[ICLR'27] ONLINE SEMANTIC GAUSSIAN BANKS FOR TEST-TIME OOD DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-trained vision-language models (VLMs) have shown strong performance for out-of-distribution (OOD) detection, while recent test-time methods further improve detection by adapting to unlabeled test streams. However, incorrect ID/OOD assignments during adaptation can introduce unreliable test samples and degrade detection performance. We propose **OSGB** (**O**nline **S**emantic **G**aussian **B**anks), an optimization-free framework for online test-time OOD detection. Using frozen VLM features and positive and negative textual semantics, **OSGB** selectively updates separate Gaussian banks for ID and OOD samples. As test-time statistics accumulate, a global Gaussian likelihood-ratio score is used for OOD detection. Each sample is scored before updating the banks, preserving causal online inference. Experiments on ImageNet-1K, shifted ImageNet variants, CIFAR-100, and OpenOOD show consistent improvements over the compared methods. `https://anonymous.4open.science/r/OSGB-CC0E/README.md`

1 INTRODUCTION

Out-of-distribution (OOD) detection is important for the reliable deployment of machine learning models in open-world environments, where test samples may come from unseen categories Cao et al. (2024). Vision-language models, such as CLIP Radford et al. (2021), have enabled strong zero-shot OOD detection by aligning visual and textual representations. CLIP-based OOD detection methods exploit image-text similarity and textual concepts to distinguish ID from OOD samples without requiring OOD training data Ming et al. (2022); Jiang et al. (2024). More recently, test-time methods have exploited unlabeled test streams to adapt OOD detection to the encountered test distribution Zhang & Zhang (2024).

However, online adaptation from unlabeled test streams containing both ID and OOD samples remains challenging Gong et al. (2023). Existing test-time methods adopt different strategies: AdaND Cao et al. (2025) learns an adaptive noise detector from pseudo-labels, OODD Yang et al. (2025) maintains a dynamic dictionary of OOD features for score calibration, and TTL Ye et al. (2026) learns OOD textual knowledge from test samples. Despite these advances, incorrect ID/OOD assignments during adaptation can introduce noisy samples into the learned test-time knowledge and reduce the separation between ID and OOD scores. As shown in Fig. 1, when evaluated on the same ImageNet$\rightarrow$SUN stream, these representative methods show greater ID/OOD score overlap than our approach. This motivates an online adaptation method that updates its ID and OOD representations using reliable test samples while suppressing unreliable updates.

To address this challenge, we propose **OSGB** (**O**nline **S**emantic **G**aussian **B**anks), an online semantic Gaussian adaptation framework that maintains separate ID and OOD Gaussian banks. Each incoming sample is scored immediately using only the current input and statistics accumulated from past test samples, while frozen CLIP features and positive/negative text semantics guide reliable online bank construction. As the banks accumulate evidence, the detector progressively exploits Gaussian distributional information to improve ID/OOD separation. The VLM remains frozen throughout adaptation, and each sample is scored before it can update the banks.

Extensive experiments on ImageNet-scale and CIFAR-scale OOD benchmarks demonstrate the effectiveness of the proposed framework. On ImageNet-1K with four OOD datasets, our method achieves an average FPR95 of 6.11% and AUROC of 98.34%, outperforming prior test-time OOD methods. Consistent gains are also observed across shifted ImageNet variants and CIFAR-100, with

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.

ablation studies confirming the contribution of the proposed online adaptation components. Our main contributions are summarized as follows:

- We propose an optimization-free online test-time OOD detection framework that models ID and OOD feature distributions with separate semantic Gaussian banks while keeping the pretrained VLM frozen.
- We introduce reliable online bank construction based on semantic guidance and Gaussian statistics, together with selective update mechanisms that reduce errors from unreliable ID/OOD assignments and a global Gaussian likelihood-ratio score for detection.
- We conduct extensive experiments on ImageNet-1K, shifted ImageNet variants, and CIFAR-100, demonstrating consistent improvements over existing OOD detection methods, with controlled stream comparisons and ablation studies validating the proposed framework.

2 Related Works

OOD Detection with Vision-Language Models. Traditional OOD detection methods use different scores to separate ID and OOD samples. Early methods use maximum softmax confidence Hendrycks & Gimpel (2016), temperature scaling and input perturbation Liang et al. (2017), Mahalanobis distance in feature space Lee et al. (2018), and energy-based scores Liu et al. (2020). Other methods use auxiliary outlier data during training, such as Outlier Exposure Hendrycks et al. (2018). With the development of vision-language models, recent methods use textual information for OOD detection. MCM Ming et al. (2022) uses image-text similarity with ID class names, while NegLabel Jiang et al. (2024) introduces negative labels to provide OOD-related information. CLIPN Wang et al. (2023) introduces a learnable "no" prompt and text encoder to model negation semantics, while NegPrompt Li et al. (2024) learns transferable negative prompts for OOD detection. LoCoOp Miyai et al. (2023) studies few-shot OOD detection through prompt learning. Other methods use external semantic knowledge to improve VLM-based OOD detection Cao et al. (2024). However, these methods mainly use detection knowledge that is fixed before the test stream is observed.

Test-Time Adaptation for OOD Detection. Test-time adaptation uses unlabeled test samples to adapt the model during testing. Early methods include test-time training with self-supervision Sun et al. (2020) and entropy minimization Wang et al. (2020). For vision-language models, TPT Shu et al. (2022) updates prompts at test time, while TDA Karmanov et al. (2024) uses a training-free dynamic cache. In open-world test streams, unreliable samples can reduce adaptation performance, which has motivated methods that select test samples more carefully Gong et al. (2023). AdaND Cao et al. (2025) trains an additional noise detector using pseudo-labels while keeping the VLM frozen. OODD Yang et al. (2025) maintains a dynamic dictionary of OOD visual features for score calibration, while AdaNeg Zhang & Zhang (2024) stores selected test features to build adaptive negative proxies. TTL Ye et al. (2026) learns OOD textual information from pseudo-labeled test samples and maintains a textual knowledge bank for score calibration. TT-RAA Fan et al. (2025) uses a streaming Gaussian database to model test-time feature statistics for training-free adaptation.

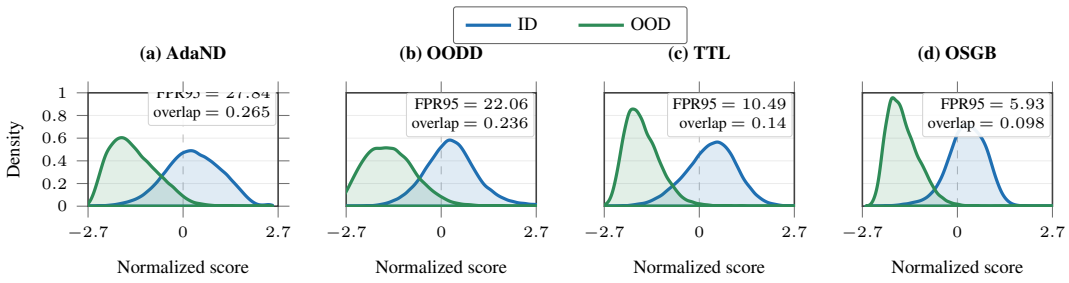


Figure 1: **ID/OOD score separation on ImageNet→SUN.** AdaND, OODD, TTL, and **OSGB** are evaluated on the same 50K ImageNet ID and 10K SUN OOD samples under the same three stream permutations. Scores are standardized independently within each method for visualization.

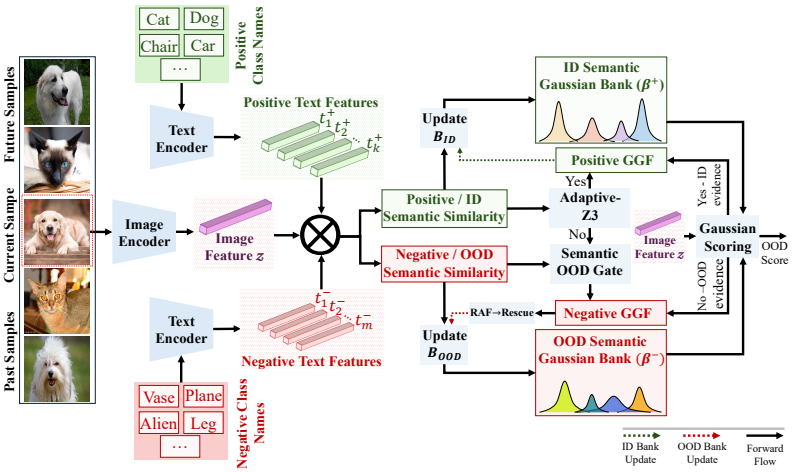


Figure 2: Overview of the proposed online semantic Gaussian adaptation framework. Positive and negative semantic evidence selects ID/OOD candidates for updating the corresponding Gaussian banks, while Gaussian scoring controls bank updates and produces the final OOD score.

3 METHODOLOGY

3.1 PRELIMINARIES

OOD Detection Setup. Let $\mathcal{Y}_{\text{ID}} = \{c_1, \ldots, c_K\}$ denote the known in-distribution (ID) label space. At test time, samples arrive sequentially as a stream $\{x_t\}_{t=1}^{T}$, where each x_t may belong to an ID class $c_k \in \mathcal{Y}_{\text{ID}}$ or to an unknown class outside the ID label space. The goal of OOD detection is to distinguish ID from OOD samples without access to target-domain annotations during testing.

Frozen CLIP and Semantic Prototypes. We use a pretrained CLIP model Radford et al. (2021) with frozen image and text encoders f_{img} and f_{txt}. For each test image x_t, the image encoder produces an ℓ_2-normalized visual feature $z_t \in \mathbb{R}^D$. Using a fixed prompt template, the ID class names are encoded into normalized positive text prototypes $T^+ = [t_1^+, \ldots, t_K^+]^\top \in \mathbb{R}^{K \times D}$. Following negative-label-based OOD detection Jiang et al. (2024), we construct a fixed set of $M = 10{,}000$ negative semantic concepts from a WordNet-derived Miller (1995) vocabulary of adjective and noun concepts. These concepts are encoded by the frozen CLIP text encoder as $T^- = [t_1^-, \ldots, t_M^-]^\top \in \mathbb{R}^{M \times D}$. Both prototype sets are constructed before processing the test stream and remain fixed throughout adaptation. Since all embeddings are ℓ_2-normalized, we obtain the positive and negative cosine-similarity vectors as $s_t^+ = T^+ z_t \in \mathbb{R}^K$ and $s_t^- = T^- z_t \in \mathbb{R}^M$, respectively.

3.2 OVERVIEW

OSGB performs training-free online OOD adaptation while keeping the CLIP encoders and semantic prototypes frozen. Positive and negative semantic similarities are used to identify reliable ID and OOD candidates, which selectively update separate semantic Gaussian banks. Gaussian evidence and contamination filtering prevent ambiguous samples from corrupting the banks. Once the banks are established, detection transitions from the semantic score to an aligned global Gaussian likelihood gap. Each sample is scored using only past information and updates the banks only afterward.

3.3 SEMANTIC CANDIDATE SELECTION

For each incoming sample, we use its similarities to the positive and negative text prototypes to select candidates for updating the two semantic banks. These candidate decisions are used only for bank adaptation and are separate from the final OOD prediction.

Positive Candidate Selection. From the positive similarity vector s_t^+, we compute three statistics. First, $p_t^{\max} = \max_j s_{t,j}^+$ measures the strongest similarity to an ID class. Second, $u_t = (p_t^{\max} -$

Table 1: OOD detection performance on ImageNet-1K and representative distribution shifts using CLIP ViT-B/16. Each ID dataset is evaluated against iNaturalist, SUN, Places, and DTD. Results are reported as AUROC (%) and FPR95 (%). Best and second-best results are shown in bold and underlined, respectively. Additional results on ImageNet-Sketch and ImageNet-V2 are reported in the appendix.

ID Dataset	Method	iNaturalist Van Horn et al. (2018)		SUN Xiao et al. (2010)		Places Zhou et al. (2017)		DTD Cimpoi et al. (2014)		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
ImageNet Deng et al. (2009)	MCM	94.611	30.910	92.570	37.590	89.766	44.690	86.107	57.766	90.763	42.739
	OODD	99.332	2.240	94.959	22.060	86.963	44.733	93.081	30.975	93.584	25.002
	NegLabel	99.472	2.040	95.475	20.930	91.543	36.600	89.917	45.337	94.102	26.227
	AdaND	98.359	7.240	94.129	27.840	92.347	30.507	91.305	30.934	94.035	24.130
	AdaNeg	99.723	0.510	97.367	10.570	93.933	35.970	95.497	26.423	96.630	18.368
	TTL	99.589	1.180	97.803	10.487	95.548	18.473	92.973	31.944	96.478	15.521
	OSGB	**99.793**	**0.423**	**98.407**	**5.933**	**96.762**	**12.090**	**98.389**	**6.005**	**98.338**	**6.113**
ImageNet-A Hendrycks et al. (2021b)	MCM	79.995	76.790	75.738	80.340	70.696	80.830	62.847	85.035	72.319	80.749
	OODD	98.444	5.780	91.902	29.567	80.070	53.623	91.981	30.165	90.599	29.784
	NegLabel	98.801	4.140	90.147	43.470	83.129	59.530	79.884	65.266	87.990	43.101
	AdaND	94.556	22.523	90.974	38.490	87.410	47.630	91.472	31.596	91.103	35.060
	AdaNeg	98.860	**3.893**	94.153	28.037	89.647	48.113	93.620	28.033	94.070	27.019
	TTL	97.208	9.990	94.098	27.117	88.724	39.300	82.741	55.124	90.693	32.883
	OSGB	**99.130**	4.240	**95.960**	**16.960**	**91.560**	**33.560**	**97.190**	**11.420**	**95.960**	**16.540**
ImageNet-R Hendrycks et al. (2021a)	MCM	83.708	71.750	79.975	76.130	75.395	77.280	68.526	82.677	76.901	76.959
	OODD	99.267	2.870	93.487	25.747	83.447	47.570	92.274	30.201	92.119	26.597
	NegLabel	99.579	1.630	96.166	15.510	92.036	29.330	90.325	37.145	94.526	20.904
	AdaND	99.673	0.810	98.485	6.900	96.976	13.637	93.690	26.483	97.206	11.958
	AdaNeg	99.903	**0.037**	99.320	2.643	97.023	18.493	95.953	24.140	98.050	11.328
	TTL	99.655	0.797	98.747	5.153	96.163	15.100	85.583	60.975	95.037	20.506
	OSGB	**99.970**	0.070	**99.650**	**1.410**	**99.070**	**3.440**	**98.850**	**4.810**	**99.390**	**2.430**

$\mu_t^+)/\sigma_t^+$ measures how much this maximum differs from the remaining positive similarities, where μ_t^+ and σ_t^+ are the mean and standard deviation of s_t^+. Third, we compute the positive log-sum-exp score $\ell_t^+ = \tau \log \sum_{j=1}^{K} \exp(s_{t,j}^+/\tau)$, where τ is the pretrained CLIP temperature. After an initial calibration prefix of B_{cal} samples, these three statistics are standardized using only previously observed samples. Let $\tilde{\ell}_t^+$, $\tilde{u}_t$, and $\tilde{p}_t^{\max}$ denote their standardized values. We combine them as

$$e_t = \frac{\tilde{\ell}_t^+ + \tilde{u}_t - \tilde{p}_t^{\max}}{3}. \tag{1}$$

Using multiple statistics reduces the dependence on a single CLIP similarity measure. The negative sign on $\tilde{p}_t^{\max}$ avoids giving additional weight to the absolute maximum similarity, which is already reflected in ℓ_t^+, while u_t captures how distinctive the strongest positive response is relative to the remaining classes. We then estimate an adaptive threshold θ_t from the previous evidence values by selecting the split that minimizes the total within-group variance. The standardization statistics and threshold are updated at fixed intervals after the initial calibration stage using all previously observed raw statistics. At each update, the previous evidence values are recomputed using the updated standardization, while the current sample is excluded. A sample is selected as a positive candidate when $e_t > \theta_t$.

Negative Candidate Selection. We use the negative text prototypes to identify candidate OOD samples. The M negative prototypes Jiang et al. (2024) are divided into G groups. We compute the positive evidence $L_t^+ = \log \sum_{j=1}^{K} \exp(s_{t,j}^+/\tau)$ and, for each negative group $\mathcal{G}_g$, $L_{t,g}^- = \log \sum_{m \in \mathcal{G}_g} \exp(s_{t,m}^-/\tau)$. The semantic score is

$$q_t = \frac{1}{G} \sum_{g=1}^{G} \sigma\big(L_t^+ - L_{t,g}^-\big), \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function. The group-wise scores are averaged to reduce the effect of individual negative concepts. A larger q_t indicates stronger ID evidence, while a sample with $q_t < \tau_{\text{OOD}}$ is selected as a negative candidate. The selected candidates are then passed to the bank-management procedure described next.

3.4 Online Semantic Gaussian Banks

We maintain a positive bank $\mathcal{B}^+$ and a negative bank $\mathcal{B}^-$ to model the test-time feature distributions. The positive bank contains components associated with the K ID prototypes, while the negative

Table 2: OOD detection on CIFAR-100 Krizhevsky et al. (2009) using CLIP ViT-B/16. Results are reported as AUROC (%) and FPR95 (%); higher AUROC and lower FPR95 are better. Best and second-best results are shown in bold and underlined, respectively.

Method	SVHN Netzer et al. (2011)		LSUN-R Yu et al. (2015)		LSUN-C Yu et al. (2015)		iSUN Xu et al. (2015)		Texture Cimpoi et al. (2014)		Places365 Zhou et al. (2017)		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
MCM	88.911	70.400	85.604	74.200	91.838	48.490	85.665	72.392	72.943	92.039	61.001	97.820	80.994	75.890
OODD	90.520	60.600	84.610	75.200	92.080	44.720	83.850	75.980	73.120	90.850	59.090	98.090	80.550	74.240
AdaND	98.118	3.213	96.247	16.833	98.130	6.350	95.409	21.296	92.473	24.970	68.421	79.893	91.633	25.426
AdaNeg	91.994	27.050	90.499	30.480	91.990	27.050	89.950	31.320	89.081	59.740	77.471	81.800	88.498	42.907
TTL	**99.980**	**0.010**	99.470	**2.520**	99.540	**2.010**	**99.670**	**1.270**	98.910	1.680	97.980	6.650	99.260	2.360
OSGB	99.934	0.140	**99.488**	3.243	**99.570**	3.120	99.425	3.686	**99.928**	**0.112**	**99.876**	**0.320**	**99.703**	**1.770**

bank contains components associated with the M negative prototypes. Both banks are initialized empty. During the initial calibration prefix, no bank updates are performed. After the calibration stage, semantic candidates are used to populate the two banks. Before Gaussian filtering becomes available, these candidates are admitted using only the semantic criteria. If both positive and negative semantic criteria are satisfied, the positive candidate takes precedence.

Semantic Bank Construction. Rather than assigning an accepted sample to a single prototype, we distribute its feature across the k most similar semantic prototypes. For a positive-bank update, let $\mathcal{I}_t^+$ denote the indices of the top-k entries of s_t^+. The corresponding assignment weights are

$$a_{t,j}^+ = \frac{\exp(s_{t,j}^+/\tau)}{\sum_{r\in\mathcal{I}_t^+}\exp(s_{t,r}^+/\tau)}, \qquad j \in \mathcal{I}_t^+, \tag{3}$$

and the same operation is applied to s_t^- for a negative-bank update. Each component maintains the accumulated assignment weight, weighted feature sum, and weighted squared-feature sum. These quantities are updated online using $a_{t,j}^+$, from which the component mean and diagonal variance are computed. Components with zero accumulated assignment weight are not used for Gaussian scoring. To stabilize components with limited observations, the local variance $\mathbf{v}_j^{\mathrm{loc}}$ is combined with the global variance $\mathbf{v}^{\mathrm{glob}}$ of the corresponding bank as $\mathbf{v}_j = \frac{W_j}{W_j+1}\mathbf{v}_j^{\mathrm{loc}} + \frac{1}{W_j+1}\mathbf{v}^{\mathrm{glob}}$, where W_j is the accumulated assignment weight of component j. The resulting diagonal variance is lower-bounded by 10^{-6} for numerical stability. Thus, components with fewer observations rely more on the global bank statistics.

Gaussian-Guided Updates. Gaussian filtering is activated only after sufficient online statistics have been accumulated, since Gaussian estimates can be unreliable in the early low-sample stage Fuchs et al. (2025). In our implementation, this occurs after Adaptive-Z3 has been initialized, both banks contain at least 32 accepted samples, and at least 20 valid Gaussian observations have been accumulated. These readiness values are selected once on the development setting and kept fixed across all experiments.

Before these conditions are satisfied, the method remains in the bootstrap stage and uses the semantic criteria for bank updates. Once Gaussian filtering is activated, semantic candidates are further checked before updating the banks. Let $p^+(z_t)$ and $p^-(z_t)$ denote the component-mixture likelihoods under the positive and negative banks. Each likelihood is computed as the uniform average over the active Gaussian components of the corresponding bank; components with zero accumulated

Table 3: CIFAR-100 results under the official OpenOOD near- and far-OOD protocol Yang et al. (2022); Zhang et al. (2023). Near-OOD consists of CIFAR-10 and Tiny ImageNet (TIN), while far-OOD consists of MNIST, SVHN, Texture, and Places365. Results are reported as FPR95↓ / AUROC↑ (%). Best and second-best results are shown in bold and underlined, respectively.

Method	Near-OOD		Far-OOD				Near Avg.	Far Avg.
	CIFAR-10	TIN	MNIST	SVHN	Texture	Places365		
MCM	73.48 / **73.59**	78.33 / 66.58	28.02 / 88.60	37.18 / 85.11	83.37 / 70.63	89.99 / 58.84	75.91 / 70.08	59.64 / 75.80
NegLabel	**71.72** / 72.51	76.81 / 78.23	16.56 / 92.82	16.78 / 95.18	61.11 / 88.59	77.48 / 78.06	74.27 / 75.37	42.98 / 88.66
AdaND	85.29 / 61.55	56.41 / 84.57	0.52 / 99.89	1.39 / 99.60	24.72 / 92.14	61.91 / 74.47	70.85 / 73.06	22.13 / 91.52
AdaNeg	72.24 / 72.45	75.74 / 79.16	16.49 / 93.11	16.61 / 95.24	56.48 / 90.18	75.59 / 80.07	73.99 / 75.80	41.29 / 89.65
TTL	84.94 / 69.49	38.51 / 90.96	0.31 / 99.86	0.22 / 99.80	15.01 / 96.85	71.44 / 70.98	61.73 / 80.23	21.75 / 91.87
OSGB	75.11 / 70.31	**2.46 / 99.66**	**0.00 / 99.98**	**0.04 / 99.96**	**0.15 / 99.92**	**0.02 / 99.97**	**38.79 / 84.99**	**0.05 / 99.96**

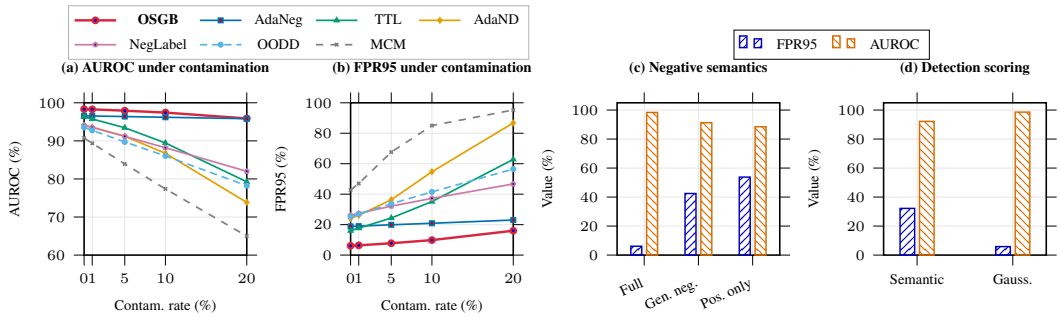


Figure 3: Additional analysis of OSGB on ImageNet-1K. **(a)–(b)** Robustness to adversarial stream contamination, averaged over iNaturalist, SUN, Places, and DTD. **(c)** Ablation of negative semantics, comparing the full semantic bank with cardinality-matched generic negatives and a positive-only variant. **(d)** Ablation of the detection scoring strategy, comparing semantic-only scoring with the global Gaussian likelihood-ratio score under the same online adaptation procedure.

assignment weight are excluded. We compute the normalized log-likelihood gap

$$g_t = \frac{1}{D}\left[\log p^+(z_t) - \log p^-(z_t)\right]. \tag{4}$$

A positive candidate is accepted when $g_t > 0$, indicating stronger support from the positive bank. For a negative candidate, we first require $g_t < 0$ and then apply an additional rival-aware check.

Some negative concepts can be semantically close to an ID class. For the strongest negative prototype, we therefore find its closest positive prototype and standardize the corresponding positive similarity using the mean and standard deviation of the current sample's positive similarity vector, yielding r_t. The Gaussian gap is adjusted as

$$g_t^{\mathrm{RAF}} = g_t + \beta \max(r_t, 0), \tag{5}$$

where β controls the contribution of the rival evidence. A negative candidate is accepted when $g_t^{\mathrm{RAF}} < 0$. The rescue rule is considered only when a negative candidate satisfies $g_t < 0$ but is rejected by the rival-aware check. In this case, the sample is assigned to the negative bank when $p_t^{\max} < \mu_p - z_{\mathrm{rescue}}\sigma_p$, where μ_p and σ_p are computed from previously accepted positive samples. Samples that satisfy none of these conditions are not used to update either bank.

3.5 Global Gaussian OOD Detection

The component Gaussians described above are used to control bank updates. For the final OOD detection score, we instead maintain one global diagonal Gaussian for each bank using all accepted positive and negative features. Let $\boldsymbol{\mu}^+$ and $\mathbf{v}^+$ denote the mean and diagonal variance of the positive bank, and let $\boldsymbol{\mu}^-$ and $\mathbf{v}^-$ denote the corresponding statistics of the negative bank.

For an incoming feature z_t, we compute its log-likelihood under the positive and negative global Gaussians, denoted by $\log p^+_{\mathrm{glob}}(z_t)$ and $\log p^-_{\mathrm{glob}}(z_t)$. The Gaussian detection score is defined as

$$S_t = \frac{1}{D}\left[\log p^+_{\mathrm{glob}}(z_t) - \log p^-_{\mathrm{glob}}(z_t)\right]. \tag{6}$$

A larger S_t indicates that the sample is better explained by the positive bank and is therefore more ID-like, while a smaller score indicates stronger agreement with the negative bank and is more OOD-like.

During the initial stage, before sufficient Gaussian statistics are available, we use the semantic score q_t from Sec. 3.3. Because q_t and S_t have different numerical scales, we align the two score regimes before combining them for evaluation. We first transform the semantic score as $H_t = \log \frac{\mathrm{clip}(q_t,\epsilon,1-\epsilon)}{1-\mathrm{clip}(q_t,\epsilon,1-\epsilon)}$. Once Adaptive-Z3 is active and both banks contain sufficient statistics for Gaussian scoring, we estimate an affine mapping $\hat{S}_t = aS_t + b$ from paired semantic and Gaussian scores observed only before the transition, where $a = \mathrm{IQR}(H)/(\mathrm{IQR}(S) + \epsilon)$ and

$b = \text{median}(H) - a\ \text{median}(S)$. The mapping is fixed at the transition and used for all subsequent samples. The final online score is therefore H_t before the transition and $\hat{S}_t$ afterward. All samples, including those from the initial stage, are included in the reported metrics. Importantly, each sample is scored using only information available before its arrival-time update, so the current sample cannot affect the model used to score itself.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Evaluation Protocol. We evaluate **OSGB** on ImageNet- and CIFAR-based benchmarks. For ImageNet-1K, we use iNaturalist, SUN, Places, and DTD as OOD datasets. We further evaluate under ID distribution shifts using ImageNet-Sketch (ImageNet-K), ImageNet-A, ImageNet-V2, and ImageNet-R as ID datasets with the same four OOD datasets. For CIFAR-100, we use SVHN, LSUN-R, LSUN-C, iSUN, Texture, and Places365 as OOD datasets. We additionally evaluate CIFAR-100 under the official OpenOOD near- and far-OOD protocol Yang et al. (2022); Zhang et al. (2023), using the released test lists and full evaluation sets without OOD subsampling. We report FPR95 and AUROC, where lower FPR95 and higher AUROC indicate better OOD detection. Unless stated otherwise, experiments use three stream-order seeds and report the average over the three runs. The exact stream construction and additional protocol details are provided in the supplementary material.

Implementation Details. We use CLIP ViT-B/16 Radford et al. (2021) as the backbone in all experiments, with both image and text encoders kept frozen during testing. **OSGB** operates online without gradient-based optimization, and each test sample is scored before it is used to update the semantic Gaussian banks. All samples, including the initial calibration and bootstrap samples, are included in the reported metrics using the causally aligned arrival-time scores defined in Sec. 3.5. Hyperparameters are selected once using a separate ImageNet→SUN development setting and are then fixed for all ImageNet variants and CIFAR-100 experiments.

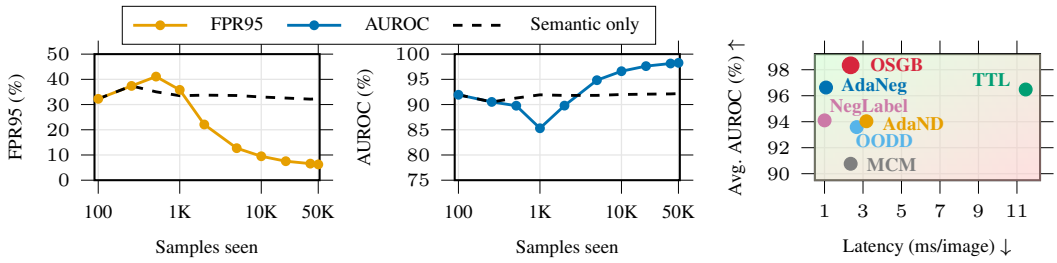
Comparison Methods. We compare **OSGB** with MCM Ming et al. (2022), NegLabel Jiang et al. (2024), AdaND Cao et al. (2025), OODD Yang et al. (2025), AdaNeg Zhang & Zhang (2024), and TTL Ye et al. (2026). We run all compared methods under the same datasets, test splits, backbone, and stream orders, using their official implementations and configurations when available. Since the official implementation of TTL was unavailable at the time of our experiments, we reimplemented it following the paper and supplementary material. Implementation details and a comparison with the originally reported results are provided in the supplementary material.

4.2 MAIN RESULTS

ImageNet-Based Benchmarks. As shown in Table 1, **OSGB** achieves the best average OOD detection performance on the standard ImageNet-1K setting, with an AUROC of 98.34% and an FPR95 of 6.11%. The improvement remains consistent under the evaluated shifted-ID benchmarks, where the method obtains the best average performance across all four shifted-ID settings. These results demonstrate the effectiveness of the proposed online adaptation across the evaluated ID distributions.

CIFAR-100 Benchmark. Table 2 presents the results across six OOD datasets. **OSGB** achieves the best average performance, with an AUROC of 99.70% and an FPR95 of 1.77%. Although TTL performs better on several individual datasets, **OSGB** achieves the strongest overall average and notably low FPR95 on Texture and Places365. These results further show the effectiveness of the proposed framework across the evaluated CIFAR-based OOD benchmarks.

OpenOOD Near- and Far-OOD Benchmark. Table 3 shows that **OSGB** achieves the best average performance for both near- and far-OOD detection. The gain is especially strong on far-OOD, where **OSGB** obtains 0.05% FPR95 and 99.96% AUROC. For near-OOD, performance varies across the two datasets: **OSGB** performs strongly on TIN, while on CIFAR-10 MCM achieves higher AUROC and NegLabel achieves lower FPR95.



(a) FPR95 over time **(b)** AUROC over time **(c)** Performance–efficiency

Figure 4: Efficiency and online behavior of OSGB on ImageNet. **(a)–(b)** Cumulative FPR95 and AUROC over the test stream, averaged over four OOD datasets and three stream-order seeds; dashed curves show semantic-only scoring. **(c)** Average AUROC versus latency on the controlled ImageNet→SUN stream.

4.3 Ablation Studies

Component Ablation. Table 4 evaluates the contribution of the main components of **OSGB**. Performance improves as SGB, GGF, and RAF are added, with GGF providing the largest improvement among the bank-update components. Using the global Gaussian likelihood ratio for final detection further improves performance, reaching 6.11% FPR95 and 98.34% AUROC. These results show that the filtering components improve online adaptation, while the global Gaussian score provides stronger final ID–OOD separation.

Importance of Negative Semantics. Figure 3(c) studies the effect of the negative semantic concepts used by **OSGB**. To keep the number of negative entries the same, the generic-negative variant uses 10,000 entries organized into the same five groups of 2,000 as the full semantic bank. Compared with generic negatives, the full semantic bank reduces FPR95 by 36.37 points and improves AUROC by 7.04 points. Removing the negative bank leads to a further drop in performance. These results show that the gain does not come only from the number or grouping of negative entries, and that meaningful negative semantic concepts are important for OOD detection.

Effect of Adaptive Gaussian Scoring. Figure 3(d) studies the detection scoring strategy while keeping the online adaptation procedure unchanged. Switching from semantic-only scoring to the global Gaussian likelihood-ratio score reduces FPR95 by 26.09 points and improves AUROC by 6.14 points. This shows that the Gaussian statistics collected during adaptation provide a stronger detection score after enough samples have been observed.

4.4 Online and Efficiency Analysis

Performance over Time. Figure 4(a)–(b) reports cumulative FPR95 and AUROC computed from the arrival-time scores of all samples observed up to each stream position, averaged over four OOD datasets and three stream-order seeds. During the initial bootstrap period, the Gaussian banks contain insufficient statistics and performance is governed primarily by the semantic score. As the banks accumulate reliable statistics, average FPR95 decreases from 22.09% at 2K samples to 12.70% at 5K, 9.49% at 10K, and 6.57% at 40K. AUROC increases from 89.79% to 94.83%, 96.61%, and

Table 4: Component ablation on the ImageNet-1K OOD benchmark. Results are averaged over iNaturalist, SUN, Places, and DTD using three stream seeds. All adaptive variants use the same Adaptive-Z3 positive bootstrap.

SGB	GGF	RAF	Global	FPR95 ↓	AUROC ↑
✗	✗	✗	✗	32.20	92.20
✓	✗	✗	✗	26.03	93.05
✓	✓	✗	✗	12.84	96.15
✓	✓	✓	✗	12.43	96.29
✓	✓	✓	✓	**6.11**	**98.34**

98.14%, respectively. The final benchmark result is 6.11% FPR95 and 98.34% AUROC, compared with 32.20% FPR95 and 92.20% AUROC for the semantic-only score. Thus, the method improves as reliable online statistics accumulate, while the initial stage before sufficient bank statistics are available remains more challenging.

Robustness to Adversarial Stream Contamination. Figure 3(a)–(b) evaluates robustness as the percentage of adversarial samples in the online test stream increases from 0% to 20%. Adversarial examples are generated offline using frozen CLIP ViT-B/16 Radford et al. (2021) with localized ℓ_∞ PGD Madry et al. (2017); Goodfellow et al. (2014), using $\epsilon = 4/255$, a step size of $1/255$, and 10 steps. The perturbation is limited to a fixed square region covering 25% of the image. For ID samples, the attack reduces the MCM confidence score Ming et al. (2022), while for OOD samples it increases the score. The same attacked samples and stream orders are used for all methods, and the attack does not use the online state of any method. As the contamination level increases, **OSGB** shows a smaller drop in performance than the compared methods. At 20% contamination, **OSGB** achieves 15.99% FPR95 compared with 23.02% for AdaNeg, while maintaining a similar AUROC. We also evaluate a fixed clean subset containing only samples that are never attacked. On this subset, FPR95 increases from 6.21% to 8.54% as contamination increases from 0% to 20%, showing that the effect on clean samples remains small.

Performance–Efficiency Trade-off. Figure 4(c) compares average ImageNet OOD detection performance with inference latency measured on the same hardware and controlled ImageNet→SUN stream. Under this measurement, **OSGB** achieves an average AUROC of 98.34% with a latency of 2.36 ms per image. Its latency is comparable to MCM and lower than OODD, AdaND, and TTL, while providing higher OOD detection performance. The exact timing convention, post-encoder overhead, and memory measurements are provided in the supplementary material.

Hyperparameter Sensitivity. As shown in Figure 5, we analyze the sensitivity of **OSGB** to its main hyperparameters on ImageNet→SUN. Performance remains relatively stable over the tested ranges for most parameters. Larger changes are observed for the calibration-prefix size and the number of semantic groups, while the remaining parameters show only moderate variation around their default values. We use the same default configuration throughout the main experiments.

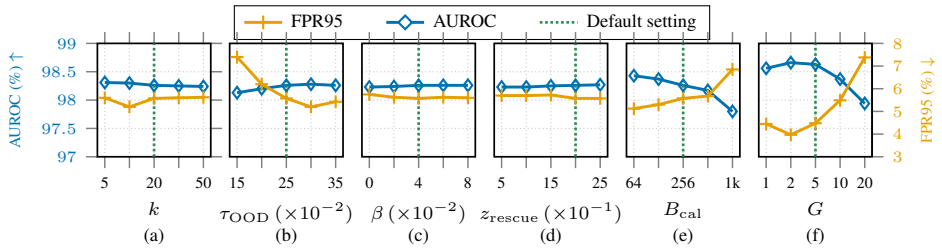


Figure 5: Hyperparameter sensitivity of **OSGB** on ImageNet→SUN.

4.5 Limitations

OSGB is designed for mixed ID/OOD streams in which the observed distributions remain sufficiently stable for reliable statistics to accumulate. Performance is weaker during the initial online stage before sufficient bank statistics have accumulated and under abrupt and persistent changes in the OOD distribution. In the non-stationary experiment, AUROC remains 94.15% but FPR95 increases to 23.86%, showing that changing OOD sources remain challenging. The method also depends on structured negative semantic concepts and performs best when the stream is not strongly OOD-dominated. These settings are important directions for future work.

5 Conclusion

In this work, we proposed **OSGB**, an optimization-free framework for online test-time OOD detection. **OSGB** uses frozen VLM features and positive and negative textual semantics to construct and update Gaussian banks for ID and OOD samples. As test-time statistics accumulate, a global Gaussian likelihood-ratio score is used for detection. Experiments across standard OOD benchmarks show consistent improvements over the compared methods.

REFERENCES

Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. Envisioning outlier exposure by large language models for out-of-distribution detection. *arXiv preprint arXiv:2406.00806*, 2024.

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 conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Xinqi Fan, Xueli Chen, Luoxiao Yang, Chuin Hong Yap, Rizwan Qureshi, Qi Dou, Moi Hoon Yap, and Mubarak Shah. Test-time retrieval-augmented adaptation for vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8810–8819. IEEE, 2025.

Clément Fuchs, Maxime Zanella, and Christophe De Vleeschouwer. Online gaussian test-time adaptation of vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 128–137. IEEE, 2025.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. *Advances in Neural Information Processing Systems*, 36:14070–14093, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021a.

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

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171. IEEE, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17584–17594. IEEE, 2024.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36:76298–76310, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

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.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

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.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8769–8778. IEEE, 2018.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1802–1812. IEEE, 2023.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, and Nanyang Ye. Oodd: Test-time out-of-distribution detection with dynamic dictionary. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30630–30639. IEEE, 2025.

Jinlun Ye, Jiang Liao, Runhe Lai, Xinhua Lu, Jiaxin Zhuang, Zhiyong Gan, and Ruixuan Wang. Ttl: Test-time textual learning for ood detection with pretrained vision-language models. *arXiv preprint arXiv:2604.15756*, 2026.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided ood detection with vision-language models. *Advances in Neural Information Processing Systems*, 37:38744–38768, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

AI USE STATEMENT

Generative AI tools were used to assist in drafting portions of the manuscript and to improve the clarity, grammar, and presentation of the writing. All AI-assisted text was reviewed, revised, and verified by the authors. The authors take full responsibility for the final content of the paper.

A ADDITIONAL EXPERIMENTS AND ANALYSIS

A.1 ADDITIONAL IMAGENET DISTRIBUTION SHIFTS

Table 5 reports additional results on ImageNet-Sketch and ImageNet-V2 using the same evaluation protocol as the main ImageNet study. On ImageNet-Sketch, **OSGB** achieves the best average performance with 0.85% FPR95 and 99.54% AUROC, clearly outperforming the compared methods across all four OOD datasets. On ImageNet-V2, **OSGB** also obtains the best average result with 9.35% FPR95 and 97.01% AUROC, while remaining competitive across the individual OOD datasets.

Table 5: Additional OOD detection results on ImageNet-Sketch (ImageNet-K) and ImageNet-V2 using CLIP ViT-B/16. Each ID dataset is evaluated against iNaturalist, SUN, Places, and DTD. Results are reported as AUROC (%) and FPR95 (%). Best and second-best results are shown in bold and underlined, respectively.

ID Dataset	Method	iNaturalist		SUN		Places		DTD		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
ImageNet-K Wang et al. (2019)	MCM	88.033	63.220	84.952	68.880	80.953	71.530	75.277	78.067	82.304	70.424
	OODD	94.716	23.270	76.023	59.713	57.673	81.360	69.361	75.757	74.443	60.025
	NegLabel	99.354	2.200	95.172	21.850	91.022	37.830	89.317	46.348	93.716	27.057
	AdaND	94.985	29.373	88.378	54.167	83.976	66.530	77.783	70.981	86.280	55.263
	AdaNeg	99.823	0.577	97.567	9.137	94.057	36.693	<u>94.023</u>	<u>30.243</u>	<u>96.368</u>	19.163
	TTL	<u>99.843</u>	<u>0.303</u>	<u>99.235</u>	<u>2.223</u>	<u>97.503</u>	<u>10.487</u>	86.827	59.285	95.852	<u>18.075</u>
	OSGB	**99.989**	**0.010**	**99.750**	**0.440**	**99.620**	**0.610**	**98.790**	**2.340**	**99.540**	**0.850**
ImageNet-V2 Recht et al. (2019)	MCM	92.143	44.890	89.686	51.170	86.435	56.730	81.981	67.624	87.561	55.104
	OODD	99.416	1.900	95.772	19.407	88.674	42.267	<u>94.830</u>	23.942	94.673	21.879
	NegLabel	99.378	2.610	94.498	26.320	89.935	43.030	87.983	51.613	92.949	30.893
	AdaND	97.982	5.623	94.412	25.323	91.528	31.663	91.597	28.978	93.880	22.897
	AdaNeg	**99.747**	**0.390**	<u>97.020</u>	12.167	93.577	40.840	94.583	35.860	96.232	22.314
	TTL	99.318	1.723	**97.820**	**9.523**	**95.313**	<u>18.690</u>	94.617	<u>21.531</u>	<u>96.767</u>	<u>12.867</u>
	OSGB	<u>99.690</u>	<u>0.570</u>	96.870	<u>9.780</u>	<u>94.680</u>	**17.920**	**96.820**	**9.150**	**97.010**	**9.350**

A.2 ADDITIONAL ABLATION STUDIES

Effect of Adaptive-Z3. Table 6 evaluates the contribution of the three statistics used in Adaptive-Z3. Removing either $z_{\max}$ or LSE_+ causes a substantial performance drop, while removing $p_{\max}$ leads to a smaller degradation. The full Adaptive-Z3 achieves an FPR95 of 6.11% and an AUROC of 98.34%.

Table 6: Ablation of the Adaptive-Z3 positive bootstrap on ImageNet-1K. All variants use the same downstream adaptation and detection components.

Positive Bootstrap	FPR95 ↓	AUROC ↑
Semantic confidence	7.39	98.18
Adaptive-Z3 w/o $z_{\max}$	40.76	90.07
Adaptive-Z3 w/o $p_{\max}$	6.33	98.19
Adaptive-Z3 w/o LSE_+	49.78	85.83
Full Adaptive-Z3	**6.11**	**98.34**

Effect of Semantic Gaussian Decomposition. Table 7 compares Semantic Gaussian Banks with a single global Gaussian for online bank management while keeping the remaining adaptation and detection procedure unchanged. Semantic decomposition reduces average FPR95 from 6.42% to 6.11% and increases OOD coverage from 66.54% to 71.42%, while AUROC remains comparable. These results indicate that semantic decomposition improves online bank construction, particularly at the high-ID-recall operating point.

Table 7: Ablation of Gaussian bank structure on ImageNet-1K. Both variants use the same adaptation and final detection settings; only the bank structure differs.

Bank Structure	FPR95 ↓				Average		
	iNat	SUN	Places	DTD	FPR95 ↓	AUROC ↑	OOD Cov. ↑
Global Gaussian	0.47	6.60	12.10	6.50	6.42	98.33	66.54
Semantic Gaussian Banks	**0.42**	**5.93**	**12.09**	**6.01**	**6.11**	**98.34**	**71.42**

Purity–Coverage Analysis. Table 8 examines the roles of RAF and the 2σ rescue in OOD-bank construction. Adding RAF increases OOD-bank purity from 82.96% to 93.63%, while reducing coverage from 73.04% to 66.91%. The rescue recovers coverage to 71.42% while maintaining higher purity than GGF alone, leading to an FPR95 of 6.11% and an AUROC of 98.34%. These results show that RAF improves the reliability of OOD-bank updates, while the rescue recovers useful OOD samples rejected by RAF.

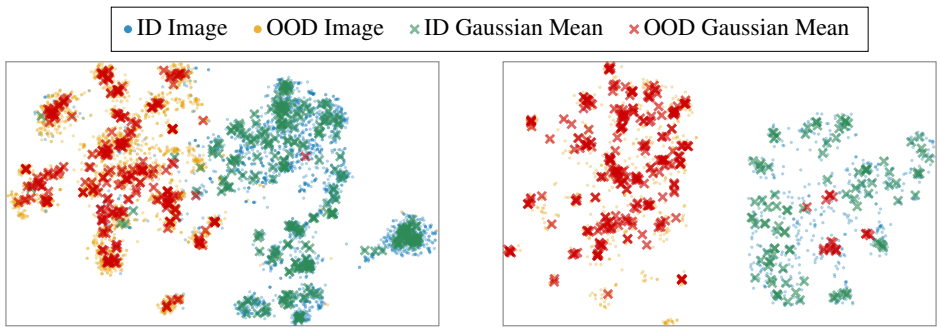
Table 8: Purity–coverage analysis of RAF and the 2σ rescue on ImageNet-1K.

Configuration	FPR95 ↓	AUROC ↑	OOD-Bank Purity ↑	OOD Coverage ↑
SGB + GGF	6.24	98.26	82.96	**73.04**
SGB + GGF + RAF	6.44	98.29	**93.63**	66.91
SGB + GGF + RAF + 2σ Rescue	**6.11**	**98.34**	87.83	71.42

Visualization of Semantic Gaussian Banks. Figure 6 presents t-SNE visualizations Van der Maaten & Hinton (2008) of the frozen CLIP image features and the Gaussian means learned by SGB on ImageNet-1K→SUN and CIFAR-100→Texture. In both settings, the learned ID Gaussian means mostly lie within the ID feature regions, while the OOD Gaussian means tend to follow the OOD feature regions. This provides qualitative evidence that the online Gaussian banks represent the corresponding ID and OOD feature distributions. Since the CLIP image encoder remains frozen, the learned Gaussian means come from the online bank updates rather than changes to the CLIP features.

A.3 Sensitivity Analysis

Sensitivity to Stream Composition. Figure 7 evaluates **OSGB** in a separate stream-composition sensitivity experiment under different ID:OOD ratios, averaged over four OOD datasets and three stream-order seeds. Performance is relatively stable from the ID-dominant $10{:}1$ setting through the balanced $1{:}1$ setting, with AUROC improving from 95.19% to 96.64% and FPR95 decreasing from



(a) ImageNet-1K → SUN **(b)** CIFAR-100 → Texture

Figure 6: t-SNE visualization of frozen CLIP image features and learned SGB Gaussian means on (a) ImageNet-1K→SUN and (b) CIFAR-100→Texture.

12.90% to 8.81%. Performance decreases under the more extreme OOD-dominated 1:5 stream, reaching 92.80% AUROC and 22.28% FPR95. Thus, persistent OOD dominance remains a challenging regime.

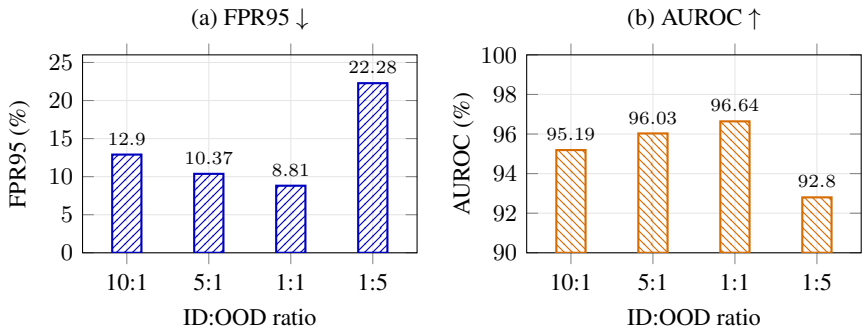


Figure 7: Stream-composition sensitivity of OSGB on ImageNet. Results are averaged over four OOD datasets and three stream-order seeds. OSGB remains stable from ID-dominant streams through balanced ID/OOD streams, while performance decreases when OOD samples strongly dominate the stream.

Sensitivity to Calibration-Prefix Length. Table 9 varies the number of initial stream samples used to bootstrap Adaptive-Z3. Performance is stable across prefix lengths from 64 to 512 samples: average FPR95 varies from 5.81% to 6.18%, while AUROC remains between 98.22% and 98.50%. Increasing the prefix to 1,024 samples produces a modest degradation to 6.70% FPR95 and 97.85% AUROC. The default prefix of 256 therefore lies within a broad region of stable performance.

Table 9: Sensitivity of OSGB to the Adaptive-Z3 calibration-prefix length. Results are averaged over four OOD datasets and three stream-order seeds.

Calibration prefix	FPR95 ↓	AUROC ↑
64	**5.81**	**98.50**
128	5.99	98.41
256	6.11	98.34
512	6.18	98.22
1024	6.70	97.85

A.4 Online Evaluation Protocol and Reproducibility

Stream Protocol. Table 10 summarizes the stream construction. For each ImageNet–OOD pair, we construct a finite online stream from the full 50,000-image ImageNet validation set and the corresponding fixed OOD evaluation pool. iNaturalist, SUN, and Places each contain 10,000 OOD samples, producing 60,000-sample streams with an ID:OOD ratio of 5:1. DTD contains 5,640 samples, producing a 55,640-sample stream with an ID:OOD ratio of approximately 8.87:1. The same samples are used across repeated runs, while only their stream order is changed. Each stream is traversed once without replacement. At test time, **OSGB** first produces the detection score for the current sample and only afterward allows that sample to affect the online banks or calibration state. Ground-truth ID/OOD labels are used only after inference to compute evaluation metrics and are never provided to the adaptation procedure.

Table 10: Exact ImageNet stream-generation protocol used for OSGB evaluation. The three seeds change only stream order; dataset composition is fixed. Each sample appears once and is scored before it can update the online state.

OOD dataset	ID	OOD	ID:OOD	OOD (%)
iNaturalist	50,000	10,000	5.00:1	16.67
SUN	50,000	10,000	5.00:1	16.67
Places	50,000	10,000	5.00:1	16.67
DTD	50,000	5,640	8.87:1	10.14

Leakage and Causality Audit. Table 11 summarizes checks of the online evaluation pipeline. Both semantic Gaussian banks are empty at the start of the stream. In a 4,000-sample ImageNet–SUN replay, flipping all evaluation labels produced identical detection scores and identical bank-update actions for every sample. Each sample is scored before it can update the banks, Adaptive-Z3 is fitted only from preceding stream samples, and no future samples are accessed. We also found zero literal matches to the evaluated OOD dataset names among the selected negative concepts. These checks support the causal implementation of the online evaluation protocol.

Table 11: Runtime leakage and initialization audit for OSGB. The label-independence test replays 4,000 samples from the ImageNet–SUN stream (seed 35) after flipping all evaluation labels.

Audit check	Result
Initial positive-bank size	0
Initial negative-bank size	0
Label-flip replay samples	4,000
Scores unchanged after label flip	100%
Update actions unchanged after label flip	100%
Current sample scored before update	Yes
Adaptive-Z3 uses past samples only	Yes
Future samples used	No
Benchmark OOD-name matches in negative concepts	0

Negative-Concept Reproducibility. Table 12 audits the complete negative semantic bank used by **OSGB** on ImageNet. The bank contains exactly 10,000 unique concepts, comprising 1,646 adjective-based and 8,354 noun-based concepts, with no missing or duplicated entries. All 10,000 prompts match their prescribed templates. The deterministic grouping produces five groups of exactly 2,000 concepts, with valid and non-duplicated within-group positions. Every negative concept has a valid ImageNet rival assignment, collectively covering 705 of the 1,000 ImageNet classes. The complete concept, prompt, group, and rival mapping is retained with the evaluation artifacts.

Table 12: Audit of the negative semantic bank used by OSGB on ImageNet. The bank contains 10,000 unique concepts with exact prompt, group, and ImageNet-rival assignments. All integrity checks pass.

Property	Audited result
Negative concepts	10,000
Unique concepts	10,000/10,000
Adjective / noun concepts	1,646/8,354
Correctly reconstructed prompts	10,000/10,000
Semantic groups	5
Concepts per group	2,000
Valid within-group positions	10,000/10,000
Duplicate group-position pairs	0
Valid ImageNet rival assignments	10,000/10,000
Distinct ImageNet rival classes	705/1,000 (70.5%)
Benchmark-specific identifier matches	0

A.5 TTL Reimplementation and Reproducibility

At the time of our experiments, an official implementation of TTL Ye et al. (2026) was not available to us. We therefore reimplemented TTL following the method, hyperparameters, and experimental details described in the paper and supplementary material. We use the same CLIP ViT-B/16 backbone and evaluation data employed throughout our experiments. All reproduced TTL results are averaged over the same three stream-order seeds $\{35, 62, 71\}$.

Table 13: Reproducibility analysis of TTL and comparison with our method. Results are reported as FPR95↓ / AUROC↑ (%). TTL Reproduced denotes our reimplementation averaged over three stream-order seeds. Δ denotes reproduced minus reported TTL performance.

ID Dataset	OOD Dataset	TTL Reported	TTL Reproduced	Δ (Reprod. − Reported)	OSGB
ImageNet	iNaturalist	**0.420 / 99.870**	1.180 / 99.589	+0.760 / −0.281	0.423 / 99.793
	SUN	7.180 / **98.450**	10.487 / 97.803	+3.307 / −0.647	**5.933** / 98.407
	Places	15.860 / 96.220	18.473 / 95.548	+2.613 / −0.672	**12.090 / 96.762**
	DTD	26.390 / 94.600	31.944 / 92.973	+5.554 / −1.627	**6.005 / 98.389**
	Average	**12.460 / 97.290**	**15.521 / 96.478**	**+3.061 / −0.812**	**6.113 / 98.338**
CIFAR-100	SVHN	**0.010 / 99.980**	0.363 / 99.702	+0.353 / −0.278	0.140 / 99.934
	LSUN-R	**2.520** / 99.470	2.570 / 99.272	+0.050 / −0.198	3.243 / **99.488**
	LSUN-C	**2.010** / 99.540	2.533 / 99.204	+0.523 / −0.336	3.120 / **99.570**
	iSUN	**1.270 / 99.670**	2.674 / 99.226	+1.404 / −0.444	3.686 / 99.425
	Texture	1.680 / 98.910	13.735 / 97.336	+12.055 / −1.574	**0.112 / 99.928**
	Places365	6.650 / 97.980	30.840 / 92.997	+24.190 / −4.983	**0.320 / 99.876**
	Average	**2.360 / 99.260**	**8.786 / 97.956**	**+6.426 / −1.304**	**1.770 / 99.703**

Reproducibility Discussion. As shown in Table 13, our implementation closely recovers the reported TTL results on ImageNet, with average FPR95/AUROC changing from 12.46/97.29 to 15.52/96.48. On CIFAR-100, the reproduced results are close on SVHN, LSUN-R, LSUN-C, and iSUN, while larger differences are observed on Texture and Places365, resulting in an average change from 2.36/99.26 to 8.79/97.96. Because the official implementation was unavailable, implementation details that could not be matched exactly may contribute to these differences. We therefore distinguish reproduced TTL results from those reported in the original paper.

Comparison with Reported TTL Results. Our conclusions do not depend on the difference between the reported and reproduced TTL results. Using the originally reported TTL results, our method still achieves better average performance on both benchmarks. On ImageNet, average FPR95 decreases from 12.46 to 6.11 and AUROC increases from 97.29 to 98.34. On CIFAR-100, average FPR95 decreases from 2.36 to 1.77 and AUROC increases from 99.26 to 99.70.

Reporting in the Main Tables. For the standard ImageNet-1K setting, we report our reproduced TTL results under the same stream construction and stream-order seeds used for the other methods. For CIFAR-100, we report the results provided in the original TTL paper. For the additional ImageNet-K, ImageNet-A, ImageNet-V2, and ImageNet-R settings, which were not evaluated in the original TTL paper, we report results from our reimplementation and denote them by †.

†TTL results for ImageNet-K, ImageNet-A, ImageNet-V2, and ImageNet-R are obtained using our reimplementation because these settings were not reported in the original TTL paper.