AIM: Augmentation-Invariant Memory for Robust Open-set Test-Time Adaptation

1st Can Zhang
Beijing University of Chemical Technology
College of Information Science and Technology
Beijing, China
alexlessend@gmail.com

2nd Yunfeng Liu
Beijing University of Chemical Technology
College of Information Science and Technology
Beijing, China
2025200819@buct.edu.cn

3rd Ye Wang
Chongqing University of Posts and Telecommunications
School of Information Science and Technology
Chongqing, China
wangye@cqupt.edu.cn

4th Ruirui Li
Beijing University of Chemical Technology
College of Information Science and Technology
Beijing, China
ilydouble@gmail.com

***Abstract*—Vision-language models (VLMs) suffer severe performance degradation when the test distribution deviates from the pre-training data. Test-time adaptation (TTA) is an effective paradigm to mitigate such distribution shifts, among which cache-based methods have drawn increasing attention due to their training-free property and remarkable efficacy. However, these methods are vulnerable in open-set scenarios, where out-of-distribution (OOD) samples readily poison the cache and mislead subsequent predictions. We present AIM (Augmentation-Invariant Memory), a robust, open-set TTA framework that equips a VLM with an on-the-fly key-value cache. AIM admits a test sample only when its prediction is both high-confidence and consistent across augmentations, guaranteeing cache purity. Leveraging this clean memory, AIM refines logits via cache retrieval and simultaneously separates ID/OOD data through a lightweight linear head trained with progressively self-thresholded pseudo-labels. Evaluated on 20 ID-OOD pairs, AIM establishes a new state-of-the-art, raising mean harmonic accuracy from 66.1% to 70.5% on CUB-200-2011.**

***Index Terms*—Vision-Language Models, Test-Time Adaptation, Robust Learning, Open-Set**

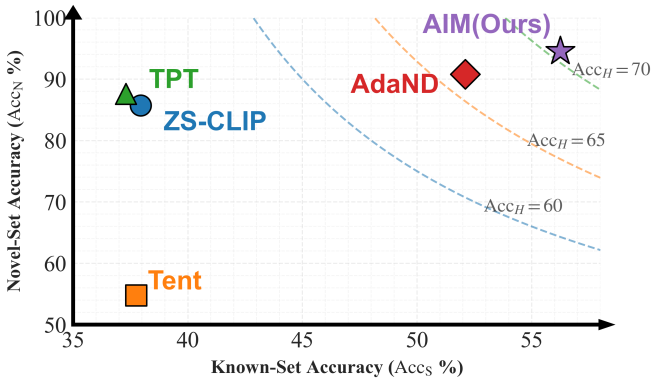

Fig. 1. Performance comparison on CUB-200-2011 averaged over four OOD datasets (iNaturalist, Places, SUN, and Texture). AIM achieves state-of-the-art performance by securing the best results in both Known-Set (Acc_S) and Novel-Set (Acc_N) accuracy, leading to the highest Harmonic Mean (Acc_H).

I. Introduction

Vision-Language Models (VLMs), exemplified by CLIP [1], have revolutionized computer vision by learning aligned visual and textual representations from massive-scale web data. Unlike traditional supervised models restricted to a fixed set of categories, VLMs demonstrate remarkable zero-shot generalization capabilities, enabling open-vocabulary recognition across diverse tasks. This paradigm shift holds the promise of deploying universal perception systems in the wild without the need for task-specific retraining.

However, despite their strong generalization, VLMs still suffer from performance degradation when deployed in target domains that differ significantly from their pre-training distributions (e.g., domain shifts caused by lighting, style, or geography). To bridge this gap, Test-Time Adaptation (TTA) [2], [3] has emerged as a practical paradigm. TTA aims to adapt a pre-trained model to the target domain on the fly, utilizing only the unlabeled test data stream. By minimizing unsupervised objectives such as prediction entropy [2] or maximizing self-consistency [4], TTA effectively aligns the model features with the test distribution, yielding significant performance gains in standard benchmarks.

While effective in controlled environments, existing TTA methods typically operate under a strict **closed-set assumption**, presuming that all test samples belong to the known categories defined by the source model. This assumption is fundamentally flawed in real-world open-set scenarios, where the test stream inevitably contains "Out-of-Distribution" (OOD) samples—semantic outliers that the model has never seen. Naively forcing the model to adapt to these outliers leads to disastrous consequences: the model minimizes entropy on erroneous predictions, causing error accumulation and catastrophic model collapse. Thus, the lack of a rejection mechanism renders standard TTA unsafe for open-world de-

ployment.

Parallel to TTA, Zero-Shot Out-of-Distribution Detection has been developed to safeguard VLMs. Recent approaches leverage the text-image alignment capability of CLIP to identify and reject unknown categories [5], [6]. While these methods excel at filtering outliers, they are predominantly **static**: they freeze the model to maintain safety, thereby sacrificing the ability to adapt to the domain shift of the *In-Distribution (ID)* samples. This creates a dilemma: TTA offers plasticity but lacks safety, while OOD detection offers safety but lacks plasticity. A unified framework that can simultaneously reject OOD samples and adapt to ID samples remains an open challenge.

To address this dilemma, we propose a robust adaptation framework named **AIM** (Augmentation-Invariant Memory). Instead of relying on unreliable prediction confidence, AIM introduces a dual-criterion mechanism to strictly purify the adaptation stream. Specifically, we utilize a retrieval-based semantic cache to maintain a history of reliable prototypes, and enforce an augmentation consistency check to filter out unstable predictions. Our contributions are summarized as follows:

- We identify the vulnerability of CLIP-based TTA in open-set scenarios and propose AIM, a novel framework that unifies robust OOD detection with effective test-time adaptation.
- We introduce a dual-criterion filtering mechanism that synergizes cache-based semantic retrieval with augmentation-based consistency checks, ensuring that only reliable ID samples are used for model updates.
- Extensive experiments on complex benchmarks demonstrate that AIM establishes new state-of-the-art results, significantly outperforming existing methods in both ID adaptation accuracy and OOD rejection performance.

To ensure reproducibility, we will make our source code publicly available upon acceptance.

II. Related Work

A. Test-Time Adaptation

Test-Time Adaptation (TTA) aims to adapt a pre-trained model to unlabeled test data on the fly. Early works primarily focused on updating model parameters by minimizing unsupervised objectives such as entropy [2] or imposing consistency regularization [7]. With the advent of Vision-Language Models (VLMs), recent approaches like TPT [4] and PromptAlign [8] leverage the strong generalization capability of CLIP by tuning learnable text prompts using stream-based unsupervised losses. In parallel to prompt tuning, cache-based adaptation offers an efficient alternative that refines CLIP predictions through retrieval in the embedding space without updating the backbone: Tip-Adapter [9] builds a key-value cache from labeled image features and injects a retrieval-induced logit correction at inference time, while Tip-Adapter-F [9] further enhances adaptability by learning a small set of lightweight parameters on top of the cache. More recently, TDA-style

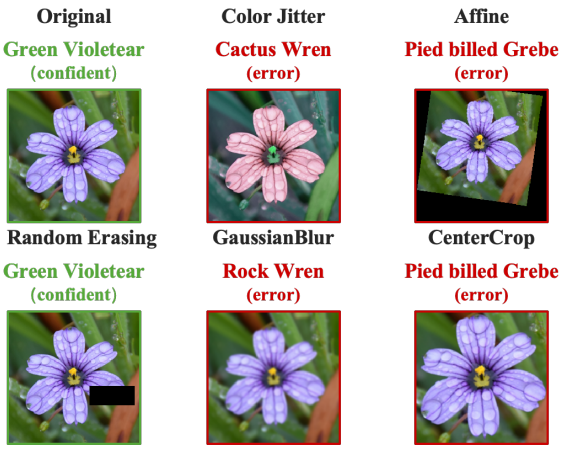

Fig. 2. Prediction instability of a high-confidence OOD sample under common augmentations: An iNaturalist image is confidently misclassified as "Green Violetear" when evaluated on CUB-200-2011 classes. As shown, common augmentations cause the pseudo-label to change to other unrelated classes.

methods [10] extend this paradigm to the test stream by updating the cache online with selected test samples and calibrating outputs via retrieval aggregation. While these methods achieve impressive performance on standard benchmarks, they predominantly operate under a **closed-set assumption**, where all test samples are assumed to belong to the known categories defined during training. This assumption is often violated in realistic deployment scenarios, where the model inevitably encounters samples from unknown semantic categories; in such open-set settings, naively adapting or updating a cache with out-of-distribution (OOD) samples can cause semantic drift, motivating the need for open-set-aware test-time adaptation.

B. Zero-Shot Out-of-Distribution Detection

Out-of-Distribution (OOD) detection is critical for the safety of AI systems. Traditional methods typically rely on output statistics from fixed classifiers, such as Maximum Softmax Probability (MSP) [11] or energy scores [12]. Recently, CLIP has revolutionized this field by enabling zero-shot OOD detection. Early approaches like MCM [5] and GL-MCM [6] utilize the maximum alignment score between images and ID text prompts to identify outliers. To further improve separability, subsequent works introduce **negative semantics** into the prompt space. For instance, CLIPN [13] learn "no-inference" prompts to absorb OOD probabilities, effectively tightening the decision boundary around ID classes. However, regardless of the detection strategy, these standard zero-shot OOD detectors are typically **static**; they focus solely on rejecting outliers and do not update the model to mitigate the domain shift of the accepted ID samples. In this work, we aim to bridge this gap by integrating robust OOD detection into the adaptation process.

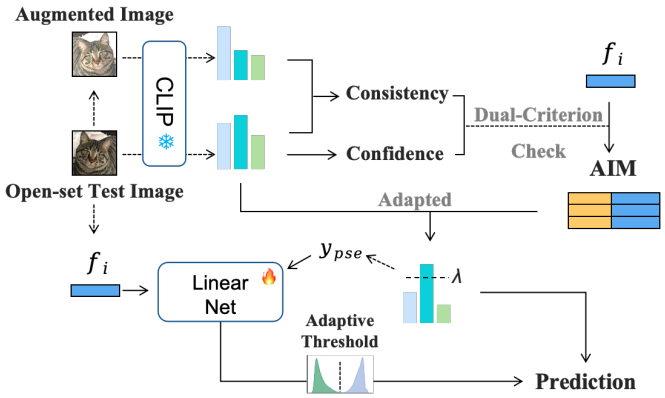

Fig. 3. Overview of the AIM framework. An incoming test image and its augmented version are fed into a frozen CLIP model. The resulting predictions are used in a Dual-Criterion Check to decide if the sample's feature f_i should be added to the AIM. The feature f_i is also used to train a linear net, which is guided by pseudo-labels y_{pse} derived from adapted outputs. The final prediction is refined using the purified AIM cache. The entire system forms a closed-loop that progressively improves both the cache purity and the prediction quality.

C. Robustness in Test-Time Adaptation

Ensuring the reliability of adaptation in non-stationary environments has attracted increasing attention. Several works attempt to filter out risky samples to prevent error propagation. For instance, EATA [14] employs an entropy-based threshold to exclude high-uncertainty samples, while SAR [15] utilizes optimization reliability to filter gradients.

More recently, AdaND [16] explicitly targets open-set scenarios by leveraging nearest-neighbor distances in the feature space to detect and reject outliers. However, these methods often struggle to distinguish between *difficult ID samples* and *confident OOD samples*. They typically rely on internal feature statistics or prediction confidence, which can be misleading when OOD samples mimic ID features.

Unlike prior works, our approach introduces a dual-criterion mechanism—leveraging both semantic cache and augmentation consistency—to strictly purify the adaptation stream. This allows the model to effectively adapt to domain shifts while robustly rejecting semantic outliers.

III. Preliminary

Zero-Shot CLIP Classification. Given a visual encoder $E_v(\cdot)$ and a textual encoder $E_t(\cdot)$, CLIP performs zero-shot classification for an image x. The image feature is $f_v = E_v(x)$ and text features for C classes are $T = \{t_1, ..., t_C\}$, where $t_c = E_t(\text{prompt}_c)$. The probability for class c is:

$$p(y = c|x) = \frac{\exp(\text{sim}(f_v, t_c)/\tau)}{\sum_{j=1}^{C} \exp(\text{sim}(f_v, t_j)/\tau)} \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity and τ is a temperature parameter.

IV. Method

Our proposed framework, **Augmentation-Invariant Memory (AIM)**, is illustrated in Fig. 3. To perform robust open-set test-time adaptation, AIM maintains an on-the-fly key–value *cache* that is continuously purified by a dual-criterion check. This purified cache is then queried to *calibrate* the CLIP logits for each incoming test sample. In parallel, a lightweight online linear net learns to separate ID/OOD signals using pseudo-labels derived from the calibrated outputs, forming a closed-loop that stabilizes both cache construction and downstream predictions. The term memory is used conceptually to describe the retained reliable knowledge over time, rather than a specific storage mechanism. While AIM's memory is practically implemented as a cache-based key–value repository for efficiency, our focus is on preserving augmentation-invariant knowledge, not on the storage structure itself.

A. The Dual-Criterion Mechanism

A central challenge in open-set test-time adaptation lies in the contamination of the cache by OOD samples. Confidence-only admission is fragile because some OOD samples can still exhibit high confidence while being semantically unstable under perturbations. To mitigate this issue, we adopt a dual-criterion admission rule that jointly checks *prediction confidence* and *augmentation consistency*.

a) Confidence criterion (entropy-based).: Given the CLIP logit vector $l(x) \in \mathbb{R}^C$ for an input x, we define the predictive distribution $p(x) = \text{softmax}(l(x))$ and its entropy

$$\mathcal{H}(x) = -\sum_{c=1}^{C} p_c(x) \log p_c(x). \tag{2}$$

A sample is regarded as confident if

$$\mathcal{H}(x) < \gamma, \tag{3}$$

where γ is a hyperparameter controlling the strictness of the confidence filter.

b) Consistency criterion (teacher–student under an augmentation).: We enforce prediction consistency between the original view and an augmented view. Given an input x, we apply an augmentation operator $a(\cdot)$ to obtain $\tilde{x} = a(x)$. The teacher prediction is obtained from the original view:

$$\hat{y}^T = \arg\max_c p_c(x), \quad p(x) = \text{softmax}(l(x)). \tag{4}$$

The student prediction is obtained by aggregating the predictive distributions from the two views:

$$\bar{p}(x) = \frac{1}{2}\left(p(x) + p(\tilde{x})\right), \quad \hat{y}^S = \arg\max_c \bar{p}_c(x). \tag{5}$$

We deem the sample consistent if $\hat{y}^T = \hat{y}^S$.

A test sample is written into the *reliable cache* C^+ only when it satisfies both criteria above. This dual gate suppresses unstable yet deceptively confident OOD samples, thereby keeping C^+ clean for subsequent logit calibration.

Remark on Consistency Boundary. It is worth noting that while augmentation consistency is not a theoretically absolute separator, it serves as an effective empirical filter in the robust CLIP feature space. We observe that semantic outliers (OOD) exhibit significantly higher prediction volatility under perturbations compared to ID samples. Our dual-criterion is

designed as a *high-precision* filter rather than a high-recall one: it may conservatively reject "hard" ID samples (e.g., severely occluded or ambiguous images) that fail the consistency check. However, this strict admission rule is crucial for maintaining the purity of the support cache C^+, ensuring that the retrieval-based logit refinement relies solely on high-confidence prototypes rather than noisy signals.

B. Memory-Calibrated Logit Refinement

To exploit reliable test-time evidence without updating the backbone, we employ a lightweight retrieval-based logit calibration module, drawing on the general idea of using test-time memories to refine model outputs. Concretely, we maintain a key–value cache that stores normalized image features $f \in \mathbb{R}^D$ as keys, together with an entropy-based score for maintenance. We organize the cache into two complementary caches: (i) a *support cache* C^+ that accumulates high-reliability exemplars to provide class-consistent evidence, and (ii) an *ambiguity cache* C^- that collects moderately uncertain observations to provide class-wise *regularization cues*. For entries in C^- we additionally store the corresponding softmax distribution $p(x)$, from which we derive class masks.

a) Support evidence from C^+.: Given a test feature f_i, we compute cosine affinities to keys in the support cache,

$$s_i^+ = (C_f^+)^\top f_i, \tag{6}$$

and convert them to non-negative weights via an exponential kernel,

$$w_i^+ = \exp\big(\beta^+(s_i^+ - 1)\big)\,. \tag{7}$$

Let Y^+ denote the matrix of one-hot labels associated with entries in C^+. The retrieved support evidence induces a logit offset through a weighted vote,

$$\Delta l_i^+ = (Y^+)^\top w_i^+, \tag{8}$$

which strengthens classes supported by nearby, high-purity exemplars in the feature space.

b) Regularization cues from C^-.: To further stabilize predictions in ambiguous regions, we optionally introduce a second cache C^- that provides class-wise regularization signals instead of direct supervision. For each item in C^-, we transform its probability vector into a binary class mask by retaining only classes whose probabilities fall into a prescribed interval,

$$m(x) = \mathbf{1}\{\, m_\ell < p(x) < m_u \,\} \in \{0,1\}^C. \tag{9}$$

Stacking these masks forms M^-. Using the same affinity-to-weight mapping for keys in C^-, we obtain

$$\Delta l_i^- = (M^-)^\top w_i^-, \quad w_i^- = \exp\big(\beta^-(s_i^- - 1)\big)\,, \tag{10}$$

which discourages classes that are repeatedly (and weakly) activated by similar uncertain observations.

c) Final calibrated logits.: We combine the retrieved support evidence and the regularization cues to produce the adapted logits:

$$l_i' = l_i + \alpha^+ \Delta l_i^+ - \alpha^- \Delta l_i^-, \tag{11}$$

where (α^+, β^+) and (α^-, β^-) control the strength and sharpness of the two contributions.

d) Capacity control and update policy.: To prevent drift and limit computation, we cap the cache size per predicted class by K and prioritize entries with lower predictive entropy, which empirically correlates with higher reliability. The ambiguity cache C^- is updated only when the normalized entropy falls within a moderate range, avoiding both overly confident samples and highly uncertain samples that may inject noise into the masks. We follow the generic cache-style maintenance protocol in prior work [9], [10] while adapting it to our purified admission rule.

C. Online Adaptive OOD Detection

In line with recent open-set TTA practice, we employ a lightweight linear detector to partition the incoming stream into ID and OOD samples. A key aspect of our framework is the dynamic feedback loop used to generate these pseudo-labels. For each incoming sample x_i, the model first produces a cache-calibrated logit l_i'. Let $S_{msp}(l) = \max(\text{softmax}(l))$ be the maximum softmax probability confidence score for a given logit vector l. The binary pseudo-label y_i^{pse} is then generated relative to an adaptive threshold λ:

$$y_i^{pse} = \begin{cases} 1 & \text{if } S_{msp}(l_i') \geq \lambda \\ 0 & \text{if } S_{msp}(l_i') < \lambda \end{cases} \tag{12}$$

The net is then trained on the image features f_i and their corresponding pseudo-labels y_i^{pse} using a standard cross-entropy loss as follows:

$$\mathcal{L}_{ood}(\theta) = \text{CE}(g_\theta(f_i),\, y_i^{pse})\,, \tag{13}$$

where $g_\theta(\cdot)$ denotes a lightweight OOD detector parameterized as a single linear classifier on top of the frozen image features, i.e., $g_\theta(f) = Wf + b$, producing two-way logits for $\{\text{ID}, \text{OOD}\}$.

To ensure the pseudo-labeling process is robust to varying test data, the threshold λ is not fixed but is adapted dynamically. The adaptive threshold λ is updated dynamically based on recent confidence scores $\mathcal{S}$. Following the strategy from OWTTT [17], the optimal threshold λ^* is found by searching for the value that minimizes the weighted sum of the variances of the two resulting subsets of scores partitioned by a candidate threshold λ:

$$\lambda^* = \arg\min_\lambda \big(w_{id}(\lambda)\sigma_{id}^2(\lambda) + w_{ood}(\lambda)\sigma_{ood}^2(\lambda)\big) \tag{14}$$

where $w_{id/ood}(\lambda)$ and $\sigma_{id/ood}^2(\lambda)$ represent the proportions and variances of the score subsets above and below λ, respectively.

The linear net complements cache calibration by learning a progressively cleaner ID/OOD separator from l_i'-guided

TABLE I

COMPARISON RESULTS ON OPEN-SET DATASETS. BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED BY **BOLD** AND UNDERLINE, RESPECTIVELY. OUR METHOD IS SHADED.

ID	Method	iNaturalist			SUN			Texture			Places			Avg		
		Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
CUB-200-2011	ZS-CLIP	38.13	88.06	53.22	38.10	87.86	53.15	37.56	79.11	50.94	38.00	87.81	53.04	37.95	85.71	52.59
	Tent	37.02	46.95	41.40	38.61	55.55	45.56	34.98	41.77	38.07	40.41	74.83	52.48	37.75	54.78	44.38
	TPT	37.41	89.57	52.78	37.49	89.67	52.87	36.88	81.67	50.81	37.44	89.45	52.79	37.30	87.59	52.31
	AdaND	52.34	96.40	67.84	52.41	**93.91**	67.27	51.82	81.24	63.28	51.82	91.51	66.17	52.10	90.77	66.14
	AIM(Ours)	**56.12**	**96.78**	**71.04**	**56.53**	91.32	**69.83**	**55.93**	**94.04**	**70.14**	**56.40**	**95.96**	**71.04**	**56.25**	**94.53**	**70.52**
Food-101	ZS-CLIP	80.60	94.76	87.11	80.75	96.08	87.75	80.51	93.12	86.36	80.62	94.62	87.06	80.62	94.65	87.07
	Tent	75.83	25.09	37.70	82.86	85.10	83.97	82.54	87.03	84.73	82.26	80.13	81.18	80.87	69.34	71.90
	TPT	79.70	94.93	86.65	79.92	96.19	87.30	79.70	93.86	86.20	79.76	95.14	86.77	79.77	95.03	86.73
	AdaND	86.50	**99.87**	92.71	86.40	99.64	92.55	86.44	96.51	91.20	86.42	**99.40**	92.46	86.44	**98.85**	92.23
	AIM(Ours)	**86.87**	99.59	**92.80**	**86.84**	**99.66**	**92.81**	**86.70**	**96.59**	**91.38**	**86.71**	99.36	**92.60**	**86.75**	98.79	**92.40**
Oxford-IIIT Pet	ZS-CLIP	78.58	88.30	83.16	79.75	87.30	83.35	80.20	91.16	85.33	79.59	84.17	81.82	79.53	87.73	83.41
	Tent	80.07	78.09	79.07	81.19	68.30	74.19	81.48	74.72	77.95	80.64	62.51	70.43	80.84	70.91	75.41
	TPT	77.56	89.71	83.19	78.87	89.82	83.99	79.17	92.26	85.22	78.62	87.32	82.74	78.56	89.78	83.78
	AdaND	85.81	98.78	91.84	85.82	98.19	91.59	85.86	98.68	91.82	85.88	96.58	90.92	85.84	98.06	91.54
	AIM(Ours)	**86.43**	**99.33**	**92.43**	**86.91**	**98.74**	**92.45**	**86.77**	**99.12**	**92.53**	**86.88**	**97.97**	**92.09**	**86.75**	**98.79**	**92.38**
ImageNet	ZS-CLIP	54.01	86.53	66.51	53.43	83.96	65.30	52.71	78.52	63.08	53.35	80.50	64.17	53.38	82.38	64.77
	Tent	48.56	35.74	41.18	55.44	75.54	63.95	54.94	70.93	61.92	55.76	73.98	63.59	53.67	64.05	57.66
	TPT	52.58	88.93	66.09	51.91	86.09	64.77	51.11	80.01	62.38	51.80	82.89	63.76	51.85	84.48	64.25
	AdaND	63.26	**96.87**	76.54	61.34	89.44	72.77	62.45	83.54	71.47	61.92	84.82	71.58	62.24	88.67	73.09
	AIM(Ours)	**65.44**	96.18	**77.89**	**61.53**	**89.57**	**72.95**	**62.61**	**84.64**	**71.98**	**61.97**	**86.71**	**72.28**	**62.89**	**89.28**	**73.77**
ImageNet-V2	ZS-CLIP	48.01	85.72	61.55	47.37	83.23	60.38	46.81	77.54	58.38	47.39	79.41	59.36	47.39	81.47	59.92
	Tent	47.94	76.98	59.08	48.28	80.50	60.36	47.56	74.47	58.05	48.34	77.37	59.50	48.03	77.33	59.25
	TPT	46.63	88.37	61.05	46.12	85.58	59.94	45.21	79.14	57.55	46.02	81.95	58.94	46.00	83.76	59.37
	AdaND	56.32	**97.06**	71.28	54.78	86.64	67.12	**57.28**	80.61	66.97	**55.81**	79.24	65.49	56.05	85.89	67.72
	AIM(Ours)	**57.91**	96.21	**72.30**	**54.83**	**91.44**	**68.55**	56.85	**82.89**	**67.44**	54.47	**86.75**	**66.92**	**56.02**	**89.32**	**68.80**

pseudo-labels, which further stabilizes the overall adaptation loop.

Synergy for Cold-Start Adaptation. A potential challenge in online learning is the *cold-start phase*, where the parametric linear detector g_θ is under-trained due to sparse observations. AIM addresses this through a parametric/non-parametric handover mechanism. In the early stages, the system relies primarily on the retrieval-based cache (C^+), which requires no gradient updates and provides immediate, robust logit refinement to generate reliable pseudo-labels. These high-quality pseudo-labels then "warm up" the linear detector on the fly. As adaptation proceeds and g_θ converges, it progressively enhances the ID/OOD separation boundary, forming a mutually reinforcing loop that stabilizes performance from the very first batch.

V. EXPERIMENTS

A. Experiment Setup

The evaluation is designed to rigorously demonstrate AIM's effectiveness in open-set TTA through comprehensive comparisons with state-of-the-art methods and detailed analyses under standardized metrics.

Datasets. We conduct experiments on a diverse suite of established ID–OOD dataset pairs to cover both fine-grained and large-scale recognition tasks. The ID datasets include CUB-200-2011 [18], Oxford-IIIT Pet [19], and Food-101 [20], as well as large-scale datasets such as ImageNet [21] and ImageNet-V2 [22]. The corresponding OOD datasets are drawn from iNaturalist [23], SUN [24], Places [25], and Texture [26], which together represent a broad spectrum of semantic and distributional shifts.

Comparison Methods. We benchmark AIM against three categories of baselines. (i) The non-adapting zero-shot CLIP (ZS-CLIP) [1], which provides a lower bound without adaptation. (ii) TTA approaches, including Tent [2] and TPT [4], which represent entropy-minimization and prompt-based adaptation strategies, respectively. (iii) The state-of-the-art open-set TTA method AdaND [16], which decouples classification and detection via a frozen backbone and trainable detector.

Evaluation Metrics. To holistically assess adaptation and robustness, we adopt three complementary metrics: ID classification accuracy (Acc_S), OOD detection accuracy (Acc_N), and their harmonic mean (Acc_H). Acc_S measures correct classification of ID samples, Acc_N quantifies successful rejection

of OOD samples, and Acc_H balances the two. Formally:

$$\mathrm{Acc_S} = \frac{\sum_{(x,y)\in\mathcal{D}_{\text{test}}} \mathbf{1}\{\hat{y} = y,\, y \in \mathcal{Y}_{\text{id}}\}}{|\{y \in \mathcal{Y}_{\text{id}}\}|}, \tag{15}$$

$$\mathrm{Acc_N} = \frac{\sum_{(x,y)\in\mathcal{D}_{\text{test}}} \mathbf{1}\{\hat{y} \in \mathcal{Y}_{\text{ood}},\, y \in \mathcal{Y}_{\text{ood}}\}}{|\{y \in \mathcal{Y}_{\text{ood}}\}|}, \tag{16}$$

$$\mathrm{Acc_H} = \frac{2 \cdot \mathrm{Acc_S} \cdot \mathrm{Acc_N}}{\mathrm{Acc_S} + \mathrm{Acc_N}}. \tag{17}$$

B. Experimental Settings

We instantiate the augmentation as horizontal flip and average the probabilities from the original and augmented views. It is optimized in an online manner using the Adam optimizer with a learning rate of 5×10^{-4}, weight decay 0, and momentum $\beta_1 = 0.9$. The detector minimizes the binary cross-entropy loss derived from the pseudo-labels at each test step.

To ensure robust pseudo-labeling, we maintain a queue $\mathcal{S}$ of size $N = 64$, which stores the most recent cache-calibrated confidence scores. Regarding the cache structure, we set the per-class shot capacities to $K^+ = 3$ and $K^- = 2$. The logit refinement parameters are configured as $(\alpha^+, \beta^+) = (2.0, 5.0)$ and $(\alpha^-, \beta^-) = (0.117, 1.0)$, respectively.

The Linear OOD Detector is updated in a batch-wise manner using a FIFO queue (size $Q = 64$). Gradients are computed and backpropagated only when the queue is full, ensuring stable optimization of the decision boundary and mitigating the noise from individual pseudo-labels.

C. Main Results

Table I summarizes the performance of AIM and competing methods across five ID–OOD pairs. AIM consistently establishes new state-of-the-art results, achieving the highest Acc_H scores on every benchmark.

On the fine-grained CUB-200-2011 dataset, AIM achieves an Acc_H of **70.52%**, surpassing AdaND by +4.38 points. This gain is jointly driven by an Acc_S of 56.25% and an Acc_N of 94.53%, indicating that AIM enhances ID accuracy without sacrificing OOD rejection. AIM also delivers strong improvements on Food-101 and Oxford-IIIT Pet, reaching Acc_H scores of 92.40% and 92.38%, respectively. On large-scale ImageNet and ImageNet-V2, AIM maintains superiority with 73.77% and 68.80%, confirming the scalability of the augmentation-invariance principle beyond fine-grained recognition. AIM purifies the adaptation signal at its origin by enforcing dual criteria on cache admission. This prevents error propagation and ensures robust performance.

D. Ablation Study

a) Component Effectiveness.: To quantify the contribution of each component in AIM, we perform a controlled ablation on CUB-200-2011 and report the average Acc_H across its four OOD datasets (Table II). We evaluate four configurations: (1) the zero-shot CLIP baseline (ZS-CLIP); (2) ZS-CLIP with cache boosting; (3) ZS-CLIP with cache boosting and the linear network but without the dual-criterion; and (4) the complete AIM. To isolate the benefit of the filtering mechanism from the benefit of simply viewing augmented samples, our baseline 'w/o Dual-Criterion' utilizes the averaged prediction of the two views but accepts all samples. The significant gain of AIM confirms that the improvement stems from *selective adaptation*, not just multi-view ensembling.

The zero-shot CLIP baseline achieves 52.59%. Adding a purely confidence-based cache yields only a marginal gain to 52.79% (+0.2), confirming that an unscreened cache is ineffective in open-set settings. Incorporating the lightweight Linear Net increases performance by +12.9 points (65.72%), highlighting the benefit of adaptive ID–OOD separation. Finally, enabling the dual-criterion mechanism produces a further +4.8 improvement, bringing Acc_H to **70.52%**.

This progression isolates the role of cache purification. Without augmentation invariance, cached OOD samples contaminate adaptation. With AIM's dual-criterion, the cache remains clean and reliable, enabling robust logit refinement and state-of-the-art adaptation.

b) Sensitivity to Cache Capacity: We further analyze the sensitivity of AIM to the cache size hyperparameters (K^+, K^-) in Fig. 4. The results indicate that performance is relatively robust to variations in shot capacity. While extremely small capacities limit the adaptation signal, and excessively large ones may introduce noise, a moderate setting consistently yields optimal open-set performance (Acc_H), demonstrating that our method does not require meticulous hyperparameter tuning.

c) Robustness of Augmentation Strategy.: Finally, we empirically determined the optimal augmentation strategy through extensive testing on four datasets. As illustrated in the comparative results of Fig. 5, among the evaluated transformations (Affine, CenterCrop, GaussianBlur, and Horizontal Flip), the *Horizontal Flip* strategy demonstrates the most stable performance across all datasets. Consequently, we adopt it as the default configuration for our consistency check module.

TABLE II
ABLATION ON CUB-200-2011 (AVERAGE Acc_H OVER FOUR OOD DATASETS). Δ DENOTES THE ABSOLUTE GAIN OVER THE ZERO-SHOT CLIP BASELINE.

Cache	Linear Net	Dual-Criterion	Avg. Acc_H (%)	Δ (%)
×	×	×	52.59	+0.00
✓	×	×	52.79	+0.20
✓	✓	×	<u>65.72</u>	+13.13
✓	✓	✓	**70.52**	**+17.93**

VI. LIMITATIONS AND DISCUSSION

While AIM establishes a new state-of-the-art in open-set TTA, we acknowledge two primary boundaries. First, the *augmentation consistency assumption*, while effective for semantic outliers, strictly filters the adaptation stream. Consequently, difficult ID samples that are semantically correct but visually unstable under augmentation might be rejected, potentially capping the recall for fine-grained recognition tasks. Second,

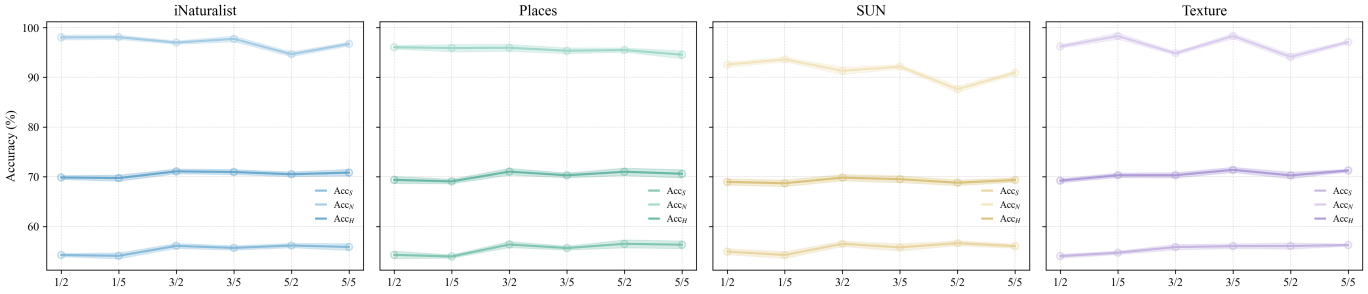

Fig. 4. Cache capacity sensitivity on four pairs (CUB-200-2011 as ID; iNaturalist, Places, SUN, and Texture as OOD). We vary the per-class shot capacities of the cache (K^+, K^-) and report open-set performance, with Acc_H as the primary metric. Shaded regions denote the fluctuation range across two experimental runs.

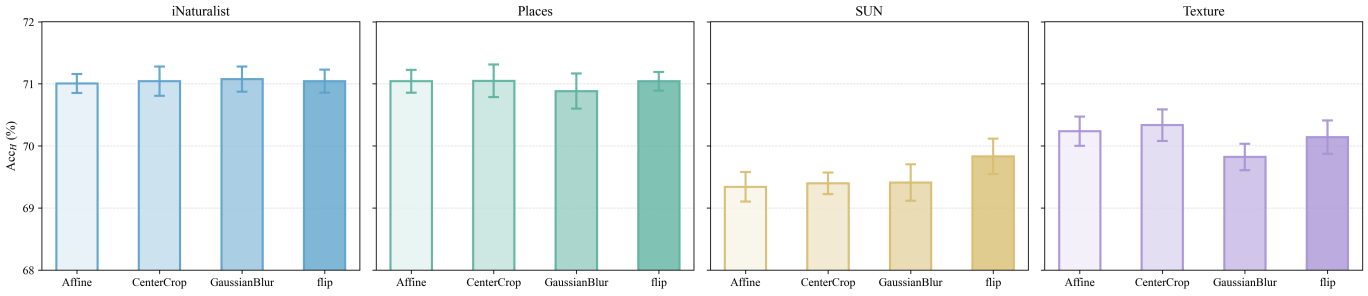

Fig. 5. Impact of augmentation strategies. We compare the robustness of different transformations (Affine, Crop, Blur, Flip) across four pairs(CUB-200-2011 as ID; iNaturalist, Places, SUN, and Texture as OOD). Horizontal flip consistently yields the highest stability (Acc_H), validating its selection for the consistency check.

our framework relies on the inherent separability of the CLIP feature space. For *near-OOD* categories that share high semantic overlap with ID classes (e.g., distinct but visually similar breeds), the dual-criterion might face ambiguity. Future work could address these by incorporating geometric constraints or contrastive learning objectives to sharpen the decision boundary in dense semantic regions.

VII. CONCLUSION

We introduce Augmentation-Invariant Memory (AIM), a robust open-set test-time adaptation framework that purifies online cache through a dual-criterion mechanism based on confidence and augmentation consistency. The purified cache is leveraged to refine prediction logits and to train a lightweight head for ID/OOD separation, establishing a virtuous cycle of adaptation and detection. Evaluated on 20 ID-OOD pairs, AIM advances the state-of-the-art harmonic accuracy from 66.1% to 70.5%, offering a model-agnostic solution for deploying vision-language models in open-world settings.

We believe AIM can be readily extended to broader settings by integrating stronger augmentation sets and more principled online criteria for memory management, which we leave for future work.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[2] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *ICLR*, 2021.

[3] J. Liang, R. He, and T. Tan, "A comprehensive survey on test-time adaptation under distribution shifts," *International Journal of Computer Vision*, vol. 133, no. 1, pp. 31–64, 2025.

[4] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," in *NeurIPS*, 2022.

[5] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li, "Delving into out-of-distribution detection with vision-language representations," in *NeurIPS*, 2022.

[6] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, "Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection," *International Journal of Computer Vision*, vol. 133, no. 6, pp. 3586–3596, 2025.

[7] M. Zhang, S. Levine, and C. Finn, "Memo: Test time robustness via adaptation and augmentation," *Advances in neural information processing systems*, vol. 35, pp. 38 629–38 642, 2022.

[8] J. Abdul Samadh, M. H. Gani, N. Hussein, M. U. Khattak, M. M. Naseer, F. Shahbaz Khan, and S. H. Khan, "Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 80 396–80 413. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/fe8debfd5a36ada52e038c8b2078b2ce-Paper-Conference.pdf

[9] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *European conference on computer vision*. Springer, 2022, pp. 493–510.

[10] A. Karmanov, D. Guan, S. Lu, A. El Saddik, and E. Xing, "Efficient test-time adaptation of vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 162–14 171.

[11] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *CoRR*, vol.

abs/1610.02136, 2016. [Online]. Available: http://arxiv.org/abs/1610.02136

[12] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *NeurIPS*, 2020.

[13] H. Wang, Y. Li, H. Yao, and X. Li, "Clipn for zero-shot ood detection: Teaching clip to say no," in *ICCV*, 2023.

[14] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," in *International conference on machine learning*. PMLR, 2022, pp. 16 888–16 905.

[15] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan, "Towards stable test-time adaptation in dynamic wild world," *arXiv preprint arXiv:2302.12400*, 2023.

[16] C. Cao, Z. Zhong, Z. Zhou, T. Liu, Y. Liu, K. Zhang, and B. Han, "Noisy test-time adaptation in vision-language models," in *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Y. Li, X. Xu, Y. Su, and K. Jia, "On the robustness of open-world test-time training: Self-training with dynamic prototype expansion," in *ICCV*, 2023.

[18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[19] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *CVPR*, 2012.

[20] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *ECCV*, 2014.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[22] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *ICML*, 2019.

[23] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *CVPR*, 2018.

[24] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010.

[25] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[26] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *CVPR*, 2014.