

Table 1: Public benchmark families configured by the released stream builder. Counts reflect commonly used public dataset descriptions; actual class overlap after mapping is logged by the builder.

Benchmark	Primary shift	Classes	Size / structure	Open-world use
ImageNet-C	synthetic corruptions	1000	15 corruptions $\times$ 5 severities	known shift stream
ImageNet-R	renditions	200	30k images	known/unknown split
ImageNet-A	natural adversarial	200	7.5k images	hard unknown rejection
ImageNet-V2	re-sampled ImageNet	1000	3 test sets, 10k images each	recurring domain
ObjectNet	viewpoint/background	313	50k images, 113 ImageNet-overlap classes	open-class stream
DomainNet	domain/style	345	approx. 0.6M images, 6 domains	multi-domain open stream

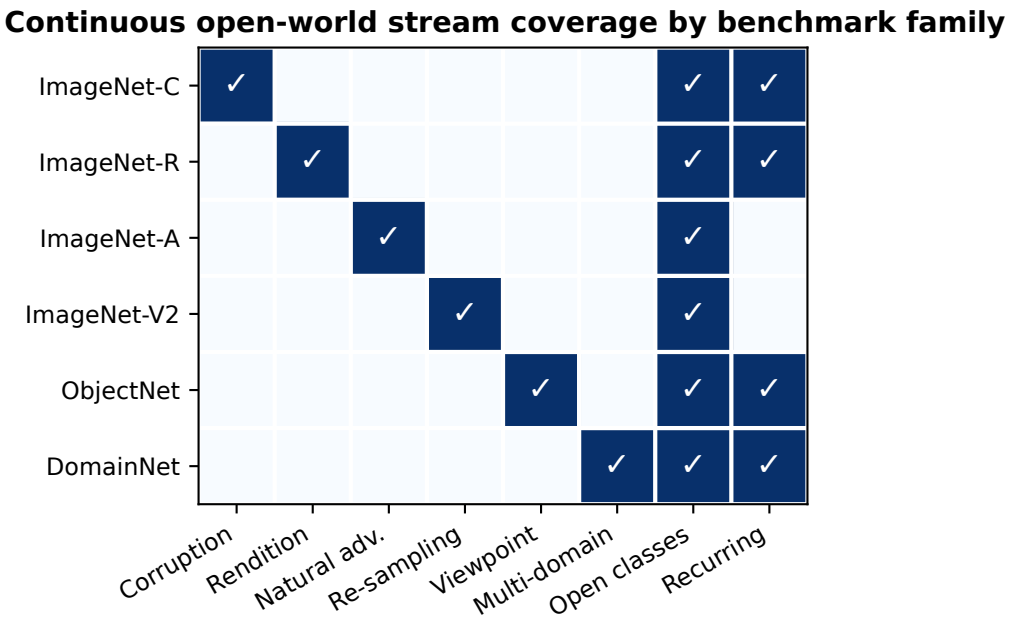

Figure 3: **Stream stress factors.** The benchmark suite combines domain drift with open classes. The code logs the exact stream ordering, class split, seed, and sample count for each run.

rejection, and VLM semantic scoring. Baseline wrappers expose a common online interface so that parameter counts, update counts, and latency can be logged consistently.

Metrics. For known samples we report top-1 accuracy over time and per-domain recovery after shifts. For unknown samples we report AUROC, AUPR, and FPR95 using unknownness scores. To measure stability, we report adaptation-induced forgetting as the drop on a held-out known validation stream after online updates. Table 2 gives the exact definitions implemented in `code/reg_tta/metrics.py`.

Ablations. We isolate the role of consistency, semantic agreement, density, dual memory, temporary unknown prototypes, energy filtering, and prompt regularization. The aggregation script writes both table-ready means over seeds and time-resolved curves for abrupt and recurring shifts. Figure 4 shows the reporting pipeline.

6 Discussion and Limitations

REG-TTA is designed for the practical setting where the user can name known classes but cannot enumerate all future categories or domains. Its conservative gate prioritizes avoiding harmful updates over adapting to every confident prediction. The main limitation is that gate thresholds and memory horizons must be calibrated for the deployment cost of false acceptance versus false rejection. In

Table 2: Logged metrics and definitions. All metrics are computed from stream-level JSONL records produced during inference.

Metric	Definition	Purpose
Known top-1	$\frac{1}{\lvert\mathcal{I}_K\rvert}\sum_{t\in\mathcal{I}_K}\mathbf{1}\{\hat{y}_t = y_t\}$	known-class recognition
Unknown AUROC	$\mathrm{AUROC}(s_t, \mathbf{1}\{y_t \notin \mathcal{Y}\})$	open-set detection
FPR95	false-positive rate at 95% unknown recall	deployment rejection risk
Forgetting	$A_{\mathrm{pre}} - A_{\mathrm{post}}$ on anchor stream	stability after adaptation
Update rate	$\lvert\{t : g_t = 1\}\rvert/T$	reliability-gate selectivity
Parameter budget	trainable parameters / total parameters	adaptation footprint

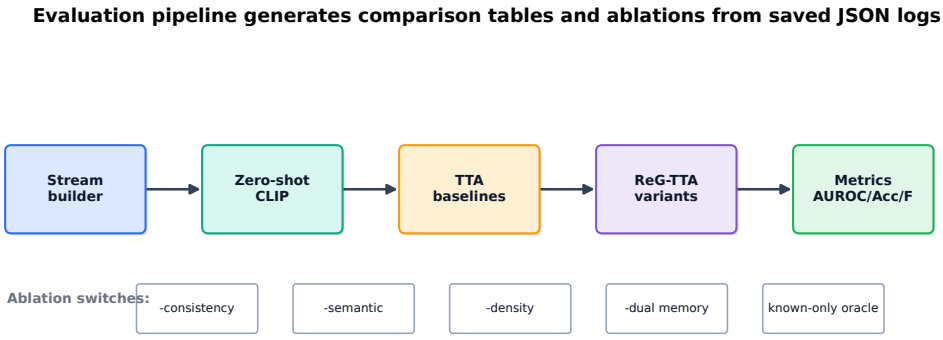

Figure 4: **Reporting pipeline.** A single online interface runs zero-shot, adaptation, open-set, and ablation variants. JSONL logs are converted to comparison tables and figures by the supplementary scripts.

addition, VLM semantic agreement can inherit biases from pretraining data, so unknown detection should be audited for classes and visual domains relevant to the deployment environment.

7 Conclusion

Open-world test-time adaptation requires deciding not only what to learn, but when learning should be withheld. ReG-TTA combines stochastic consistency, image-text semantic agreement, and online prototype density to gate prompt adaptation in VLMs. The resulting objective avoids using uncertain or unknown samples as closed-set entropy targets and maintains separate memories for persistent class semantics and short-term domain statistics.

Reproducibility Statement

The supplementary package includes source code, configuration files, dataset stream builders, metric aggregation, plotting scripts, and proof details. Runs write immutable JSONL records containing dataset path, class split, stream order, seed, checkpoint identifier, trainable-parameter count, update count, and metric values.

References

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Table 3: Default implementation budget for ViT-B/16 CLIP experiments. The prompt-adapter count is determined by context length L, embedding width d, and optional residual MLP width h.

Component	Trainable?	State scale	Default	Notes
Image encoder f_θ	no	backbone	frozen	no source gradients
Text encoder g_θ	no	backbone	frozen	class prompts re-encoded
Context prompt ϕ	yes	$L \times d$	$L = 16$	class-shared context
Residual prompt MLP	yes	$2dh + h + d$	optional	disabled by default
Persistent memory	no grad	$K \times d$	EMA $\alpha_p = 0.01$	semantic anchor
Short-term memory	no grad	$K \times d$	EMA $\alpha_s = 0.10$	fast domain tracking
Unknown buffer	no grad	$B \times d$	capped FIFO	rejection calibration

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 34, pages 7068–7081, 2021.

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

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

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

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, 2021b.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475, 2020.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, volume 35, pages 35087–35102, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16888–16905. PMLR, 2022.

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

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the 36th International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 14274–14289, 2022.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022.

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.

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*, pages 7201–7211, 2022.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In *Proceedings of the European Conference on Computer Vision*, pages 493–510. Springer, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022a.

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

A PROOF DETAILS

For logits $\ell \in \mathbb{R}^K$ and probabilities $p = \mathrm{softmax}(\ell)$, entropy is $H(p) = -\sum_k p_k \log p_k$. The derivative with respect to a logit ℓ_j is

$$\frac{\partial H}{\partial \ell_j} = -p_j \left(\log p_j + H(p)\right). \tag{10}$$

Let $J_\phi(x) = \partial \ell_\phi(x)/\partial \phi$ be the prompt-logit Jacobian. Then

$$\nabla_\phi H(p_\phi(\cdot \mid x)) = J_\phi(x)^\top b(p), \qquad b_j(p) = -p_j(\log p_j + H(p)). \tag{11}$$

When p is uniform, $b(p) = 0$. When p is non-uniform, at least one logit has a negative entropy gradient that increases its margin under gradient descent. Thus unknown samples with spuriously concentrated closed-set probabilities produce pseudo-class-attracting updates even though none of the known labels is correct. If a gate independently blocks an unknown with probability $1 - \rho_U$, the accepted-gradient expectation is $\rho_U \mathbb{E}[J^\top b(p)]$, giving the stated reduction of the unknown-induced component.

B THRESHOLD CALIBRATION WITHOUT SOURCE DATA

The implementation supports three source-free calibration modes: fixed thresholds supplied in the config; quantile thresholds estimated from an initial unlabeled warm-up stream; and memory-adaptive thresholds where τ_R is increased when the unknown buffer grows faster than the known memory. Warm-up images are not used to update prompt parameters unless they pass the same gate as later samples.

C STREAM CONSTRUCTION DETAILS

For each dataset, classes are mapped to a canonical string label. Known and unknown sets are created by deterministic hashing of class names under a seed, which prevents accidental leakage from directory ordering. A stream segment is a tuple (D, q, n) consisting of a dataset/domain name, a class mixture q over known and unknown labels, and a length n. Abrupt shifts concatenate segments with different datasets or domains. Recurring shifts reuse an earlier segment after intervening domains.

D SUPPLEMENTARY ALGORITHMIC NOTES

Algorithm 2 Temporary unknown prototype update

Require: normalized embedding z_t, unknown buffer $\mathcal{B} = \{(\nu_j, m_j)\}_{j=1}^{B}$, threshold τ_B, cap $B_{\max}$

if $\mathcal{B}$ is empty **then**
  create prototype $(\nu_1 = z_t, m_1 = 1)$
else
  $j^\star = \arg\max_j z_t^\top \nu_j$
  if $z_t^\top \nu_{j^\star} \geq \tau_B$ **then**
    $\nu_{j^\star} \leftarrow \frac{m_{j^\star}\nu_{j^\star} + z_t}{\left\| m_{j^\star}\nu_{j^\star} + z_t \right\|_2}$; $m_{j^\star} \leftarrow m_{j^\star} + 1$
  else
    append $(z_t, 1)$ and evict the least-recently used prototype if $B > B_{\max}$
  end if
end if

E HYPERPARAMETERS USED BY THE DEFAULT CONFIGS

Parameter	Default	Description
M	8	stochastic views per sample
τ	0.01	CLIP logit temperature as exposed by checkpoint
τ_R	0.65	reliability acceptance threshold
τ_U	config dependent	energy threshold for unknown rejection
α_p	0.01	persistent EMA rate
α_s	0.10	short-term EMA rate
λ_c	0.5	view-consistency weight
λ_0	10^{-3}	prompt-anchor weight