RoAR: Robust Test-Time Refinement with Unknown Priors for Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Test-time adaptation improves the zero-shot predictions of vision-language models using only the unlabeled test pool, and its gains are measured on pools whose priors are fixed: classes are roughly balanced, every class in the label set is present, and every image belongs to the label set. A deployed model receives a pool whose priors are unknown. We show that strong gains on the original benchmark pools need not persist when their class composition changes. The strongest evaluated baseline on the original pools falls below zero-shot in mean accuracy at the strongest tested level of class imbalance, class absence, and open-set contamination. We release an evaluation protocol that turns the class composition of a test pool into controlled prior axes, and we propose **RoAR** (Robust Anchored Refinement), a transductive test-time refinement for pools with unknown priors. It pairs a vision-language model with a self-supervised encoder, screens potentially unknown images before aggregation, and holds a selectively reweighted zero-shot anchor fixed during refinement. In two-view pooling, anchor confidence sets the total prototype exponent, while geometric scores determine the relative weights of the prototype views. With a single configuration shared across all datasets and prior settings, RoAR matches the strongest existing method on the original pools, improving over zero-shot by 7.7 percentage points on ImageNet and ten cross-domain datasets. Under severe class imbalance or absent classes, where that method drops below zero-shot, RoAR outperforms it by about 8 percentage points.

1 Introduction

Test-time adaptation (TTA) refines the zero-shot predictions of vision-language models (VLMs) using only unlabeled test images, and recent methods report steady gains. Online methods estimate class-conditional statistics from the test stream (Dai & Yang, 2025; Zhang et al., 2025b; Zhou et al., 2025; Han et al., 2025), and transductive methods classify the whole test pool jointly (Saha et al., 2025; Zanella et al., 2024). These gains, however, are measured on benchmark test sets whose priors are fixed and known in advance, and a deployed model receives no such guarantee. But where do their gains come from, and do they persist when the priors of the test pool change?

Most of these methods draw correction evidence from the VLM's own embedding space, and progress has been uneven. TTA-VLM (Sheng et al., 2025) re-evaluates fifteen methods under a unified protocol and finds limited improvements over early prompt tuning (Shu et al., 2022) and less consistent gains on SigLIP (Zhai et al., 2023). To address this inconsistency, recent methods, such as SOTA (Hu et al., 2026) and COSMIC (Huang et al., 2025), draw on complementary visual representations. They combine their VLMs with pretrained DINOv2 (Oquab et al., 2024), a self-supervised encoder, and obtain substantial out-of-distribution gains. However, they remain untested whether this additional evidence survives changes in the test pool without requiring a new configuration for each setting.

At deployment, the class frequencies may be skewed (class imbalance), some classes in the candidate label set may never appear (absent classes), and some images may belong to no candidate class, which we call *unknown* images (open-set pools), as illustrated in Figure 1. These deployment conditions can undermine existing mechanisms developed for TTA. For example, fixed balancing constraints can impose incorrect class frequencies, while confidence-filtered caches can still admit erroneous predictions on unknown images. Dataset-specific settings offer no assurance that gains will persist without retuning. In our evaluation (see Figure 2), even the strongest baseline on the class-balanced test pools falls below zero-shot under these deployment conditions. These results highlight the need for robustness to unknown class priors beyond balanced test pools.

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.

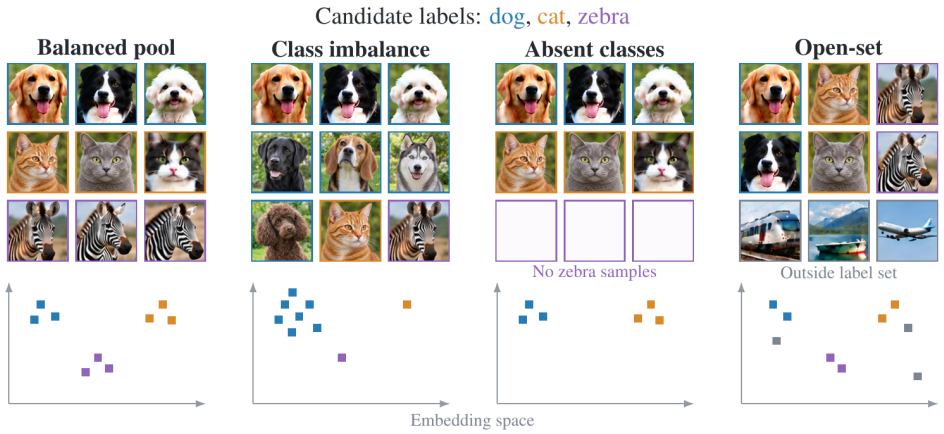


Figure 1: **Test pools under unknown priors.** All four pools share the candidate label set $\mathcal{Y} = \{\text{dog, cat, zebra}\}$ and differ only in their prior $(\mathcal{P}, \pi, \varepsilon)$ (Sec. 2). Each image is outlined by its true class (grey: outside $\mathcal{Y}$), and the bottom row places the same images in a schematic embedding space. **Balanced pool:** the benchmark setting, in which every candidate class appears in roughly equal proportion and every image belongs to $\mathcal{Y}$. **Class imbalance:** the known-class prior π is skewed, here seven dogs against one cat and one zebra. **Absent classes:** zebra stays in the label set but has no images in the pool, so $\mathcal{P} \subsetneq \mathcal{Y}$. **Open-set:** a third of the pool (a train, a boat, and an airplane) lies outside $\mathcal{Y}$, so $\varepsilon > 0$, yet closed-set zero-shot inference still labels each of these images as dog, cat, or zebra. Methods built around the balanced pool can then impose wrong class frequencies, assign predictions to the absent class, or absorb unknown images into the class statistics they estimate from the pool. The deployed model sees only the unlabeled pool and is not told which of these settings, or which combination of them, applies.

To enhance this robustness, we propose **RoAR** (Robust Anchored Refinement), a test-time refinement for VLMs under unknown priors. The VLM supplies the zero-shot predictions, while a complementary self-supervised encoder supplies pool-derived prototype predictions. We form an anchor from the zero-shot predictions, reweight it once when predictions on the retained pool are concentrated, and hold it fixed throughout prototype refinement. In two-view pooling, anchor confidence sets the total prototype exponent, while geometric scores allocate it between the raw and residual views. When neither view has positive strength, a whitened view is used if pseudo-label coverage is sufficient, and otherwise the output is the anchor. Images that may belong to none of the candidate classes are screened before aggregation. No information about the prior is supplied, and one configuration is fixed across all evaluated datasets and prior settings.

Our contributions are:

- **Evaluation under controlled priors.** An evaluation protocol that turns the class composition of a test pool into controlled prior axes, with the three-axis pool construction, draws, and seeds released so that any method can be tested under the same prior settings.
- **Gains depend on the test-pool prior.** The finding that strong gains on the original pools do not ensure robustness to changes in pool composition, with the strongest evaluated baseline falling below zero-shot in mean accuracy at the strongest tested level of each axis.
- **Robust refinement with one configuration.** RoAR, which on ImageNet and ten cross-domain datasets is on par with the strongest baseline on the original pools, improving mean accuracy over zero-shot by 7.7 percentage points, and leads it by about 8 percentage points at the strongest level of the two class axes.

2 Where do test-time gains go when the pool changes?

Test-pool priors. Let $\mathcal{Y} = \{y_1, \ldots, y_C\}$ be the candidate label set and $\mathcal{X} = \{x_1, \ldots, x_N\}$ an unlabeled test pool. An image is *known* if its true class belongs to $\mathcal{Y}$ and *unknown* otherwise. We describe the class composition of a pool by three quantities, which together form its *prior*. The *present set* $\mathcal{P} \subseteq \mathcal{Y}$ contains the candidate classes represented in the pool, the *known-class prior* π gives their relative frequencies among known images, and the *unknown fraction* ε is the share of unknown images. Class imbalance corresponds to a non-uniform π, absent classes give $\mathcal{P} \subsetneq \mathcal{Y}$, and open-set pools have $\varepsilon > 0$. Unknown images have no correct answer in $\mathcal{Y}$, yet closed-set zero-shot inference still assigns them a candidate label. They can therefore supply misleading evidence to methods that aggregate predictions across images.

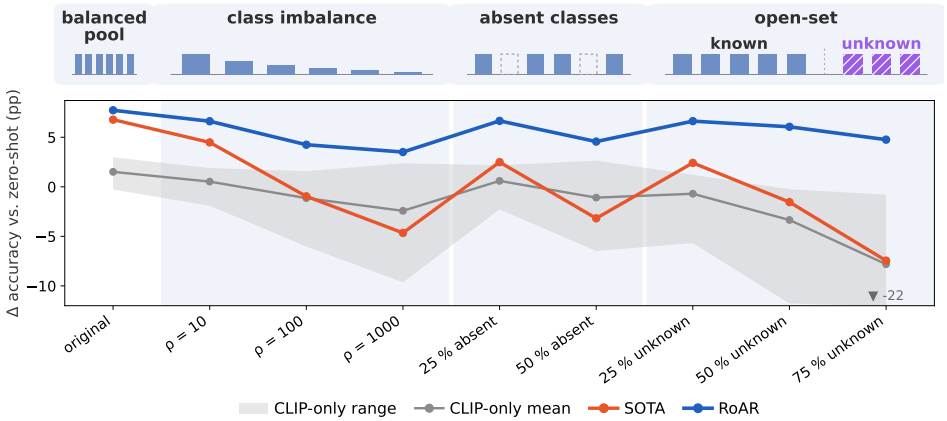


Figure 2: **Adaptation gains depend on the test-pool prior.** Accuracy change over CLIP (Radford et al., 2021) zero-shot in percentage points, averaged over ImageNet (Deng et al., 2009) and ten cross-domain datasets: Caltech101 (Fei-Fei et al., 2004), Oxford Pets (Parkhi et al., 2012), Stanford Cars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), FGVC Aircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), and UCF101 (Soomro et al., 2012). The original test pools are modified along three axes. Open-set pools are scored on the known images, with a rejected known image counted as an error. The orange line shows the state-of-the-art method SOTA (Self- adaptive Optimal Transport; Hu et al. (2026)). The grey line and band show the mean and the minimum to maximum range across eight CLIP-only TTA methods: TransCLIP (Zanella et al., 2024), SOTA with CLIP features, ADAPT (Zhang et al., 2025b), CLIP prototypes, ZLaP (Stojnić et al., 2024), GDA (Wang et al., 2024), StatA (Zanella et al., 2025), and LAME (Boudiaf et al., 2022).

None of $\mathcal{P}$, π, or ε is a property of an individual image, so the pool is the smallest unit at which the prior can be specified, varied, or evaluated. Our protocol therefore defines every setting at the pool level and scores each pool as a whole, on the known images, with a rejected known image counted as an error. Online and transductive methods are evaluated on the same pools and by the same metric. An online method consumes each pool as a stream in a fixed random order per draw, and both families appear among our baselines, with OnZeta (Qian & Hu, 2024) and ADAPT (Zhang et al., 2025b) online and TransCLIP (Zanella et al., 2024), StatA (Zanella et al., 2025), and SOTA (Hu et al., 2026) transductive (Table D.1). This equalizes the evaluation, not the information available. RoAR is transductive and observes $\mathcal{X}$ jointly. It receives no target labels, no information about the prior, and no per-dataset tuning.

Sensitivity to the prior. Methods that aggregate evidence across test images can depend on how that evidence is distributed. A prescribed class marginal can conflict with the pool's class frequencies or allocate mass to absent classes. Caches built from VLM-generated pseudo-labels, as in TDA (Karmanov et al., 2024), DMN (Zhang et al., 2024b), and COSMIC (Huang et al., 2025), can retain confident errors on unknown images despite confidence filtering, allowing those errors to influence subsequent predictions. Figure 2 examines performance across these prior settings. SOTA (Hu et al., 2026), the strongest evaluated baseline on the original pools, gains 6.8 points there against 7.7 for RoAR and keeps a positive mean gain only at the mildest level of each axis, 4.5 at $\rho = 10$, 2.5 with a quarter of the classes absent, and 2.4 with a quarter of the pool unknown. It is already below zero-shot at $\rho = 100$ and with half of the pool unknown, and at the strongest level of the two class axes it loses 4.6 and 3.2 points. RoAR retains non-negative mean gains throughout the evaluated settings. Access to complementary visual features alone therefore does not ensure robustness to unknown priors.

Toward robustness to unknown priors. These observations motivate combining pool-derived predictions with an anchor derived from the zero-shot predictions and held fixed during refinement. Geometric evidence should guide the relative weighting of complementary feature views. This refinement should neither impose uniform class frequencies nor require every candidate class to be present. Potentially unknown images should also be screened before contributing to refinement. Comparisons with additional class names can provide evidence that a cluster is better explained outside $\mathcal{Y}$. We develop a method around these principles and evaluate whether one fixed configuration can preserve gains on the original pools while avoiding negative mean gains across the tested prior settings.

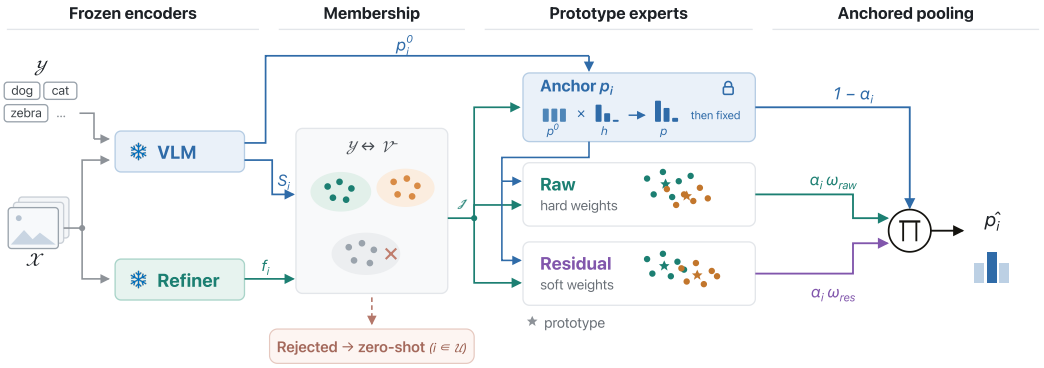


Figure 3: Our proposed method **RoAR** (Robust Anchored Refinement): Membership screening determines the retained pool. Its predicted-label histogram adjusts p_i^0 into an anchor p_i that is held fixed during refinement. Raw and residual prototype experts are pooled with this anchor. Confidence sets their total coefficient and geometry their relative weights. Rejected images $i \in \mathcal{U}$ retain p_i^0.

3 METHOD

We introduce RoAR, a transductive refinement for pools with unknown priors that screens potentially unknown images and pools prototype predictions from a self-supervised encoder with a zero-shot anchor held fixed during refinement. Figure 3 illustrates this pipeline, from membership screening and anchor formation to prototype experts and anchored pooling, and Sections 3.1–3.4 describe these stages in turn.

Given candidate labels $\mathcal{Y} = \{y_1, \ldots, y_C\}$ and an unlabeled pool $\mathcal{X} = \{x_1, \ldots, x_N\}$ with an unknown prior, RoAR combines two frozen models, a base and a refiner. The **base** VLM supplies a zero-shot distribution $p_i^0 \in \Delta_C$ over the C classes, a unit image embedding u_i, and unit text embeddings t_c. The **refiner**, a self-supervised encoder (DINOv2 in our experiments), supplies a unit visual embedding $f_i \in \mathbb{S}^{d-1}$. Our design uses pool-derived prototypes as corrections to a base-derived anchor. The anchor is adjusted once from the retained pool and then held fixed throughout prototype refinement. We first screen potentially unknown images, then account for uneven predicted class support when forming the anchor. Two complementary feature views supply prototype evidence, which is combined with that same anchor. Anchor confidence sets the total prototype coefficient, while geometry allocates it across views. The transductive procedure returns a distribution $\widehat{p}_i$ for every image and the unknown set $\mathcal{U}$ of images it rejects, uses one configuration across the evaluated datasets and prior settings, and receives neither target labels nor information about the prior. Constants are written as symbols, with values in Appendix A.

3.1 SCREENING MEMBERSHIP BEFORE AGGREGATION

Geometric coherence alone does not establish membership in $\mathcal{Y}$. Unknown images can receive confident candidate labels and form coherent clusters, and a prototype estimated on the pool would absorb them. Screening therefore has to answer a question the pool cannot answer about itself, whether an image is explained better by the label set or by the open vocabulary outside it, and it has to answer it for the pool as a whole rather than one image at a time. We combine two sources of evidence that no single image supplies. The first is text. A fixed vocabulary of nouns and adjectives is scored against the candidate text embeddings by the 95th percentile of the cosine similarities, and the lowest-scoring fraction is kept as *negatives*, words the candidate names do not describe (the mining step of NegLabel (Jiang et al., 2024)). Mining uses the class names but no test images and yields M negatives with unit embeddings $n_1, \ldots, n_M$, which a seeded random shuffle partitions into G groups $\mathcal{N}_1, \ldots, \mathcal{N}_G$ of $m = M/G$ words each. The grouping carries no semantic structure, and averaging over groups keeps a few negatives that happen to resemble an image from dominating a single ratio. From these negatives we build a membership score that is meant to be cut at a fixed value, not ranked, so it is formed from means rather than sums and lies on a common scale for every dataset. The second source is the geometry of the refiner space, in which unknown images gather into clusters of their own. The rejection of Eq. (2) is therefore decided per cluster, with an individual veto for images the text evidence clearly retains, so that known images are not discarded one by one. For image i, let A_i be its mean similarity mass over the candidate names and B_{ig} its mean similarity mass over the negatives of group g, both at logit scale s. The membership score is the share the candidate names retain, averaged over groups,

$$A_i = \frac{1}{C}\sum_{c} e^{s u_i^\top t_c}, \qquad B_{ig} = \frac{1}{m}\sum_{j\in\mathcal{N}_g} e^{s u_i^\top n_j}, \qquad S_i = \frac{1}{G}\sum_{g=1}^{G}\frac{A_i}{A_i + B_{ig}}. \tag{1}$$

An image the label set explains well has $A_i \gg B_{ig}$ and S_i near one. S_i is a relative membership score rather than a calibrated probability.

An ambiguous individual score need not indicate an unknown image. We therefore require both cluster-level evidence and an individual veto before excluding an image from refinement. Cosine k-means partitions $\{f_i\}$ into K clusters, and with $\mathcal{A}(i)$ the cluster containing i,

$$\overline{S}_{\mathcal{A}} = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} S_i, \qquad \mathcal{U} = \big\{\, i : \overline{S}_{\mathcal{A}(i)} < \tfrac{1}{2} \text{ and } S_i < \gamma \,\big\}. \tag{2}$$

The cluster cut minimizes a surrogate cost under the score (Appendix A), and the fixed veto γ retains images with strong individual evidence even in low-scoring clusters. Images in the unknown set $\mathcal{U}$ keep $\widehat{p}_i = p_i^0$, and all subsequent statistics use the retained set $\mathcal{I} = \{1,\dots,N\} \setminus \mathcal{U}$ of size n.

3.2 Accounting for uneven predicted class support

Screening addresses membership, but the retained pool can still support the classes unevenly. A class that receives many pseudo-labels yields a prototype averaged over many images, whereas a class that receives a handful yields one that a few erroneous assignments can dominate. We therefore let the predicted class frequencies adjust the anchor once, and only when predictions are concentrated. Let $h \in \Delta_C$ be the histogram of predicted labels over the retained pool, whose entry h_c is the fraction of retained images whose zero-shot prediction is class c, so that h sums to one and a class that receives no prediction has $h_c = 0$. Let $\widetilde{h}$ be its smoothed version with a fixed constant $\delta_h > 0$, which keeps every entry positive, and ν_0 the entropy-effective fraction of classes that receive predictions,

$$h_c = \frac{1}{n}\sum_{i\in\mathcal{I}} \mathbf{1}\{\arg\max_j p_{ij}^0 = c\}, \qquad \widetilde{h}_c = \frac{h_c + \delta_h/C}{1+\delta_h}, \qquad \nu_0 = \frac{\exp H(h)}{C}, \tag{3}$$

where H is the entropy of h before smoothing, so that $\nu_0 = 1$ on a uniform histogram and ν_0 approaches $1/C$ when one class takes every prediction. Concentrated predictions, $\nu_0 < \frac{1}{2}$, trigger the adjustment

$$p_{ic} \propto p_{ic}^0\, \widetilde{h}_c^{\,\eta}, \qquad \eta = \tfrac{1}{2}\,\mathbf{1}\{\nu_0 < \tfrac{1}{2}\}, \tag{4}$$

normalized over classes, which favors classes with greater predicted support. The exponent η sets its strength. Applying the histogram directly, $\eta = 1$, can feed back the errors of p^0 from which h is estimated, so we use its square root, and $\eta = 0$ leaves $p_i = p_i^0$ on a pool whose predictions are spread. The adjustment is not an estimate of the true known-class prior. Once formed, p_i is held fixed throughout refinement.

3.3 Assessing separation in complementary feature views

A prototype is only as good as the separation of the images that form it. In the refiner space, a few directions of variation that have nothing to do with class, such as lighting, background, or viewpoint, can spread the images of one pseudo-class apart and dominate that separation. We therefore build a second view of the pool with those directions removed. To find them, we look at the scatter inside the pseudo-classes rather than the scatter of the whole pool, because the leading directions of the whole pool also carry the differences between pseudo-class means that we want to keep. Let $\widehat{y}_i = \arg\max_c p_{ic}$ be the anchor's pseudo-label, n_c the number of retained images with $\widehat{y}_i = c$, $m_c = \frac{1}{n_c}\sum_{i\in\mathcal{I}:\widehat{y}_i=c} f_i$ the mean refiner embedding of those images, and $\overline{f} = \frac{1}{n}\sum_{i\in\mathcal{I}} f_i$ the mean of the retained pool. The within-pseudo-class scatter collects how far each image sits from its own class mean, and the projection P removes its k leading directions, a fixed number,

$$S_W = \sum_{i\in\mathcal{I}} (f_i - m_{\widehat{y}_i})(f_i - m_{\widehat{y}_i})^\top, \qquad P = I - V_k V_k^\top, \qquad f_i^{\mathrm{res}} = \mathrm{unit}\big(P(f_i - \overline{f})\big), \tag{5}$$

where V_k holds the k leading orthonormal eigenvectors of S_W and $\mathrm{unit}(a) = a/\|a\|$ for nonzero a. The projection has no way of knowing whether a direction it removes also separates the classes, so we keep the untouched view $f_i^{\mathrm{raw}} = f_i$ beside the residual view and never choose between the two.

Instead we measure what the projection cost. The between-pseudo-class scatter S_B collects how far the class means sit from the pool mean, and ϕ is the fraction of it that survives the projection,

$$\phi = \frac{\mathrm{tr}(PS_BP)}{\mathrm{tr}(S_B)} \in [0,1], \qquad S_B = \sum_{c:n_c>0} n_c(m_c - \overline{f})(m_c - \overline{f})^\top. \tag{6}$$

Since $\mathrm{tr}(PS_BP) = \mathrm{tr}(S_B) - \sum_{j\le k} v_j^\top S_B v_j$, the projection lowers ϕ exactly when the class means differ along the directions it removed. A value near one means the class means were left intact, and a small value means the projection removed part of the class signal together with the nuisance, in which case the residual view is discounted by ϕ in Section 3.4.

The two views need not separate the pseudo-classes equally well, and a view that does not separate them should not be trusted with a prototype. We measure separation by asking, for each image, how much closer it sits to the center of its own pseudo-class than to the nearest other one. In view v, let $\bar{c}_c^v$ be the normalized mean of the images assigned to class c, and let a_i^v and b_i^v be the cosine of image i to its own center and to the closest competing centre. The normalized centroid cosine margin (NCCM) averages their normalized difference,

$$M_v = \frac{1}{n}\sum_{i\in\mathcal{I}} \frac{a_i^v - b_i^v}{\max(|a_i^v|, |b_i^v|, 10^{-8})}, \qquad a_i^v = (f_i^v)^\top \bar{c}_{\widehat{y}_i}^v, \qquad b_i^v = \max_{c\neq\widehat{y}_i} (f_i^v)^\top \bar{c}_c^v. \tag{7}$$

It is computed once from the anchor's partition, before any prototype is built. A positive M_v says that images sit closer to their own center on average, not that every image does or that the pseudo-labels are correct. It is used only to weight the two views in Section 3.4.

3.4 Retaining the anchor during prototype refinement

Screening and geometric processing do not eliminate assignment errors, so prototype predictions are combined with the anchor rather than used on their own. In each view, the anchor assigns images to classes, the prototype of a class is the direction of the weighted sum of its images, and an image is scored by its cosine to each prototype,

$$\mu_c^v = \mathrm{unit}\Big(\textstyle\sum_{i\in\mathcal{I}} w_{ic}^v f_i^v\Big), \qquad r_{ic}^v = \mathrm{softmax}_c\big((f_i^v)^\top \mu_c^v/\tau\big), \tag{8}$$

with a fixed temperature τ and nonnegative weights w_{ic}^v derived from the anchor. The residual view builds its prototypes once from p_i, while the raw view, which retains more noise, reassigns confident images to their current best class for one to three rounds (Appendix A). A class that receives no weight keeps a zero prototype and a zero logit, which is not a masked class.

This prediction never replaces the anchor. It is pooled with it in log space at an exponent that grows as the anchor becomes less confident,

$$q_i^v = \mathrm{normalize}\big(p_i^{1-\alpha_i}(r_i^v)^{\alpha_i}\big), \qquad \alpha_i = \tfrac{1}{4} + \tfrac{1}{2}\Big(1 - \max_c p_{ic}\Big), \tag{9}$$

so higher anchor confidence gives the prototype prediction a smaller exponent, and $\alpha_i \in [\frac{1}{4}, \frac{3}{4}]$ keeps both sources active. For positive p_i and r_i^v this is the unique minimizer of $\mathrm{KL}(q\|r_i^v) + \lambda_i \mathrm{KL}(q\|p_i)$ with $\lambda_i = (1-\alpha_i)/\alpha_i$ (Proposition 1).

Finally each view is credited for the structure its margin showed. With $e_{\mathrm{raw}} = [M_{\mathrm{raw}}]_+^2$ and $e_{\mathrm{res}} = (\phi[M_{\mathrm{res}}]_+)^2$, where a non-positive margin gives zero and ϕ discounts the residual view for the structure its projection removed, the two predictions are combined in proportion to these strengths,

$$\widehat{p}_i \propto \prod_v (q_i^v)^{\omega_v} \propto p_i^{1-\alpha_i} \prod_v (r_i^v)^{\alpha_i\omega_v}, \qquad \omega_v = \frac{e_v}{e_{\mathrm{raw}} + e_{\mathrm{res}}}. \tag{10}$$

Because the weights sum to one, the anchor keeps the exponent $1-\alpha_i$ whatever the geometry says. In this pooling rule, confidence therefore sets the total prototype exponent, while geometry determines its allocation across views. When both strengths are zero, neither view has shown structure, and a whitened third view is consulted if the pseudo-labels cover enough of the label set, otherwise the output is the anchor p_i (Appendix A).

4 Experiments

Setup. We use ImageNet (Deng et al., 2009) and ten cross-domain datasets: Caltech101 (Fei-Fei et al., 2004), Oxford Pets (Parkhi et al., 2012), Stanford Cars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), FGVC Aircraft (Maji et al.,

Table 1: **Class imbalance.** Accuracy (%) on ImageNet (IN), the ten cross-domain datasets, and their mean, for CLIP ViT-B/16 with DINOv2-L.

pool	method	IN	Cal	Pets	Cars	Flow	Food	Airc	SUN	DTD	Euro	UCF	mean
original	zero-shot	70.2	95.0	90.8	68.2	73.4	85.6	30.4	64.3	54.6	52.2	71.9	68.8
	TransCLIP	67.0	93.2	90.6	64.7	81.2	83.0	27.0	67.4	54.3	55.3	76.6	69.1
	GDA	70.8	95.4	92.6	71.0	78.1	84.8	31.8	69.2	57.2	59.1	74.7	71.3
	StatA	72.5	95.3	92.2	68.8	77.9	86.4	29.0	69.5	56.3	57.9	76.5	71.1
	COSMIC	72.3	94.1	93.4	69.1	79.9	85.0	30.5	67.4	56.5	58.3	74.8	71.0
	SOTA	78.0	96.3	**94.8**	73.5	83.6	87.3	27.8	72.1	52.7	**85.0**	**80.1**	75.6
	RoAR (Ours)	**80.4**	**97.8**	94.7	**76.0**	**84.8**	**88.1**	**39.4**	**72.5**	**63.2**	65.1	79.6	**76.5**
$\rho = 10$	zero-shot	70.3	95.0	91.7	69.7	73.6	85.8	30.4	64.3	54.5	54.4	72.3	69.3
	TransCLIP	67.2	93.1	91.0	65.7	80.6	82.8	27.5	66.6	54.2	46.9	74.5	68.2
	GDA	70.7	95.4	92.6	71.9	78.3	85.3	31.2	67.7	55.1	60.8	74.0	71.2
	StatA	71.8	95.3	92.0	69.7	78.0	86.3	28.9	68.1	56.6	56.8	74.7	70.7
	COSMIC	70.5	94.2	92.5	69.9	80.4	84.7	30.4	65.2	53.7	54.0	73.1	69.9
	SOTA	76.2	96.3	93.4	74.7	84.3	86.5	28.2	69.7	49.6	**73.4**	**79.0**	73.8
	RoAR (Ours)	**79.4**	**97.8**	**94.3**	**76.8**	**84.9**	**88.2**	**37.1**	**71.3**	**61.1**	65.1	78.9	**75.9**
$\rho = 100$	zero-shot	70.1	95.3	92.4	70.8	72.3	85.8	30.8	64.3	54.6	58.1	73.0	69.8
	TransCLIP	66.0	93.2	90.3	65.6	76.6	81.3	28.2	64.8	52.1	39.7	73.6	66.5
	GDA	70.2	95.6	92.7	72.0	75.5	85.3	31.3	66.5	54.7	**63.4**	73.9	71.0
	StatA	71.1	95.7	93.0	72.0	76.2	85.7	30.2	67.1	55.9	51.1	76.6	70.4
	COSMIC	69.6	94.2	92.7	70.8	77.8	83.2	30.7	64.8	54.3	46.1	73.2	68.9
	SOTA	71.0	93.8	90.6	69.9	79.9	83.3	27.0	65.1	45.4	56.4	74.7	68.8
	RoAR (Ours)	**77.1**	**97.8**	**93.8**	**75.7**	**82.6**	**87.6**	**35.1**	**69.3**	**58.3**	59.5	**77.7**	**74.0**
$\rho = 1000$	zero-shot	70.5	95.8	92.0	70.6	71.0	85.6	30.4	64.2	54.6	56.1	73.8	69.5
	TransCLIP	66.0	91.5	89.8	65.0	74.3	79.4	28.4	63.3	51.2	31.5	71.8	64.8
	GDA	70.5	95.8	92.2	71.5	73.3	85.2	30.8	65.7	54.7	**59.4**	74.4	70.3
	StatA	71.5	96.3	93.5	72.3	75.1	85.2	29.9	66.7	55.8	43.8	**77.4**	69.8
	COSMIC	70.5	94.0	91.9	70.5	76.1	82.6	31.8	63.7	55.0	39.0	73.8	68.1
	SOTA	68.6	87.4	87.4	66.2	74.6	80.2	25.9	62.2	43.1	45.5	72.5	64.9
	RoAR (Ours)	**76.1**	**97.9**	**94.6**	**74.2**	**79.1**	**88.5**	**34.7**	**68.2**	**56.7**	56.2	77.1	**73.0**

2013), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), and UCF101 (Soomro et al., 2012). Pools are drawn without replacement from the test splits of Zhou et al. (2022), and the original pool is the full test split. Long-tailed pools decay the class counts exponentially to $1/\rho$ of the head, $\rho \in \{10, 100, 1000\}$, absent-class pools remove a random quarter or half of the classes from the pool while the full label set is retained, and open-set pools add Places images as unknowns until they form 25, 50, or 75 % of the pool. Every method receives the same pools. Each result is averaged over four seeds and over repeated pool draws. Long-tailed and absent-class pools are drawn ten times per cross-domain dataset and three times on ImageNet. Open-set pools are drawn three times. The base model is CLIP ViT-B/16 and the refiner DINOv2 ViT-L/14. The zero-shot posterior averages sixteen views of each image over the dataset prompt templates. On open-set pools accuracy is measured on the known images, with a rejected known image counted as an error, and Appendix G scores all images with the unknown images as one extra class.

Class imbalance. On the original pool RoAR gains 7.7 points and SOTA 6.8 (Table 1). As ρ grows, SOTA falls to 4.5, -1.0, and -4.6 relative to zero-shot, below it on ten datasets at $\rho = 1000$, while RoAR keeps 6.6, 4.2, and 3.5 and stays above zero-shot on every dataset. The gap in mean accuracy at $\rho = 1000$ is 8.1 points. COSMIC never moves more than 2.2 points from zero-shot, and among the CLIP-only methods of Table 1, GDA is the most stable, 2.5 to 0.8. On EuroSAT, where the balanced assumption holds, SOTA's transport gains 33 points against 13 for the anchored RoAR and then loses 11 at $\rho = 1000$.

Absent classes. A label set is usually fixed for a whole taxonomy, while a single deployment often observes only part of it. SOTA's balancing term must assign the absent classes, so its gain shrinks to 2.5 with a quarter absent and becomes a loss of 3.2 with half absent, below zero-shot on most datasets (Table 2). COSMIC and TransCLIP also fall below zero-shot at that level while GDA and StatA keep 1.4 and 1.7. RoAR gains 6.6 and 4.6 points on average and stays above zero-shot on every dataset.

Open-set. SOTA keeps 71.2 on the known images with a quarter of the pool unknown and falls to 67.3 and 61.4 as the unknown share grows to three quarters, below zero-shot at 68.8 (Table 3). RoAR keeps 75.4, 74.9, and 73.6 and leads on every dataset at 25 and 50 % unknown and on nine of eleven at 75 %. Its rejection step recalls 60 to 69 % of the unknown images while rejecting 0.2 to 0.9 % of the known ones, whereas ZS-NTTA reaches the same recall by rejecting 19 to 26 % of them

Table 2: **Absent classes.** Accuracy (%) when a random quarter or half of the label set is removed from the original pool.

pool	method	IN	Cal	Pets	Cars	Flow	Food	Airc	SUN	DTD	Euro	UCF	mean
25 % absent	zero-shot	70.4	95.0	91.0	68.0	73.6	85.7	30.7	64.4	54.4	53.9	72.0	69.0
	TransCLIP	65.9	89.4	88.8	63.4	78.5	80.5	27.1	66.1	52.7	53.3	73.1	67.2
	GDA	70.8	95.3	92.4	70.5	77.4	84.4	32.0	68.8	56.1	60.9	74.2	71.2
	StatA	72.3	95.1	92.3	69.7	78.4	85.9	29.6	69.0	55.9	59.0	75.8	71.2
	COSMIC	71.1	94.5	92.2	68.5	79.4	83.2	31.5	66.5	54.5	57.8	73.3	70.2
	SOTA	75.1	87.3	91.1	70.7	78.8	84.6	27.5	68.9	50.3	**76.7**	75.4	71.5
	RoAR (Ours)	**79.6**	**97.4**	**93.8**	**74.6**	**83.1**	**87.6**	**38.9**	**71.6**	**61.4**	65.5	**78.8**	**75.6**
50 % absent	zero-shot	69.8	94.8	90.8	68.3	72.8	85.8	30.5	64.2	55.5	51.2	72.0	68.7
	TransCLIP	64.3	87.1	86.6	61.3	74.4	78.6	27.2	63.6	51.6	42.2	67.6	64.1
	GDA	70.1	94.9	91.7	70.3	75.6	84.3	31.8	67.8	56.5	55.6	72.7	70.1
	StatA	71.7	95.3	92.3	71.4	78.0	85.5	30.1	67.9	56.9	50.8	74.5	70.4
	COSMIC	69.7	94.0	92.1	67.9	76.8	82.6	31.1	65.1	56.7	43.5	73.0	68.4
	SOTA	69.8	79.9	85.2	65.0	72.7	80.8	25.7	63.9	47.1	**62.1**	68.6	65.5
	RoAR (Ours)	**77.4**	**96.9**	**92.5**	**73.3**	**80.3**	**87.0**	**37.2**	**69.8**	**60.6**	54.9	**76.1**	**73.3**

Table 3: **Open-set.** Accuracy (%) on the known images when Places images form 25, 50, or 75 % of the pool. A rejected known image counts as an error. H-score is the harmonic mean of the mean known accuracy and the mean unknown recall of each dataset, averaged over the eleven datasets. Methods without a rejection step have zero recall and thus an H-score of 0.0.

pool	method	IN	Cal	Pets	Cars	Flow	Food	Airc	SUN	DTD	Euro	UCF	mean	H-score
25 % unknown	zero-shot	70.5	95.0	90.9	68.1	73.4	85.8	30.4	64.2	54.6	52.1	71.9	68.8	0.0
	TransCLIP	69.8	91.6	87.7	64.5	77.9	79.6	26.6	67.1	53.8	35.8	70.2	65.9	0.0
	GDA	70.4	94.8	90.4	67.6	76.9	85.7	29.8	66.4	57.6	58.1	72.6	70.0	0.0
	StatA	69.9	93.7	87.5	62.9	75.6	85.2	27.2	67.1	57.3	54.0	70.5	68.3	0.0
	COSMIC	68.6	93.5	92.8	66.3	79.9	85.0	28.2	62.8	55.3	47.1	72.8	68.4	0.0
	ZS-NTTA	64.3	88.0	88.4	62.9	67.1	82.2	7.5	57.9	33.9	17.4	65.7	57.8	55.4
	SOTA	74.6	94.6	91.5	71.4	82.2	85.9	26.3	70.2	53.3	57.0	76.4	71.2	0.0
	RoAR (Ours)	**75.7**	**97.7**	**94.7**	**75.1**	**84.1**	**88.3**	**39.3**	**71.1**	**62.9**	**61.7**	**79.1**	**75.4**	**60.4**
50 % unknown	zero-shot	70.5	95.0	90.9	68.1	73.4	85.8	30.4	64.2	54.6	52.1	71.9	68.8	0.0
	TransCLIP	68.6	89.5	83.9	64.4	75.7	74.0	26.6	66.0	53.9	31.1	67.7	63.8	0.0
	GDA	69.8	93.8	88.0	64.1	75.0	84.5	27.4	66.3	57.3	57.1	70.8	68.6	0.0
	StatA	69.5	92.8	84.0	57.1	71.1	83.9	25.6	66.6	53.1	40.8	68.8	64.8	0.0
	COSMIC	68.6	93.7	92.2	66.1	79.1	85.0	28.1	62.8	55.1	37.0	71.8	67.2	0.0
	ZS-NTTA	64.7	88.2	89.2	64.3	68.8	82.8	8.9	58.3	35.3	8.8	66.8	57.8	57.8
	SOTA	74.2	93.1	87.0	67.0	80.8	83.9	22.6	69.3	48.5	41.1	72.4	67.3	0.0
	RoAR (Ours)	**75.2**	**96.9**	**94.6**	**75.1**	**83.4**	**88.3**	**39.3**	**70.8**	**61.9**	**59.4**	**78.5**	**74.9**	**63.0**
75 % unknown	zero-shot	70.5	95.0	90.9	68.1	73.4	85.8	30.4	64.2	54.6	52.1	71.9	68.8	0.0
	TransCLIP	66.2	86.4	74.8	63.2	70.8	62.6	26.7	64.1	50.5	26.4	66.8	59.9	0.0
	GDA	68.5	91.6	76.2	56.1	71.4	79.6	23.1	65.7	55.0	**55.3**	68.2	64.6	0.0
	StatA	68.6	91.3	68.2	49.3	68.2	78.4	21.7	65.7	50.3	22.5	63.0	58.8	0.0
	COSMIC	68.5	93.7	91.2	66.0	78.6	85.0	27.5	62.8	54.6	32.0	70.6	66.4	0.0
	ZS-NTTA	64.8	90.7	89.8	65.0	70.1	83.5	15.4	60.4	41.5	8.2	68.7	59.8	60.1
	SOTA	**73.4**	92.2	80.1	57.4	77.9	79.9	18.2	67.0	40.3	22.1	66.3	61.4	0.0
	RoAR (Ours)	72.7	**95.9**	**94.0**	**75.0**	**83.3**	**88.3**	**39.0**	**69.5**	**60.5**	53.8	**77.3**	**73.6**	**65.3**

and keeps 57.8 to 59.8 (Table G.1), so the harmonic mean H of known accuracy and unknown recall is 60.4, 63.0, and 65.3 for RoAR against 55.4, 57.8, and 60.1 for ZS-NTTA.

Ablation. A naive DINOv2 prototype classifier built from the same pseudo-labels, without screening, anchor adjustment, or the residual view, keeps 0.5 and 1.6 points at $\rho = 1000$ and with 50 % absent and loses 5.3 with 75 % unknown, against 3.5, 4.6, and 4.8 for RoAR (Table 4). Removing the anchor ($\alpha_i = 1$) costs 0.9 to 1.8 points, most under strong skew, and the residual view is worth 0.1 to 1.7 points. The rejection step rejects no image on the class-axis pools and changes nothing there, but with 75 % unknown it adds 1.8 points on the known images alone, as unknown images are kept out of the prototypes.

5 Related Work

Test-time adaptation for VLMs. VLM adaptation methods include prompt tuning in TPT (Shu et al., 2022), C-TPT (Yoon et al., 2024), and DiffTPT (Feng et al., 2023), reward-based updates in RLCF (Zhao et al., 2024), and update-free view voting in ZERO (Farina et al., 2024). TDA (Karmanov et al., 2024), DMN (Zhang et al., 2024b), and BoostAdapter (Zhang et al., 2024a) use test-feature caches, while GDA (Wang et al., 2024), FreeTTA (Dai & Yang, 2025), DOTA (Han et al., 2025), ADAPT (Zhang et al., 2025b), and BCA (Zhou et al., 2025) use statistical models. OnZeta combines online label and proxy learning with a class-balancing constraint (Qian & Hu, 2024). ZLaP (Stojnić

Table 4: **Ablation.** Accuracy change over zero-shot (pp), mean over the eleven datasets. Open-set pools are scored on the known images as in Table 3.

variant	original	$\rho = 1000$	50 % absent	75 % unknown
naive DINOv2 prototype	+4.0	+0.5	+1.6	−5.3
RoAR w/o the rejection step	+7.7	+3.5	+4.6	+3.0
RoAR w/o the residual view	+6.6	+3.4	+4.1	+3.1
RoAR w/o the anchor ($\alpha_i = 1$)	+6.8	+1.8	+2.8	+3.5
RoAR	**+7.7**	**+3.5**	**+4.6**	**+4.8**

et al., 2024), SCAP (Zhang et al., 2025a), GTA-CLIP (Saha et al., 2025), and TransCLIP (Zanella et al., 2024) exploit relationships among unlabeled samples. LAME performs joint output refinement within a batch (Boudiaf et al., 2022). TTA-VLM (Sheng et al., 2025) reports limited progress over early prompt tuning and less consistent gains on SigLIP (Zhai et al., 2023). RoAR complements these approaches with self-supervised prototype evidence and examines whether the resulting gains persist under changes in test-pool composition.

Cross-model refinement with self-supervised features. DINOv2 (Oquab et al., 2024) provides visual representations complementary to VLM embeddings. COSMIC integrates text, CLIP, and DINOv2 features through caches and clique-based graphs to mitigate noisy pseudo-labels (Huang et al., 2025). SOTA combines foundation-model predictions through optimal transport with prescribed marginals and learned model weights (Hu et al., 2026). Complementary representations alone, however, do not ensure robustness to changes in pool composition. RoAR addresses this sensitivity by screening potentially unknown images and combining prototype predictions with a fixed anchor without prescribing class frequencies. In our evaluation, it retains positive mean gains at the strongest tested level of each axis, where SOTA falls below zero-shot.

Adaptation under non-standard test priors. StatA studies variable class coverage and non-i.i.d. batches, anchoring class statistics to text-encoder knowledge (Zanella et al., 2025). Cao et al. (2025) introduce zero-shot noisy test-time adaptation and propose AdaND to detect images outside the label set. Our protocol examines class imbalance, absent classes, and open-set contamination individually and in compound settings, evaluating RoAR with one configuration across all datasets and compositions.

6 Conclusion

This paper studies VLM test-time refinement under unknown priors. Our evaluation shows that strong gains on the original pools do not ensure robustness to class imbalance, absent classes, or open-set pools. The strongest evaluated baseline falls below zero-shot in mean accuracy at the strongest tested level of each axis. We present RoAR, which pools prototype predictions from a self-supervised encoder with a selectively reweighted zero-shot anchor held fixed during refinement. It screens potentially unknown images before aggregation. In two-view pooling, anchor confidence sets the total prototype exponent, while geometric scores determine the relative view weights. With one fixed configuration on ImageNet and ten cross-domain datasets, RoAR matches the strongest baseline on the original pools and gains 7.7 percentage points over zero-shot. It leads that baseline by about 8 percentage points under the strongest class imbalance and absent classes, and it is the only evaluated method above zero-shot on every axis. We release the pool construction, draws, and seeds so that robustness to unknown priors can be evaluated for any method.

Limitations and future work. The fixed anchor trades peak gains for robustness. On balanced or mildly imbalanced pools, a method that imposes a uniform marginal can gain more, as SOTA (Hu et al., 2026) does on EuroSAT (85.0 against 65.1) and, by 0.3 to 1.0 percentage points, with SigLIP backbones (Table E.1); letting geometric evidence also scale the total prototype exponent, not only allocate it across views, may recover part of these gains. RoAR is transductive and runs a second frozen encoder, and extending it to streams, small pools, and lighter refiners is left for future work. Our unknowns come from Places and iNaturalist; those closer to the label set are harder to reject, as scene images are for SUN397 (Table G.2), and near-OOD pools such as held-out classes of the same dataset are a natural next test. The hyperparameters of RoAR are set by hand and transfer unchanged across datasets and three VLMs; the anchor trigger and rejection thresholds matter most (Table F.1), and setting them from the pool itself is left for future work. Finally, the gains over zero-shot are empirical, and characterizing when anchored pooling preserves zero-shot accuracy, as well as testing other prior shapes such as drift within a stream, remains open.

AI USE STATEMENT

In this work, we used generative AI tools to help develop the conceptual framework, to assist in the writing of proofs, to provide feedback on research methodology and experiments, to assist in interpreting results, to implement methods, and to assist with translation. We have not used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design research methodology or experiments, or clean or reformat datasets, and the generation of synthetic data sets and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the text for clarity and consistency and to adjust LaTeX formatting. We have reviewed all AI-assisted work. The conceptual framework and all definitions were revised and finalized by the authors, every step of the proofs was checked by the authors, AI-assisted code was reviewed and tested by the authors, all reported results were reproduced with the final code, and translated text was reviewed by the authors for technical accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461. Springer, 2014.

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*, pp. 8344–8353, 2022.

Chentao Cao, Zhun Zhong, Tongliang Liu, Yang Liu, Kun Zhang, Bo Han, et al. Noisy test-time adaptation in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 59008–59039, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Qiyuan Dai and Sibei Yang. Free on the fly: Enhancing flexibility in test-time adaptation with online em. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9538–9548, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178–178. IEEE, 2004.

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*, pp. 2704–2714, 2023.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 38:142772–142798, 2025.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Zhanxuan Hu, Qiyu Xu, Yu Duan, Yonghang Tai, and Huafeng Li. Sota: Self-adaptive optimal transport for zero-shot classification with multiple foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26624–26634, 2026.

Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, and Zhi Wang. Cosmic: Clique-oriented semantic multi-space integration for robust clip test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9772–9781, 2025.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 14913–14941, 2024.

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*, pp. 14162–14171, 2024.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729. IEEE, 2008.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Qi Qian and Juhua Hu. Online zero-shot classification with clip. In *European Conference on Computer Vision*, pp. 462–477. Springer, 2024.

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*, pp. 8748–8763. PmLR, 2021.

Oindrila Saha, Logan Lawrence, Grant Van Horn, and Subhransu Maji. Generate, transduct, adapt: Iterative transduction with vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1369–1379, 2025.

Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? a critical look at test-time adaptation for vision-language models. *Advances in Neural Information Processing Systems*, 38, 2025.

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*, 35:14274–14289, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Vladan Stojnić, Yannis Kalantidis, and Giorgos Tolias. Label propagation for zero-shot classification with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23209–23218. IEEE, 2024.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free clip-based adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 28841–28859, 2024.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, pp. 8636–8660, 2024.

Maxime Zanella, Benoît Gérin, and Ismail B Ayed. Boosting vision-language models with transduction. *Advances in Neural Information Processing Systems*, 37:62223–62256, 2024.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11941–11952, 2023.

Chenyu Zhang, Kunlun Xu, Zichen Liu, Yuxin Peng, and Jiahuan Zhou. Scap: Transductive test-time adaptation via supportive clique-based attribute prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30032–30041, 2025a.

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, 2024a.

Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28718–28728, 2024b.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. *Advances in Neural Information Processing Systems*, 38:76280–76310, 2025b.

Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. In *International Conference on Learning Representations*, pp. 3597–3613, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29999–30009, 2025.

A ALGORITHM, CONSTANTS, AND DESIGN NOTES

Algorithm A.1 and Table A.1 give the rule and its constants. $\overline{H} = -\sum_{i\in\mathcal{I},c} p_{ic}\log\max(p_{ic},10^{-8})/(n\log C)$ is the mean normalized anchor entropy and $\nu_p = \exp H(\mathrm{hist}(\arg\max p))/C$ the class coverage after the anchor adjustment. Pools retaining at most ten images are returned unrefined.

Algorithm A.1 RoAR on an unlabeled pool

Require: zero-shot $\{p_i^0\}$, VLM and refiner embeddings $\{u_i, f_i\}$, class names
Ensure: predictions $\widehat{p}$ and unknown set $\mathcal{U}$
$\widehat{p} \leftarrow p^0$; compute S (Eq. 1) and $\mathcal{U}$ (Eq. 2); $\mathcal{I} \leftarrow \{1,\dots,N\}\setminus\mathcal{U}$
if $|\mathcal{I}| \le 10$ **then return** $(\widehat{p},\mathcal{U})$
end if
Anchor p on $\mathcal{I}$ (Eq. 4); pseudo-labels $\widehat{y}_i = \arg\max_c p_{ic}$
Raw and residual views; margins M_{raw}, M_{res} and ϕ
q^{raw}: $L(\overline{H})$ rounds of prototype refinement pooled with p (Eq. 9)
q^{res}: one prototype refinement pooled with p (Eq. 9)
if $E = e_{\mathrm{raw}} + e_{\mathrm{res}} > 0$ **then**
    $\widehat{p}_{\mathcal{I}} \leftarrow$ weighted geometric mean of $q^{\mathrm{raw}}, q^{\mathrm{res}}$ (Section 3.4)
else
    $\widehat{p}_{\mathcal{I}} \leftarrow$ Eq. (A.3) if $\nu_p \ge \kappa_{\mathrm{cov}}$, else p
end if
return $(\widehat{p},\mathcal{U})$

Table A.1: **Constants.** Values of the constants that Section 3 leaves as symbols, shared across every dataset, backbone, and prior setting.

stage	constant
negatives	bottom 30 % of nouns and of adjectives, 95th-percentile similarity
rejection	$K = \min(10, N)$ clusters, veto $\gamma = 0.7$
geometry	$k = 7$ removed directions
prototypes	temperature $\tau = 0.03$; raw view: confidence $\theta = 0.1$, $L(\overline{H}) = 1, 2, 3$ rounds at entropy cutoffs 0.5 and 0.7; residual view: weight cutoff $\theta' = 0.01$
whitened view	coverage $\kappa_{\mathrm{cov}} = 0.3$

The cluster cut. Rejection is decided per cluster rather than per image. Accepting a cluster costs $\sum_{i\in\mathcal{A}}(1-S_i)$ and rejecting it costs $\sum_{i\in\mathcal{A}} S_i$, so a cluster is rejected when its mean score is below $\frac{1}{2}$. A per-image cut, as in NegLabel, rejects several times more known images.

Prototype weights. The raw view uses hard weights and the residual view soft weights,

$$w_{ic}^{\mathrm{raw}} = \mathbf{1}\Big\{c = \arg\max_j q_{ij}^{(\ell-1)},\ \max_j q_{ij}^{(\ell-1)} \ge \theta\Big\}, \qquad w_{ic}^{\mathrm{res}} \propto p_{ic}\,\mathbf{1}\{p_{ic} \ge \theta'\}, \tag{A.1}$$

where the residual weights are normalized per image with 10^{-8} added to the sum.

The whitened expert. If both views show no structure ($E = 0$) and the anchor predicts enough classes ($\nu_p \ge \kappa_{\mathrm{cov}}$), the refiner features are whitened with spherical shrinkage,

$$\Sigma_{\mathrm{sh}} = \tfrac{1}{2}\Sigma + \tfrac{1}{2}\frac{\mathrm{tr}\,\Sigma}{d} I, \qquad f_i^{\mathrm{wh}} = \mathrm{unit}\Big(\Sigma_{\mathrm{sh}}^{-1/2}(f_i - \overline{f})\Big). \tag{A.2}$$

Soft prototypes with the residual weights give a prediction r_i^{wh}, which is pooled with the anchor twice,

$$q_i^{\mathrm{wh}} \propto p_i^{1/2}(r_i^{\mathrm{wh}})^{1/2}, \qquad \widehat{p}_i \propto p_i^{1/2}(q_i^{\mathrm{wh}})^{1/2} \propto p_i^{3/4}(r_i^{\mathrm{wh}})^{1/4}, \tag{A.3}$$

so the whitened view gets a total weight of $\frac{1}{4}$. Otherwise the output is p.

Numerical conventions. The negatives are split into $G = 100$ groups with shuffle seed 0, cosine k-means runs 30 updates from initialization seed 0, and the histogram is smoothed with $\delta_h = 0.01$. $\mathrm{unit}(v) = v/\max(\|v\|, 10^{-12})$. A prototype is used only if its weight is positive and its norm exceeds 10^{-8}. Classes with no weight keep zero prototypes and logits. The residual view is dropped if the within-class residual has rank at most k, and $\phi = 1$ if $\mathrm{tr}(S_B) \leq 10^{-12}$. Logs and probabilities are floored at 10^{-8}. The logit scale is 100 for CLIP and the native one for SigLIP. Negatives are mined with the prompts *a photo of a [noun].* and *a photo of a [adjective] object.*

B PROPERTY OF THE POOLING RULE

Proposition 1 (Geometric KL barycenter). *Let $p^{(j)}$ be strictly positive categorical distributions and let $w_j \geq 0$ with $\sum_j w_j = 1$. Then*

$$q^\star = \arg\min_{q \in \Delta_C} \sum_j w_j \mathrm{KL}(q \| p^{(j)}) = \mathrm{normalize}\left(\prod_j (p^{(j)})^{w_j} \right)$$

is the unique minimizer.

Proof. Put $a_c = \sum_j w_j \log p_c^{(j)}$ and $Z = \sum_c e^{a_c}$. The objective is $\sum_c q_c \log q_c - \sum_c q_c a_c = \mathrm{KL}(q \| q^\star) - \log Z$. Nonnegativity of KL proves uniqueness. □

Weights $(1 - \alpha_i, \alpha_i)$ give Eq. (9), ω_v the barycenter of Section 3.4, two weights of $\frac{1}{2}$ Eq. (A.3), and $(\frac{1}{2}, \frac{1}{2})$ on the uniform distribution and $\widetilde{h}$ the $\sqrt{\widetilde{h}}$ of Eq. (4). The result fixes the form of the combination, not its accuracy.

C EVALUATION PROTOCOL

Pools are sampled without replacement from the CLIP cross-domain test sets, with up to 3,333 known images on open-set pools and 4,000 on compound pools. Table C.1 gives the realized composition.

Table C.1: **Realized pool composition.** Original pool size N_d, and per composition the entropy-effective class fraction with the head-to-tail ratio in parentheses. With at least one image per class, the ratio at $\rho = 1000$ is capped by the largest class.

dataset	N_d	original	ρ 10	ρ 100	ρ 1000	25 % absent	50 % absent
ImageNet	50,000	1.00 (1)	0.82 (10)	0.56 (50)	0.45 (50)	0.75 (1)	0.50 (1)
Caltech101	2,465	0.66 (27)	0.67 (27)	0.58 (120)	0.41 (240)	0.49 (26)	0.33 (24)
Pets	3,669	1.00 (1)	0.82 (10)	0.55 (100)	0.40 (100)	0.76 (1)	0.49 (1)
Cars	8,041	0.99 (3)	0.85 (10)	0.58 (68)	0.44 (68)	0.75 (3)	0.50 (2)
Flowers102	2,463	0.88 (6)	0.86 (10)	0.57 (77)	0.43 (77)	0.66 (6)	0.44 (6)
Food101	30,300	1.00 (1)	0.82 (10)	0.55 (100)	0.39 (300)	0.75 (1)	0.50 (1)
Aircraft	3,333	1.00 (1)	0.82 (11)	0.57 (34)	0.47 (34)	0.75 (1)	0.50 (1)
SUN397	19,850	1.00 (1)	0.82 (10)	0.56 (50)	0.45 (50)	0.75 (1)	0.50 (1)
DTD	1,692	1.00 (1)	0.82 (9)	0.56 (36)	0.46 (36)	0.74 (1)	0.51 (1)
EuroSAT	8,100	0.99 (2)	0.79 (10)	0.52 (100)	0.36 (900)	0.79 (1)	0.50 (1)
UCF101	3,783	0.99 (2)	0.82 (10)	0.56 (49)	0.45 (49)	0.74 (2)	0.49 (2)

D WHICH MECHANISM FAILS WHERE

Failures follow the mechanism (Table D.1). Class-marginal constraints (TransCLIP, SOTA, OnZeta) collapse under long tails and absent classes. Stored or propagated pool statistics (ADAPT, COSMIC, ZLaP, CLIP prototypes) fall below zero-shot on every altered axis. Light re-estimation (GDA, StatA, LAME) is stable on the class axes but fails with 75 % unknown. Only RoAR stays above zero-shot everywhere. Our rejection step alone leaves the class-axis pools unchanged and costs 0.5 points on the known images, so the gain of RoAR on the known images comes from refinement, while most of its gain over all images comes from rejection.

Table D.1: **Which mechanism fails where.** Accuracy change over zero-shot (pp), mean over the eleven datasets, on the original pool and at the strongest level of each axis, computed before rounding. Shaded cells are below zero-shot.

					75 % unknown	
mechanism	method	original	$\rho = 1000$	50 % absent	known images	all images
class-marginal constraint	TransCLIP	+0.3	-4.8	-4.6	-8.9	-2.2
	SOTA (CLIP features)	+2.1	-5.6	-4.1	-4.3	-1.1
	SOTA	+6.8	-4.6	-3.2	-7.5	-1.9
	OnZeta	-0.4	-17.4	-13.8	–	–
pool statistics or caches	ADAPT	+2.5	-3.3	-1.9	-8.8	-2.2
	COSMIC	+2.2	-1.4	-0.3	-2.4	-0.6
	CLIP prototypes	-0.3	-1.9	-1.4	-22.0	-5.5
	ZLaP	-0.2	-9.6	-6.5	-3.6	-0.9
light re-estimation	GDA	+2.6	+0.8	+1.4	-4.2	-1.0
	StatA	+2.3	+0.3	+1.7	-10.0	-2.5
	LAME	+0.2	+2.4	+2.6	-0.8	-0.2
prototypes with rejection	**RoAR**	+7.7	+3.5	+4.6	+4.8	+53.2

Replacing only SOTA's uniform marginal by $\widetilde{h}$ recovers most of its loss under long tails and absent classes (Table D.2), so the marginal accounts for most of the collapse. The repair costs 0.8 points on the original pool and still trails RoAR on every pool.

Table D.2: **SOTA with its uniform marginal replaced by the smoothed histogram $\widetilde{h}$.** Mean accuracy (%) over the eleven datasets on the pools of Tables 1 and 2.

method	original	$\rho = 10$	$\rho = 100$	$\rho = 1000$	25 % absent	50 % absent
zero-shot	68.8	69.3	69.8	69.5	69.0	68.7
SOTA (uniform marginal, released)	75.6	73.8	68.8	64.9	71.5	65.5
SOTA (marginal $\widetilde{h}$)	74.8	73.5	70.8	69.1	72.8	70.0
RoAR	76.5	75.9	74.0	73.0	75.6	73.3

E Other backbones and ImageNet variants

With SigLIP and SigLIP2 and unchanged constants, SOTA leads by 0.3 to 1.0 points on the original and $\rho = 10$ pools, and RoAR leads on every other pool, where SOTA falls below zero-shot at the strongest level of each axis (Table E.1). On the ImageNet variants, RoAR leads on every dataset and pool (Table E.2).

Table E.1: **Other backbones.** Accuracy (%), mean over the eleven datasets, with SigLIP-B/16 and SigLIP2-B/16 in place of CLIP and the same DINOv2-L refiner and constants.

	SigLIP-B/16			SigLIP2-B/16		
pool	zero-shot	SOTA	RoAR	zero-shot	SOTA	RoAR
original	74.0	**79.8**	78.8	75.8	**81.4**	80.7
$\rho = 10$	74.0	**78.1**	77.8	76.1	**80.0**	79.7
$\rho = 100$	74.1	75.3	**76.4**	76.4	76.2	**78.6**
$\rho = 1000$	74.0	73.4	**75.8**	75.4	74.2	**77.2**
25 % absent	74.2	76.5	**77.9**	75.9	78.2	**79.8**
50 % absent	74.1	72.4	**76.2**	75.4	73.0	**77.5**
25 % unknown, known images	74.0	75.9	**77.1**	75.8	76.4	**78.0**
50 % unknown, known images	74.0	72.8	**76.6**	75.8	72.5	**77.6**
75 % unknown, known images	74.0	67.1	**76.5**	75.8	66.9	**77.1**
25 % unknown, all images	55.5	56.9	**74.4**	56.8	57.3	**75.6**
50 % unknown, all images	37.0	36.4	**74.2**	37.9	36.3	**75.1**
75 % unknown, all images	18.5	16.8	**74.8**	18.9	16.7	**76.1**

Table E.2: **ImageNet variants.** Accuracy (%) with CLIP ViT-B/16, the same DINOv2-L refiner, and the same constants.

pool	method	IN-A	IN-V2	IN-R	IN-Sketch	mean
original	zero-shot	57.8	64.3	80.8	50.9	63.4
	SOTA	70.6	72.1	84.8	64.5	73.0
	RoAR	**76.8**	**73.6**	**89.3**	**66.2**	**76.5**
$\rho = 10$	zero-shot	57.6	64.4	81.1	50.0	63.3
	SOTA	70.5	69.0	84.8	61.8	71.5
	RoAR	**76.5**	**70.6**	**89.4**	**63.5**	**75.0**
$\rho = 100$	zero-shot	58.5	64.1	83.0	49.5	63.8
	SOTA	70.1	66.7	82.7	56.2	68.9
	RoAR	**77.6**	**68.4**	**89.2**	**60.2**	**73.9**
$\rho = 1000$	zero-shot	60.1	64.7	84.0	50.4	64.8
	SOTA	69.8	66.5	80.1	53.7	67.5
	RoAR	**78.6**	**68.0**	**90.3**	**59.1**	**74.0**
25 % absent	zero-shot	57.5	64.5	80.8	50.8	63.4
	SOTA	67.2	69.5	81.5	62.5	70.2
	RoAR	**74.8**	**72.6**	**87.5**	**65.1**	**75.0**
50 % absent	zero-shot	57.3	64.1	80.6	50.5	63.1
	SOTA	62.0	64.8	77.5	57.8	65.5
	RoAR	**71.5**	**70.2**	**85.2**	**62.8**	**72.4**

F SENSITIVITY TO THE CONSTANTS

Most constants change accuracy only marginally (Table F.1). The exceptions are the anchor trigger, which matters only under strong skew, the whitened view, mainly on the original pool, and the rejection thresholds, which affect only the open-set column.

Table F.1: **Sensitivity to the constants.** Change in mean accuracy (pp) when one constant is moved and the rest are fixed. The 75 % unknown column scores all images with the unknown images as one extra class.

constant (paper value)	set to	original	$\rho = 1000$	50 % absent	75 % unknown
residual directions $k = 7$	3	-0.2	0.0	+0.1	0.0
	15	-0.2	-0.2	-0.3	0.0
anchor exponent $\eta = \frac{1}{2}$	0.25	0.0	-0.2	0.0	0.0
	1.0	0.0	+0.2	0.0	0.0
anchor trigger $\nu_0 < \frac{1}{2}$	$\frac{1}{3}$	0.0	-1.0	0.0	0.0
	$\frac{2}{3}$	0.0	+0.9	+0.9	0.0
exponent band $[\frac{1}{4}, \frac{3}{4}]$	$[0, 1]$	-0.2	-0.1	0.0	0.0
	fixed $\frac{1}{2}$	0.0	-0.1	-0.1	0.0
entropy cutoffs $0.5/0.7$	$0.4/0.6$	0.0	0.0	+0.1	-0.1
	$0.6/0.8$	-0.1	0.0	0.0	0.0
residual cutoff $\theta' = 0.01$	0.005	0.0	0.0	0.0	0.0
	0.02	0.0	0.0	0.0	0.0
temperature $\tau = 0.03$	0.02	+0.2	-0.1	0.0	0.0
	0.05	-0.4	+0.1	-0.1	0.0
confidence cutoff $\theta = 0.1$	0.05	0.0	0.0	0.0	0.0
	0.2	0.0	0.0	+0.1	0.0
whitened coverage $\kappa_{\text{cov}} = 0.3$	0.2	0.0	0.0	0.0	0.0
	0.5	0.0	0.0	0.0	0.0
whitened weight $\frac{1}{4}$	none	-0.8	0.0	-0.5	-0.2
	0.25	-0.3	0.0	-0.1	-0.1
	0.75	+0.1	0.0	0.0	0.0
shrinkage $\frac{1}{2}$	0.25	0.0	0.0	0.0	0.0
	0.75	-0.1	0.0	0.0	0.0
vocabulary fraction 30 %	0.15	0.0	0.0	0.0	-7.9
	0.45	0.0	0.0	0.0	+2.7
clusters $K = 10$	5	0.0	0.0	0.0	+0.3
	20	0.0	0.0	0.0	+0.1
cut $\frac{1}{2}$	0.4	0.0	0.0	0.0	-9.9
	0.6	0.0	0.0	0.0	+3.7
veto $\gamma = 0.7$	0.6	0.0	0.0	0.0	-1.8
	0.8	0.0	0.0	0.0	+1.5

G REJECTION, DETECTION, AND COMPOUND POOLS

On the Places pools, the rejection step recalls most unknown images while rejecting almost no known ones, whereas ZS-NTTA reaches the same recall by rejecting about a fifth to a quarter of them (Table G.1). Per-dataset accuracy over all images is in Table G.2. Giving every method our rejection step raises it by 15 to 54 points over all images, and RoAR still leads the best of them (Table G.3). The lead comes from the known images.

Table G.1: **Unknown and known images** on the pools of Table 3: unknown recall, known images rejected (rej.), and known accuracy (acc.). Methods without rejection have zero recall.

	25 % unknown			50 % unknown			75 % unknown		
method	recall	rej.	acc.	recall	rej.	acc.	recall	rej.	acc.
ZS-NTTA	59.7	25.8	57.8	64.3	24.4	57.8	68.9	18.9	59.8
RoAR	60.3	0.2	75.4	65.0	0.4	74.9	69.3	0.9	73.6

Table G.2: **Open-set.** Accuracy (%) over all images when Places images form 25, 50, or 75 % of the pool, with the unknown images as one extra class. A known image is correct only if it receives its label and is not rejected, and an unknown image only if it is rejected.

pool	method	IN	Cal	Pets	Cars	Flow	Food	Airc	SUN	DTD	Euro	UCF	mean
25 % unknown	zero-shot	52.8	71.2	68.2	51.0	55.0	64.4	22.8	48.2	41.0	39.1	54.0	51.6
	TransCLIP	52.3	68.7	65.7	48.4	58.4	59.7	19.9	50.3	40.4	26.9	52.7	49.4
	GDA	52.8	71.1	67.8	50.7	57.6	64.3	22.4	49.8	43.2	43.6	54.4	52.5
	StatA	52.4	70.2	65.6	47.2	56.7	63.9	20.4	50.3	43.0	40.5	52.9	51.2
	ZS-NTTA	54.7	**84.5**	89.6	71.5	66.3	84.9	12.9	48.4	44.2	16.9	66.8	58.3
	COSMIC	51.4	70.1	69.6	49.7	59.9	63.8	21.1	47.1	41.5	35.3	54.6	51.3
	SOTA	55.9	71.0	68.6	53.5	61.6	64.5	19.8	52.7	39.9	42.7	57.3	53.4
	RoAR	**71.7**	82.9	**93.7**	**79.7**	**82.3**	**89.8**	**52.4**	**53.3**	**57.8**	**48.6**	**75.9**	**71.7**
50 % unknown	zero-shot	35.2	47.5	45.4	34.0	36.7	42.9	15.2	32.1	27.3	26.0	36.0	34.4
	TransCLIP	34.3	44.7	42.0	32.2	37.9	37.0	13.3	33.0	27.0	15.6	33.9	31.9
	GDA	34.9	46.9	44.0	32.1	37.5	42.3	13.7	33.2	28.7	28.6	35.4	34.3
	StatA	34.7	46.4	42.0	28.6	35.5	41.9	12.8	33.3	26.5	20.4	34.4	32.4
	ZS-NTTA	48.1	**85.9**	92.7	81.0	73.6	89.6	19.7	**38.1**	56.5	9.0	**77.5**	61.1
	COSMIC	34.3	46.8	46.1	33.0	39.6	42.5	14.0	31.4	27.5	18.5	35.9	33.6
	SOTA	37.1	46.6	43.5	33.5	40.4	41.9	11.3	34.6	24.3	20.6	36.2	33.6
	RoAR	**73.1**	76.0	**93.5**	**84.8**	**80.6**	**92.1**	**67.2**	35.4	**59.3**	**32.6**	74.8	**69.9**
75 % unknown	zero-shot	17.6	23.7	22.7	17.0	18.3	21.5	7.6	16.1	13.7	13.0	18.0	17.2
	TransCLIP	16.6	21.6	18.7	15.8	17.7	15.6	6.7	16.0	12.6	6.6	16.7	15.0
	GDA	17.1	22.9	19.0	14.0	17.8	19.9	5.8	16.4	13.8	13.8	17.1	16.2
	StatA	17.2	22.8	17.1	12.3	17.0	19.6	5.4	16.4	12.6	5.6	15.8	14.7
	ZS-NTTA	59.6	**88.4**	**94.7**	89.8	72.9	92.8	53.9	**21.8**	**68.4**	6.1	**84.4**	66.6
	COSMIC	17.1	23.4	22.8	16.5	19.6	21.3	6.9	15.7	13.7	8.0	17.7	16.6
	SOTA	18.3	23.1	20.0	14.4	19.5	20.0	4.6	16.7	10.1	5.5	16.6	15.3
	RoAR	**71.3**	76.6	92.7	**90.2**	**84.4**	**94.8**	**82.5**	17.9	68.1	**19.5**	76.0	**70.4**

On compound pools, SOTA falls below zero-shot while RoAR gains on every combination (Table G.4). With iNaturalist unknowns, the rejection step recalls most unknown images at a similarly low known-rejection rate (Table G.5).

Table G.3: **The same rejection step for every method.** Accuracy (%) over all images on the pools of Table G.2, without and with our rejection step, mean over the eleven datasets. Known-image accuracy is in parentheses.

	25 % unknown		50 % unknown		75 % unknown	
method	w/o step	+ step	w/o step	+ step	w/o step	+ step
zero-shot	51.6 (68.8)	66.6 (68.7)	34.4 (68.8)	66.8 (68.5)	17.2 (68.8)	69.0 (68.3)
TransCLIP	49.4 (65.9)	65.2 (66.8)	31.9 (63.8)	65.4 (65.8)	15.0 (59.9)	68.0 (64.2)
GDA	52.5 (70.0)	68.0 (70.6)	34.3 (68.6)	67.6 (70.2)	16.2 (64.6)	69.3 (69.3)
StatA	51.2 (68.3)	67.1 (69.4)	32.4 (64.8)	66.2 (67.5)	14.7 (58.8)	68.3 (65.3)
SOTA	53.4 (71.2)	69.6 (72.7)	33.6 (67.3)	67.5 (70.0)	15.3 (61.4)	68.8 (67.1)
RoAR	56.3 (75.1)	**71.7** (75.4)	37.1 (74.2)	**69.9** (74.9)	17.9 (71.8)	**70.4** (73.6)

Table G.4: **Compound priors.** Accuracy (%) with 50 % absent classes, $\rho = 100$, and 50 % unknown images combined. Pools with unknown images are scored over all images as in Table G.2.

method	$\rho = 100$ and 50 % absent	$\rho = 100$ and 50 % unknown	$\rho = 100$, 50 % absent, and 50 % unknown
zero-shot	69.0	34.7	34.5
TransCLIP	60.3	32.4	31.4
GDA	69.5	35.0	35.2
StatA	69.4	34.2	35.0
SOTA	57.4	34.1	31.7
RoAR w/o rejection	73.2	36.9	36.3
RoAR	73.2	69.8	68.9

Table G.5: **iNaturalist as the unknown source.** Accuracy (%) over all images, as in Table G.2, with Places replaced by iNaturalist.

method	50 % unknown	75 % unknown
zero-shot	34.4	17.2
TransCLIP	32.7	15.6
GDA	34.5	16.4
StatA	32.2	15.0
SOTA	33.8	15.2
RoAR w/o rejection	36.6	17.6
RoAR	81.5	86.4
unknown recall (%)	87.7	90.2
known rejected (%)	0.68	1.18