DP-BOA: Dirichlet-Process Birth-or-Assign for On-the-Fly Category Discovery

Peiyan Gu[1], Zixin Teng[1], and Xuming He[1,2⋆]

[1] ShanghaiTech University, Shanghai, China
[2] Shanghai Engineering Research Center of Intelligent Vision and Imaging
peiyangu@outlook.com, {tengzx2025,hexm}@shanghaitech.edu.cn

Abstract. On-the-fly category discovery requires deciding for each incoming test sample whether to assign it to an existing category or spawn a new one. Existing methods typically implement this decision through matching-based heuristics, such as radius- or hash-based rules. While effective in practice, these methods usually treat category birth implicitly as a fallback when no existing category matches confidently, rather than as an explicit alternative supported by its own statistical evidence. To address this, we propose DP-BOA, a posterior-predictive decision framework based on an online Dirichlet-process Gaussian mixture model with a Normal–Inverse–Wishart prior. During training, we use labeled data to calibrate a shared NIW prior over category Gaussians and warm-start the known-category posteriors. At test time, for each incoming sample, DP-BOA compares the posterior predictive evidence for assignment to existing categories against the evidence for spawning a new category induced by the DP prior, and then updates category statistics online after the decision. The method captures anisotropic category geometry and naturally adapts decision confidence as evidence accumulates. Across standard OCD benchmarks, DP-BOA consistently outperforms strong baselines and delivers particularly strong novel-class discovery performance while maintaining competitive known-class accuracy. The project page is available at DP-BOA.

Keywords: On-the-Fly Category Discovery· Generalized Category Discovery· Bayesian Nonparametrics

1 Introduction

Deep learning has achieved remarkable progress in visual recognition [10, 19, 25], but most models still assume a closed world, i.e., all test categories are seen during training. This assumption rarely holds in dynamic real-world environments. Open-world settings such as Novel Category Discovery (NCD) [17] and Generalized Category Discovery (GCD) [43] partially address this issue by using labeled known classes to organize and discover novel ones in an unlabeled set.

⋆ Corresponding author.

However, they typically assume offline access to the full unlabeled dataset, which is impractical in many applications (e.g., autonomous driving, robotics, live content moderation) where data arrive continuously and decisions must be made immediately. To address this gap, On-the-fly Category Discovery (OCD) [11] considers a strict streaming setting: samples arrive one by one, and the system must instantly decide whether each sample belongs to an existing category or should start a new one. This is particularly challenging because early mistakes can propagate and degrade later decisions.

A key challenge in OCD is the *assign-versus-birth* decision. It involves two coupled questions: how to define "new", and how to represent "old" under a stream. For the birth, the system needs a reliable criterion for deciding when a sample should start a new category, and labeled known categories are the main available source of prior guidance for this decision. For the assignment, category representations should evolve with the stream: they should be updated with newly assigned samples, remain uncertain when evidence is scarce, and become more confident as support accumulates. These suggest that an effective OCD method should combine an explicit birth criterion with online category representations whose uncertainty adapts to the amount of observed evidence.

Most existing OCD methods make online decisions via distance- or similarity-based matching rules [3, 11, 28, 57]: a sample is assigned to the nearest prototype if sufficiently close, otherwise it spawns a new one. While this design is simple and efficient, it treats category birth implicitly as a fallback from failed assignment, rather than as an explicit alternative supported by its own evidence. It also brings two limitations. First, such matching rules often induce prototype-centered regions of fixed shape that ignore direction-dependent covariance (e.g., Hamming balls [11, 57], Euclidean balls [28], or angular caps under cosine similarity [3]), which can be misaligned with heteroscedastic and anisotropic feature distributions (see Appendix M for empirical geometry statistics). Second, threshold control in these methods is typically fixed or heuristic and does not explicitly model how predictive uncertainty should contract as category evidence accumulates. As a result, it becomes harder to distinguish newly formed categories from stable ones, increasing the risk of error propagation after early false births.

To address these limitations, we propose **DP-BOA**, a probabilistic OCD framework that formulates assign-versus-birth as a Bayesian evidence comparison. For each incoming test sample, DP-BOA compares the prior-weighted predictive evidence for assignment to existing categories against that for spawning a new category, so that birth is no longer treated merely as rejection by existing categories but is explicitly evaluated within the same decision framework. We instantiate this idea with an online Dirichlet Process Gaussian Mixture Model (DP-GMM) [36, 46] and a Normal–Inverse–Wishart (NIW) prior [4]. Statistics from labeled known classes are used to initialize known-category posteriors and to estimate a shared prior over category Gaussians. Full-covariance category modeling captures anisotropic geometry, posterior-predictive updates allow predictive uncertainty to contract as evidence accumulates, and the DP prior enables open-ended category growth.

We validate these design choices through streaming experiments on standard OCD benchmarks, consistently outperforming baselines. Ablations further show that support-set prior calibration, the size- and concentration-dependent DP category prior, full-covariance modeling, and evidence-adaptive posterior updates each contribute meaningfully to the overall gains.

In summary, our contributions are:

- We reformulate OCD as a Bayesian evidence comparison between assigning a sample to an existing category and spawning a new one, making birth an explicit alternative rather than an implicit fallback from failed assignment.
- We instantiate this formulation with an online DP-GMM and an NIW prior, yielding an online posterior-predictive algorithm that leverages known-class statistics for prior initialization, captures anisotropic category geometry, and adapts predictive uncertainty as evidence accumulates.
- We demonstrate strong performance on standard OCD benchmarks, particularly for novel classes.

2 Related Work

2.1 Generalized / Novel Category Discovery

Novel Class Discovery (NCD) [17] aims to cluster novel classes in an unlabeled set by leveraging a labeled known-class set. Early work focused on pairwise similarity prediction for clustering [20, 21], while subsequent methods improved similarity estimation [16, 56], feature learning [29, 49, 58, 59], and clustering objectives [14, 53, 54]. Generalized Category Discovery (GCD) [43] extends NCD to unlabeled data containing both known and novel classes. Early GCD methods established strong pool-based baselines with contrastive representation learning and semi-supervised clustering [43, 44, 51, 55]. More recent work has explored several distinct directions, including clustering-oriented representation learning via mean-shift updates [7], efficient adaptation with spatial prompt tuning [48], robustness under domain shifts [47], debiased parametric learning with distribution guidance [30], and hierarchy-aware geometry [31]. Beyond centralized visual GCD, recent studies have also extended the problem to decentralized and multimodal settings. Fed-GCD [35] studies GCD under federated learning, where data are distributed across clients with heterogeneous label spaces, while language- or multimodal-assisted methods exploit vision-language models, LLM feedback, or multimodal alignment to improve category discovery [2, 33, 40]. Unlike these offline, pool-based methods, we target a single-pass streaming setting.

2.2 On-the-Fly Category Discovery

On-the-Fly Category Discovery (OCD) studies a single-pass test stream in which each incoming sample must be immediately assigned to an existing category or trigger a new one [11]. SMILE [11] introduces the task and uses instance-level hash codes with a Hamming-radius rule for online birth-or-assign decisions. PHE [57] improves this line with prototype-guided hash encoding to better preserve

category structure in Hamming space. More recent variants incorporate auxiliary priors: DiffGRE [28] leverages synthesized novel samples through a generate–refine–encode pipeline, while Sync [3] augments OCD with language-assisted representations and lightweight active querying. In contrast, we focus on the standard OCD setting without such auxiliary priors.

2.3 Bayesian Non-parametrics for Mixtures

Our approach builds on Dirichlet-process (DP) mixture models. Blackwell and MacQueen [5] showed that marginalizing a DP yields the Chinese Restaurant Process, whose concentration parameter governs new-component creation, and Sethuraman's stick-breaking construction provides a constructive representation [39]. Rasmussen's Infinite Gaussian Mixture Model [36] popularized DP-GMMs with Normal–Inverse–Wishart (NIW) priors and closed-form Student-t predictive densities under conjugacy. Subsequent work studied more scalable or streaming inference for non-parametric mixtures and clustering [9, 38, 46], while adjacent non-parametric deep clustering and open-world recognition methods explored related ideas in learned embedding spaces [37, 52]. Unlike these methods, which primarily address unsupervised density modeling or clustering, OCD requires single-pass online decisions with known class support-set supervision.

3 Method

3.1 Problem Formulation and Preliminaries

We follow the On-the-Fly Category Discovery (OCD) setting [11]. Let $\mathcal{D}_S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M} \subset \mathcal{X} \times \mathcal{Y}_S$ denote the labeled support set used for training, where $\mathcal{Y}_S$ is the set of known classes. For evaluation, the test stream is $\mathcal{D}_Q = \{(\mathbf{x}_j^q, y_j^q)\}_{j=1}^{N} \subset \mathcal{X} \times \mathcal{Y}_Q$, whose label space satisfies $\mathcal{Y}_Q = \mathcal{Y}_S \cup \mathcal{Y}_N$ with $\mathcal{Y}_S \cap \mathcal{Y}_N = \varnothing$, where $\mathcal{Y}_N$ denotes the novel classes. Following the standard OCD protocol [11, 57], only $\mathcal{D}_S$ is available during training; at test time, query samples are revealed sequentially, and their ground-truth labels are used only for evaluation. For each arriving sample $\mathbf{x}_t$, the model must decide whether to assign it to one of the existing categories or to declare the birth of a new one.

Feature Space. Our method operates in a feature space learned from the labeled support set. Specifically, a feature encoder f_θ maps each sample $\mathbf{x}_t$ to a d-dimensional representation $\mathbf{z}_t = f_\theta(\mathbf{x}_t)$. We train f_θ offline on the labeled support set D_S using the standard supervised cross-entropy loss, and freeze it during online inference. We deliberately adopt this simple and generic feature-learning setup to keep the focus of the paper on the birth-or-assign decision mechanism, rather than on task-specific representation objectives. All probabilistic decisions are then performed in this fixed feature space. Details are provided in Sec. 4.1.

3.2 A Probabilistic Framework for OCD

We cast the core *assign-versus-birth* decision in OCD as a Bayesian evidence comparison. At time t, let K_{t-1} be the number of existing categories after processing the first $t-1$ query samples, and let $\mathcal{D}_{t-1}$ denote the current online state.

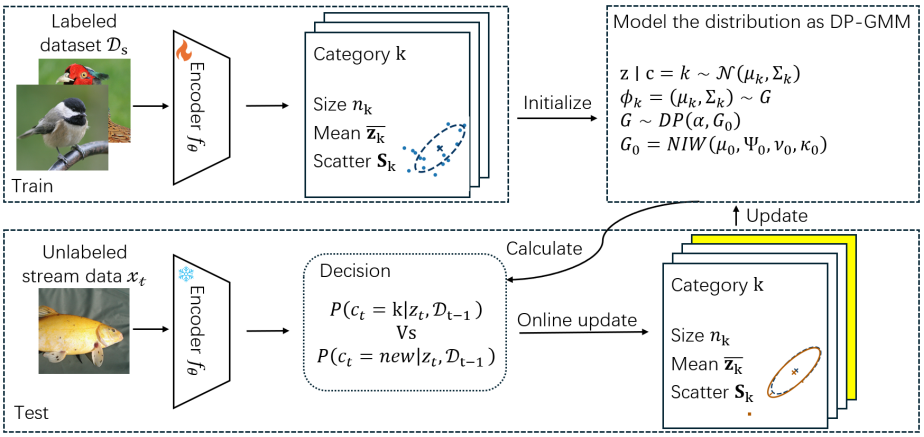

Fig. 1: Overview of our proposed DP-BOA framework. Our method consists of two phases. **(Top) Offline Initialization:** We first train a feature encoder f_θ on the labeled set $\mathcal{D}_S$. We then use the extracted features to compute the sufficient statistics $(n_k, \bar{\mathbf{z}}_k, \mathbf{S}_k)$ for all K_S known classes. These statistics are used to initialize our DP-GMM model (i.e., calibrate the global NIW hyperparameters from support-set statistics (Sec. 3.4)). **(Bottom) Online Inference:** At test time, a new sample $\mathbf{x}_t$ is passed through the frozen f_θ. Our model performs a posterior-predictive birth-or-assign decision (Eq. (12)), comparing the posterior probability of assigning $\mathbf{z}_t$ to an existing category ($P(c_t = k|...)$) versus birthing a new category ($P(c_t = \text{new}|...)$). Based on this decision, the statistics of the corresponding category are updated online.

For exposition, we write this state as a history of past observations, although in implementation it is represented only through per-category sufficient statistics. For an arriving feature vector $\mathbf{z}_t$, we choose a decision $c_t \in \{1, \ldots, K_{t-1}, \text{new}\}$ by maximizing the posterior:

$$\hat{c}_t = \arg \max_{k \in \{1,\ldots,K_{t-1},\text{new}\}} P(c_t = k \mid \mathbf{z}_t, \mathcal{D}_{t-1}). \tag{1}$$

By Bayes' rule, $P(c_t = k \mid \mathbf{z}_t, \mathcal{D}_{t-1}) \propto P(c_t = k \mid \mathcal{D}_{t-1})\, p(\mathbf{z}_t \mid c_t = k, \mathcal{D}_{t-1})$. To obtain a tractable online rule, we make two practical modeling approximations. For an existing category k, once $c_t = k$ is assumed, the predictive density depends only on that category's own history, giving $p(\mathbf{z}_t \mid c_t = k, \mathcal{D}_{t-1}) = p(\mathbf{z}_t \mid \mathcal{D}^k_{t-1})$. For a new category, we use the prior predictive induced by a shared global prior estimated offline, so $p(\mathbf{z}_t \mid c_t = \text{new}, \mathcal{D}_{t-1}) = p(\mathbf{z}_t \mid c_t = \text{new})$. This yields

$$\begin{aligned} \hat{c}_t = \arg\max \Big\{ & \max_{k \in \{1,\ldots,K_{t-1}\}} \big[P(c_t = k \mid \mathcal{D}_{t-1})\, p(\mathbf{z}_t \mid \mathcal{D}^k_{t-1})\big], \\ & \big[P(c_t = \text{new} \mid \mathcal{D}_{t-1})\, p(\mathbf{z}_t \mid c_t = \text{new})\big] \Big\}. \end{aligned} \tag{2}$$

The remaining task is to specify the two probabilistic components in Eq. (2): the category priors $P(c_t \mid \mathcal{D}_{t-1})$ and the predictive densities $p(\mathbf{z}_t \mid \cdot)$. The following section details our specifications for each.

3.3 Probabilistic Modeling via Adapted DP-GMM

As shown in Fig. 1, we instantiate the framework in Eq. (2) with a Dirichlet–Process Gaussian Mixture Model (DP-GMM) [36] and a Normal–Inverse–Wishart (NIW) prior. In contrast to standard DP-GMM usage, OCD provides a labeled support set $\mathcal{D}_S$ and requires single-pass online birth-or-assign decisions without revisiting past samples. We therefore adapt DP-GMM to this setting by using $\mathcal{D}_S$ to estimate a shared prior and initialize known categories. We briefly introduce DP-GMM in our method below.

Predictive Densities (Gaussian + NIW). To define the continuous density terms in Eq. (2) and model anisotropic category geometry, we represent each category k by a full-covariance Gaussian $\mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. We adopt a Bayesian formulation and place a conjugate Normal–Inverse–Wishart (NIW) prior over the Gaussian parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, parameterized by global hyperparameters $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0, \boldsymbol{\Psi}_0, \nu_0, \kappa_0)$. Here, $\boldsymbol{\mu}_0$ is the prior mean, $\boldsymbol{\Psi}_0$ is the scale matrix for covariance, ν_0 is the degrees-of-freedom parameter, and κ_0 controls the strength of the prior on the mean. A key property of this conjugate prior is that, after observing data, the posterior over $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ remains NIW. For an existing category k with data $\mathcal{D}_{t-1}^k$, the posterior NIW hyperparameters are $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k, \nu_k, \kappa_k)$. These can be computed in closed form from the prior $\boldsymbol{\theta}_0$ and the sufficient statistics of category k: the count $n_k = |\mathcal{D}_{t-1}^k|$, the sample mean $\bar{\mathbf{z}}_k$, and the scatter matrix $\mathbf{S}_k = \sum_{\mathbf{z} \in \mathcal{D}_{t-1}^k} (\mathbf{z} - \bar{\mathbf{z}}_k)(\mathbf{z} - \bar{\mathbf{z}}_k)^\top$. The exact NIW update equations are given in Appendix C. Integrating out $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ yields closed-form multivariate Student-t predictive densities t_d:

$$p(\mathbf{z}_t \mid \text{new}) = t_d\Big(\mathbf{z}_t \;\Big|\; \boldsymbol{\mu}_0,\, \frac{\kappa_0 + 1}{\kappa_0(\nu_0 - d + 1)} \boldsymbol{\Psi}_0,\, \nu_0 - d + 1\Big) \tag{3}$$

$$p(\mathbf{z}_t \mid \mathcal{D}_{t-1}^k) = t_d\Big(\mathbf{z}_t \;\Big|\; \boldsymbol{\mu}_k,\, \frac{\kappa_k + 1}{\kappa_k(\nu_k - d + 1)} \boldsymbol{\Psi}_k,\, \nu_k - d + 1\Big) \tag{4}$$

Because the posterior hyperparameters are updated with accumulated category evidence, the predictive density for category k becomes progressively sharper as n_k increases. This allows category uncertainty to contract naturally with support, rather than being handled only indirectly through heuristic matching rules. Appendix D provides the derivation and an illustration of how the predictive density evolves as evidence accumulates.

Category Priors (Dirichlet Process). To define the discrete prior probability terms in Eq. (2), we assume a Dirichlet Process (DP) prior [36], which nonparametrically allows for dynamically expandable category growth.

A standard result from the marginalization of the DP (the Blackwell MacQueen Polya Urn scheme [5, 13]) provides the exact prior probabilities for the decision c_t (see Appendix E for details):

$$P(c_t = k | \mathcal{D}_{t-1}) = \frac{n_k}{\alpha + N_{t-1}} \tag{5}$$

$$P(c_t = \text{new} | \mathcal{D}_{t-1}) = \frac{\alpha}{\alpha + N_{t-1}} \tag{6}$$

where n_k is the number of samples in category k, $N_{t-1} = \sum_k n_k$ is the total number of samples processed so far, and $\alpha > 0$ is the DP's concentration parameter, acting as a principled "innovation rate" for new categories.

3.4 Initialization from Known Categories

Unlike standard DP-GMM formulations, which are fully unsupervised, OCD provides labels for a subset of categories through the support set $\mathcal{D}_S$. We therefore use $\mathcal{D}_S$ to calibrate an EB-style NIW prior and to initialize posterior NIW states for known categories. This initialization plays two roles: (1) it uses $\mathcal{D}_S$ to calibrate a robust global prior $\boldsymbol{\theta}_0$ (for the "Birth"), and (2) it initializes the known categories with their posterior NIW statistics (for the "Assign").

Estimating global priors. Our global prior is defined by $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0, \boldsymbol{\Psi}_0, \nu_0, \kappa_0)$ and the DP concentration α. We use the $K_S = |\mathcal{Y}_S|$ known classes in $\mathcal{D}_S$ to estimate them. For each known category k, let $\mathcal{D}_k$ denote the set of support-set features belonging to that category. First, we define the necessary base statistics:

$$\bar{\mathbf{z}}_k = \frac{1}{n_k} \sum_{\mathbf{z} \in \mathcal{D}_k} \mathbf{z} \quad \text{(category mean)}$$

$$\mathbf{S}_k = \sum_{\mathbf{z} \in \mathcal{D}_k} (\mathbf{z} - \bar{\mathbf{z}}_k)(\mathbf{z} - \bar{\mathbf{z}}_k)^\top \quad \text{(category scatter)}$$

$$\bar{\mathbf{z}} = \frac{1}{M} \sum_{k=1}^{K_S} n_k \bar{\mathbf{z}}_k \quad \text{(global mean)}$$

$$\boldsymbol{\Sigma}_{\text{within}} = \frac{1}{M - K_S} \sum_{k=1}^{K_S} \mathbf{S}_k \quad \text{(pooled within-category covariance)}$$

$$\boldsymbol{\Sigma}_{\text{means}} = \frac{1}{K_S - 1} \sum_{k=1}^{K_S} (\bar{\mathbf{z}}_k - \bar{\mathbf{z}})(\bar{\mathbf{z}}_k - \bar{\mathbf{z}})^\top \quad \text{(covariance of category means)}$$

$$\overline{n^{-1}} = \frac{1}{K_S} \sum_{k=1}^{K_S} \frac{1}{n_k} \quad \text{(average inverse category size).}$$

We use these statistics to estimate the hyperparameters via an empirical-Bayes method [32](see Appendix F for derivations):

- **Prior mean ($\boldsymbol{\mu}_0$):** We set the prior mean to the global mean:

$$\boldsymbol{\mu}_0 = \bar{\mathbf{z}} \tag{7}$$

- **Covariance scale ($\boldsymbol{\Psi}_0$):** We choose the expected value of the prior covariance $\mathbb{E}[\boldsymbol{\Sigma}]$ to match the pooled within-category covariance:

$$\mathbb{E}[\boldsymbol{\Sigma}] = \boldsymbol{\Sigma}_{\text{within}} \quad \Rightarrow \quad \boldsymbol{\Psi}_0 = (\nu_0 - d - 1)\, \boldsymbol{\Sigma}_{\text{within}} \tag{8}$$

- **Mean strength (κ_0):** We estimate κ_0 by matching the trace of the observed covariance of category means $\boldsymbol{\Sigma}_{\text{means}}$ to its theoretical expectation under the NIW prior:

$$\boldsymbol{\Sigma}_{\text{means}} \;\approx\; \boldsymbol{\Sigma}_{\text{within}} \left(\overline{n^{-1}} + \frac{1}{\kappa_0} \right) \tag{9}$$

We use the trace approximation to solve this equation.

– **Tunable priors (ν_0, α):** While our EB-style calibration sets most NIW parameters ($\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Psi}_0$) from $\mathcal{D}_S$, two scalars must still be specified: the NIW degrees of freedom ν_0 and the DP concentration α.

We first focus on ν_0, which we reparameterize as $n_0 = \nu_0 - d - 1$ for interpretability. This n_0 primarily controls the prior predictive density $p(\mathbf{z} \mid \text{new})$ for the "Birth" hypothesis. Given $\boldsymbol{\Psi}_0 = n_0\, \boldsymbol{\Sigma}_{\text{within}}$, the prior-predictive scale $\mathbf{V}_0$ in Eq. (3) simplifies to:

$$\mathbf{V}_0 \;=\; \frac{\kappa_0 + 1}{\kappa_0(\nu_0 - d + 1)}\, \boldsymbol{\Psi}_0 \;=\; \frac{\kappa_0 + 1}{\kappa_0} \cdot \frac{n_0}{n_0 + 2}\, \boldsymbol{\Sigma}_{\text{within}} \tag{10}$$

and the predictive degrees of freedom in Eq. (3) become $\nu_0' = \nu_0 - d + 1 = n_0 + 2$. Thus, n_0 jointly tunes the "peakiness" (via $\mathbf{V}_0$) and "tail thickness" (via ν_0') of the evidence required to birth a new category. Since a single fixed constant (e.g., $n_0 = 10$) fails to generalize across datasets with different scales (see Tab. 4), we propose a data-driven heuristic:

$$n_0 \;=\; \min\!\big(\tfrac{1}{2}\,\bar{n},\; n_{\text{cap}}\big)\,, \qquad \bar{n} = \frac{1}{K_S}\sum_{k=1}^{K_S} n_k, \tag{11}$$

where n_{cap} is a fixed hyperparameter that limits the maximum prior strength. This rule is guided by two principles: (1) tying n_0 to $\bar{n}$ interprets it as a *pseudo-sample size*, calibrating the prior's strength to the scale of the observed data; and (2) the cap at n_{cap} prevents an over-confident prior on datasets with large $\bar{n}$. This is particularly important in OCD, where an overly strong prior derived from $\mathcal{D}_S$ could suppress the discovery of novel categories with different covariance structures. Empirically, Eq. (11) yields consistently strong performance and obviates per-dataset tuning.

For the DP concentration α, we find that model performance is highly robust to its value. We therefore set a single constant across all datasets; detailed sensitivity analyses of ν_0 and α are provided in Sec. 4.3.

Initializing known-category posteriors. Finally, we initialize the first $K_S = |\mathcal{Y}_S|$ known categories. Instead of starting them as empty components, we use their sufficient statistics $(n_k, \bar{\mathbf{z}}_k, \mathbf{S}_k)$ computed from $\mathcal{D}_S$ to derive their posterior NIW parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k, \nu_k, \kappa_k)$ using the update rules in Appendix C. These K_S initialized categories form the initial set for the online assignment process.

This "warm start" ensures not only that the known "Assign" categories are modeled with high confidence from the outset, but also that the "Birth" hypothesis is governed by a meaningful, data-driven standard (the estimated $\boldsymbol{\theta}_0$).

3.5 Online Inference and Update

With the K_S known categories and the global prior initialized, the model begins processing the query stream $\mathcal{D}_Q$. For each arriving sample $\mathbf{z}_t$, which could belong to either an existing category or a novel one, the model must perform the two-stage process: (1) Inference and (2) Update.

Decision Rule. First, the model executes the Bayesian evidence comparison (Eq. (2)) to determine if $\mathbf{z}_t$ belongs to an existing category (either one of the K_S known ones or a previously discovered novel one) or represents a new, unseen category. We substitute the components from Sec. 3.3, yielding the decision $\hat{c}_t$:

$$\arg\max \Big\{ \big[\alpha \cdot t_d(\mathbf{z}_t|\boldsymbol{\mu}_0, \frac{\kappa_0+1}{\kappa_0(\nu_0-d+1)}\boldsymbol{\Psi}_0, \nu_0-d+1)\big], \\ \max_{k\in\{1\dots K_{t-1}\}} n_k t_d(\mathbf{z}_t|\boldsymbol{\mu}_k, \frac{\kappa_k+1}{\kappa_k(\nu_k-d+1)}\boldsymbol{\Psi}_k, \nu_k-d+1)\big] \Big\} \tag{12}$$

Online Update. Second, based on the decision $\hat{c}_t$, the model updates its state online. This update step is critical for the OCD setting, as it allows the model to learn from and refine its understanding of novel categories as they emerge. Due to the conjugacy of the NIW prior, we only need to update the relevant statistics for the chosen category.

If Assign ($\hat{c}_t = k$): We update the sufficient statistics $(n_k, \bar{\mathbf{z}}_k, \mathbf{S}_k)$ of the chosen category k using standard online update rules [50] (see Appendix G):

$$n_k^{\text{new}} = n_k^{\text{old}} + 1 \tag{13}$$

$$\bar{\mathbf{z}}_k^{\text{new}} = \bar{\mathbf{z}}_k^{\text{old}} + \frac{1}{n_k^{\text{new}}}(\mathbf{z}_t - \bar{\mathbf{z}}_k^{\text{old}}) \tag{14}$$

$$\mathbf{S}_k^{\text{new}} = \mathbf{S}_k^{\text{old}} + (\mathbf{z}_t - \bar{\mathbf{z}}_k^{\text{old}})(\mathbf{z}_t - \bar{\mathbf{z}}_k^{\text{new}})^\top \tag{15}$$

If Birth ($\hat{c}_t = \textbf{new}$): We increment the category count $K_t = K_{t-1} + 1$, and initialize this new category $k' = K_t$ with its own sufficient statistics:

$$n_{k'} = 1, \quad \bar{\mathbf{z}}_{k'} = \mathbf{z}_t, \quad \mathbf{S}_{k'} = \mathbf{0} \tag{16}$$

These updated statistics $(n_{\hat{c}_t}, \bar{\mathbf{z}}_{\hat{c}_t}, \mathbf{S}_{\hat{c}_t})$ are then used to compute the category's posterior parameters $(\boldsymbol{\mu}_{\hat{c}_t}, \boldsymbol{\Psi}_{\hat{c}_t}, \nu_{\hat{c}_t}, \kappa_{\hat{c}_t})$ via the NIW update rules in Appendix C, which in turn define the category's predictive density $p(\mathbf{z}_{t+1}|\mathcal{D}_{\hat{c}_t})$ in subsequent decisions, thereby fulfilling the "evidence-adaptive" requirement.

Complexity. Since DP-BOA replaces lightweight Hamming/radius checks with full-covariance Student-t predictive densities and DP-style updates, its online head incurs additional overhead. Let d be the feature dimension and K the number of active categories. Because the backbone cost is the same as in prior OCD methods, we focus on the online head. For each incoming feature, evaluating full-covariance Student-t predictive densities for all existing categories plus the "new" costs $O(Kd^2)$ time, and updating the sufficient statistics and NIW posterior for the selected category has worst-case cost $O(d^3)$. Thus, DP-BOA has per-sample complexity $O(d^3 + Kd^2)$ time with $O(Kd^2)$ memory. In practice, we cache covariance inverses and log-determinants, so only the affected category is recomputed after each update. To improve scalability, we further consider DP-BOA-L, a lightweight variant that replaces full covariance with a rank-r approximation ($r \ll d$) maintained by Frequent Directions [27]:

$$\Sigma_k \approx \sigma_k^2 I + U_k \operatorname{diag}(\lambda_k) U_k^\top, \quad U_k \in \mathbb{R}^{d\times r},\ r \ll d. \tag{17}$$

Here, the columns of U_k are the r approximate dominant covariance directions, λ_k contains the corresponding approximate eigenvalues, and σ_k^2 is the scalar residual variance used to summarize the covariance outside the low-rank subspace. With fixed small r, its per-sample complexity becomes $O(Kdr+dr^2)$ time with $O(Kdr)$ memory, reducing the dependence on representation dimensionality from quadratic to linear. This preserves the same probabilistic birth-or-assign principle while bringing DP-BOA-L to the same linear order in both representation dimensionality and the number of categories as recent OCD heads [3,28,57]. DP-BOA-L thus serves as a scalability-oriented approximation to full DP-BOA: it preserves the same online birth-or-assign decision rule, while the low-rank covariance representation substantially reduces latency and memory at the cost of only a small accuracy drop. Appendix K provides the implementation and detailed complexity analysis of DP-BOA-L.

4 Experiments

4.1 Setup

Datasets. Following [11, 57], we conduct experiments on multiple datasets, including two generic datasets—CIFAR100 [24] and ImageNet100 [8]—as well as eight fine-grained datasets: CUB [45], Stanford Cars [23], Herbarium19 [41], Oxford-IIIT Pet [34], and four super-categories from iNaturalist [42], including Fungi, Arachnida, Animalia, and Mollusca. Following OCD [11], the categories of each dataset are split into subsets of known and novel classes. Specifically, 50% of the samples from the known classes are used to form the labeled set $\mathcal{D}_S$ for training, while the remainder forms the query stream $\mathcal{D}_Q$ for on-the-fly testing. The details of the datasets are provided in Appendix A.

Evaluation Protocol. We follow the protocol [43,57] and evaluate using clustering accuracy on unlabeled datasets. This metric is calculated by finding an optimal one-to-one mapping $\mathcal{P}$ between predicted cluster indices $\hat{y}_i$ and ground-truth labels y_i with the Hungarian algorithm [26]: $\text{ACC} = \frac{1}{N}\sum_{i=1}^{N} 1\{y_i = \mathcal{P}(\hat{y}_i)\}$. We report model performance separately for known, novel, and all classes.

Implementation Details. We implement our method in PyTorch and run all experiments on NVIDIA TITAN RTX GPUs. Following OCD [11,28,57], we use a DINO-pretrained ViT-B/16 backbone [6] and fine-tune only the last transformer block for 100 epochs. We train with SGD [1] (momentum 0.9) using a cosine learning-rate schedule from 1.0 down to 10^{-4} and a batch size of 256. Unlike prior OCD methods [3,11,28,57], we do not rely on sophisticated representation learning. We simply train the encoder with standard supervised cross-entropy on the labeled support set and keep it frozen during online inference, so that any performance gain mainly comes from our probabilistic birth-or-assign head. After training, we fix the backbone as the encoder f_θ with feature dimension $d = 768$, and set $\alpha = 10^{-9}$ and $n_{\text{cap}} = 50$. All results are averaged over 3 runs with different seeds.

Table 1: Comparison with State-of-the-Art methods on the first five datasets (Part I). We report accuracy for All, Known, and Novel classes. **Bold** indicates the best performance. Rows highlighted in gray (marked with †) use external knowledge.

Method	Animalia			Arachnida			CIFAR100			CUB			Fungi		
	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
SLC [18]	32.4	61.9	19.3	25.4	44.6	11.4	44.4	59.0	15.1	28.6	44.0	20.9	27.7	60.0	13.4
RankStat [16]	31.4	54.9	21.6	26.6	51.0	10.0	35.0	44.0	17.0	21.2	26.9	18.4	23.8	50.5	12.0
WTA [22]	33.4	59.8	22.4	28.1	55.5	10.9	40.8	52.9	16.7	21.9	26.9	19.4	27.5	65.5	12.0
SMILE [11]	35.9	49.4	30.3	29.9	57.9	12.2	51.6	61.5	31.7	32.2	50.9	22.9	29.3	64.6	13.6
PHE [57]	40.3	55.7	31.8	37.0	75.7	12.6	57.4	72.1	27.9	36.4	55.8	27.0	31.4	**67.9**	15.2
DiffGRE† [28]	43.5	63.2	35.3	47.7	**76.6**	29.4	-	-	-	42.5	54.4	36.5	-	-	-
Sync† [3]	-	-	-	-	-	-	56.1	68.4	31.5	45.3	54.3	40.9	-	-	-
DP-BOA	**50.7**	**67.6**	**43.7**	**50.6**	53.4	**49.4**	**60.7**	**75.0**	**32.0**	**53.4**	**57.2**	**51.6**	**51.8**	65.0	**46.0**

Table 2: Comparison with SOTA methods on the remaining five datasets (Part II).

Method	Herbarium19			ImageNet100			Mollusca			OxfordPets			StanfordCars		
	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
SLC [18]	14.9	27.4	8.1	32.9	**86.5**	5.2	31.1	59.8	15.0	35.5	41.3	33.1	14.0	23.0	9.7
RankStat [16]	13.8	20.6	10.2	31.1	73.3	9.8	29.3	55.2	15.5	33.2	42.3	28.4	14.8	19.9	12.3
WTA [22]	14.6	21.2	11.1	30.8	72.9	19.4	30.3	55.4	17.0	35.2	46.3	29.3	17.1	24.4	13.6
SMILE [11]	22.9	39.3	14.1	33.8	74.2	13.4	33.3	44.5	27.2	41.2	42.1	40.7	26.2	46.6	16.2
PHE [57]	22.6	40.5	12.9	34.0	80.2	10.9	39.9	**65.0**	26.5	48.3	53.8	45.4	31.3	**61.9**	16.8
DiffGRE† [28]	-	-	-	-	-	-	42.6	62.0	32.3	49.6	50.1	49.3	27.7	48.1	17.8
Sync† [3]	-	-	-	**44.0**	86.2	**22.8**	-	-	-	**61.6**	**69.5**	**57.5**	24.6	34.8	19.5
DP-BOA	**30.2**	**46.4**	**21.5**	33.8	75.8	12.7	**48.9**	53.9	**46.3**	59.0	63.6	56.6	**32.4**	58.8	**19.7**

4.2 Comparison with the State of the Art

Compared Methods. In Tabs. 1 and 2, we compare DP-BOA against representative methods on ten benchmarks: three from adjacent settings that are standard OCD baselines (SLC [18], RankStat [16], WTA [22]) and four OCD-specific methods (SMILE [11], PHE [57], DiffGRE† [28], Sync† [3]). Methods marked with † use *external knowledge* (e.g., diffusion models or CLIP+text), making the comparison conservative for DP-BOA. For DiffGRE, we report its strongest variant —PHE + DiffGRE with online clustering inference—which achieves the highest reported average. "–" denotes metrics not reported in the original works.

Results. Without using any external knowledge, DP-BOA attains the best *All* score on **8/10** datasets. It is particularly strong on novel classes: on CUB, DP-BOA reaches **51.6%** Novel accuracy, a **+10.7** point gain over the next-best method (Sync), and on Mollusca it achieves **46.3%** vs **32.3%** for DiffGRE (**+14.0**). This indicates that our probabilistic framework is highly effective for the core "on-the-fly discovery" task. Sync obtains higher *All* accuracy on ImageNet100 and OxfordPets, but relies on CLIP and text. Among methods that do not use external knowledge, DP-BOA (33.8%) is competitive with PHE (34.0%) on ImageNet100 and clearly outperforms PHE on OxfordPets (59.0% vs 48.3%). Overall, DP-BOA sets a new state-of-the-art among OCD methods that operate purely on the visual stream, offering stronger and more balanced novel-category discovery without external knowledge bases.

Table 3: Ablation on probabilistic design choices in DP-BOA. Each row disables one component while keeping others fixed: *w/o* μ_0: zero prior mean (no data–driven centering); *w/o* Ψ_0: identity prior covariance (no scale matching); *w/o* κ_0: set κ_0=1 (no EB estimation of κ_0); *w/o DP prior*: remove the DP category prior terms and score hypotheses using predictive densities only; *w/o adapt*: freeze NIW posteriors at test time (no evidence-adaptive updating); *Spherical*: use an isotropic covariance instead of a full matrix. Best in **bold**.

Method	Animalia			CUB			OxfordPets		
	All	Known	Novel	All	Known	Novel	All	Known	Novel
w/o μ_0	50.1	65.4	**43.7**	52.9	51.1	**53.8**	58.1	63.0	55.5
w/o Ψ_0	47.9	70.6	38.4	40.8	39.8	41.3	56.0	52.0	**58.2**
w/o κ_0	45.2	59.1	39.4	52.2	51.7	52.4	57.5	60.1	56.2
w/o DP prior	48.0	**71.3**	38.4	52.3	**57.3**	49.8	57.1	62.7	54.2
w/o adapt	40.2	63.7	30.5	35.9	53.6	27.0	54.7	55.8	54.2
Spherical	41.7	61.7	33.5	40.4	49.0	36.1	47.1	45.5	47.9
DP-BOA	**50.7**	67.6	**43.7**	**53.4**	57.2	51.6	**59.0**	**63.6**	56.6

4.3 Ablation and Analysis

Ablation on probabilistic design choices. We ablate our probabilistic design by disabling one component at a time while keeping all others fixed (Tab. 3).

Prior initialization. The first three rows remove individual support-set calibration components of the birth prior. Overall, the full calibrated initialization (DP-BOA) attains the best *All* accuracy on all three datasets. Removing the data-driven prior mean μ_0 has only a mild effect, but discarding the learned covariance scale Ψ_0 or mean strength κ_0 is clearly harmful. In particular, *w/o* Ψ_0 produces the largest drops (e.g., CUB *All* 53.4→40.8), and *w/o* κ_0 consistently reduces *All* by 5.5/1.2/1.5 points, indicating that EB scale matching and mean shrinkage are key to a well-calibrated prior.

Dirichlet–process prior. Removing the DP prior in Eqs. (5) and (6) (*w/o DP prior*) and scoring categories purely by predictive density—treating all existing clusters as equally likely a priori—degrades performance on all datasets. Novel accuracy drops from 43.7 to 38.4 on Animalia and from 56.6 to 54.2 on OxfordPets, suggesting that the DP size- and concentration-dependent terms help stabilize the birth rate and avoid over-fragmenting categories.

Evidence-adaptive updates. Freezing NIW posteriors at test time (*w/o adapt*) is particularly damaging for novel-class discovery. Novel accuracy drops from 43.7→30.5 on Animalia and 51.6→27.0 on CUB, while Known accuracy remains relatively high, showing that continuously updating cluster posteriors with incoming evidence is crucial.

Covariance geometry. Finally, we compare our full-covariance model with the *Spherical* variant that enforces isotropy for both prior and posteriors. Concretely, we (i) match the prior scale to the full model by setting $\Psi_0^{\mathrm{sph}} = (\nu_0 - d - 1)\,\bar{\sigma}^2\mathbf{I}$, where $\bar{\sigma}^2 = \frac{1}{d}\mathrm{tr}(\Sigma_{\mathrm{within}})$, and (ii) after each NIW update, project the posterior to spherical by $\Psi_k^{\mathrm{sph}} \leftarrow \big(\frac{1}{d}\mathrm{tr}(\Psi_k)\big)\mathbf{I}$. Even under this calibrated construction, Spherical lags far behind Full: *All* drops by 9.0/13.0/11.9 points

Table 4: Sensitivity to n_0. Bold = best per column, underlined = second-best. Row $\bar{n}/2^*$ is our default rule $n_0 = \min(\bar{n}/2, n_{\text{cap}})$.

n_0	Animalia ($\bar{n}/2$=19)			CUB ($\bar{n}/2$=7)			OxfordPets ($\bar{n}/2$=24)		
	All	Known	Novel	All	Known	Novel	All	Known	Novel
5	40.7	60.7	32.5	**55.4**	**60.0**	**53.2**	58.5	62.0	56.6
10	46.5	63.3	39.5	50.7	52.3	49.9	57.8	55.6	58.9
15	48.0	**69.7**	39.0	47.6	47.4	46.1	58.4	55.4	60.0
20	49.9	66.6	42.9	46.2	43.0	48.8	58.7	61.2	57.5
25	**51.4**	63.5	**46.4**	42.6	41.7	43.0	**60.8**	60.9	**61.2**
50	46.7	58.9	41.6	39.9	38.3	40.7	54.4	63.2	49.8
$\bar{n}/2^*$	50.7	67.6	43.7	53.4	57.2	51.6	59.0	**63.6**	56.6

Table 5: Sensitivity to α. Bold = best per column, underlined = second-best. The row marked * is our default.

$\log\alpha$	Animalia			CUB			OxfordPets		
	All	Known	Novel	All	Known	Novel	All	Known	Novel
-6	48.8	65.4	41.9	52.4	53.8	51.7	61.0	60.9	**61.1**
-7	49.5	67.1	42.3	53.0	53.3	**52.9**	**61.4**	62.9	60.6
-8	50.2	**68.5**	42.7	52.8	55.0	51.6	59.7	63.2	57.9
-9^*	50.7	67.6	43.7	53.4	**57.2**	51.6	59.0	63.6	56.6
-10	**51.2**	66.8	**44.7**	52.9	55.1	51.7	58.9	64.8	56.2
-11	49.0	64.0	42.9	53.2	56.8	51.4	59.0	65.0	55.9
-12	50.6	66.7	43.9	**53.5**	56.1	52.2	58.3	**65.1**	54.7

and *Novel* by 10.2/15.5/8.7 on Animalia/CUB/OxfordPets. This highlights that modeling anisotropic category shapes is essential for accurate posterior predictives and reliable birth-or-assign decisions.

Sensitivity to n_0 and α. The reparameterization $n_0 = \nu_0 - d - 1$ controls the tail heaviness of the Student-t prior predictive, while the DP concentration α sets the prior odds of birthing a new category. Very small n_0 makes the birth predictive diffuse, while large n_0 concentrates it around the global prior; similarly, too small α suppresses novel-class creation, whereas too large α can over-fragment the stream. In Tab. 4, the data-scaled rule $n_0 = \min(\bar{n}/2, n_{\text{cap}})$ remains close to the best *All* accuracy, within 0.7, 2.0, and 1.8 points on Animalia, CUB, and OxfordPets. Table 5 shows similar stability for α: the default $\log\alpha = -9$ is within 0.5, 0.1, and 2.4 *All* points of the best value on the same datasets. Although more sophisticated dataset- or stream-adaptive schedules may exist, we adopt these rules for their simplicity, stability, and near-optimal performance.

Cluster Growth and Estimated Number of Classes. We analyze the evolution of the number of discovered clusters along the OCD stream. Fig. 2 plots $K(t)$ on CUB, showing that DP-BOA exhibits smooth, controlled growth that broadly tracks the ground-truth trend while remaining slightly conservative. Consistently, Tab. 6 reports K_{final}=175/177 on CUB/StanfordCars versus 200/196 ground-truth classes, substantially closer than other baselines, indicating a more calibrated birth rate without severe stream fragmentation.

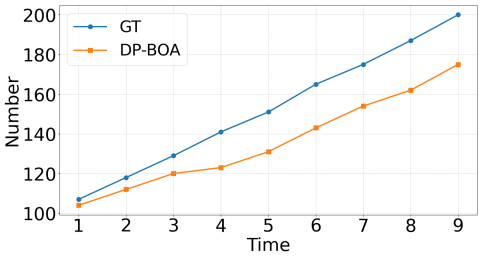

Fig. 2: Cluster growth on CUB.

Table 6: Final cluster count.

Method	CUB	StanfordCars
SMILE-16bit	924	896
PHE-16bit	318	709
DiffGRE	116	171
Sync-AL	255	447
DP-BOA	175	177
Ground Truth	200	196

Table 7: Runtime and accuracy–efficiency trade-off. Accuracy, maximum per-sample latency (Lat.) in milliseconds, and head memory (Mem.) in MB. Numbers in parentheses denote the number of categories.

	Animalia (77)					CUB (200)					StanfordCars (196)				
Method	All	Known	Novel	Lat.	Mem.	All	Known	Novel	Lat.	Mem.	All	Known	Novel	Lat.	Mem.
DP-BOA	**50.7**	**67.6**	**43.7**	46.1	299.2	**53.4**	**57.2**	**51.6**	81.7	619.2	**32.4**	**58.8**	19.7	82.9	626.7
DP-BOA-L	49.2	64.2	42.9	**34.0**	**40.5**	52.2	54.6	50.9	**50.3**	**51.0**	31.5	54.7	**20.2**	**44.2**	**49.7**

Efficiency. To complement the theoretical complexity in Sec. 3.5, we report the wall-clock cost and head memory footprint of DP-BOA and its low-rank variant in Tab. 7. Although full DP-BOA uses full-covariance posterior predictive and therefore incurs a quadratic memory cost in the feature dimension, its absolute overhead remains moderate in the OCD setting: the maximum per-sample latency is below 83 ms on the reported benchmarks, and the head memory remains below 0.7 GB. Overall, DP-BOA trades additional computation and memory for accuracy, but the resulting overhead remains moderate and practically acceptable in the OCD setting.

To further mitigate the cost of full-covariance modeling, we evaluate DP-BOA-L. As shown in Tab. 7, DP-BOA-L preserves most of the accuracy of full DP-BOA while substantially reducing both latency and head memory. Its All accuracy drops by only 1.5, 1.2, and 0.9 points on Animalia, CUB, and StanfordCars, respectively, while head memory is reduced from 299.2/619.2/626.7 MB to 40.5/51.0/49.7 MB and maximum latency decreases from 46.1/81.7/82.9 ms to 34.0/50.3/44.2 ms. The reduction is especially clear on CUB and StanfordCars, where the larger number of categories makes the full-covariance head more expensive. Overall, DP-BOA-L offers a favorable accuracy-efficiency trade-off and provides a practical approximation for larger-K or resource-constrained settings.

Robustness to stream order. We use the standard OCD protocol following prior work [57] and evaluate all methods on the same query stream for controlled comparison. To test whether DP-BOA relies on this ordering, we construct bursty-correlated streams by shuffling samples within each ground-truth class, grouping them into chunks of up to $B = 10$ samples, and then randomly shuffling these chunks, while keeping the same support/query split, query set, and evaluation metric. This induces stronger local temporal correlation than the standard protocol. As shown in Tab. 8, DP-BOA outperforms SMILE and PHE in *All* and

Table 8: Robustness to bursty-correlated stream orders. We use local class bursts of size $B = 10$ to induce strong temporal correlation. Best results are in **bold**.

	Pets			StanfordCars			CUB		
Method	All	Known	Novel	All	Known	Novel	All	Known	Novel
SMILE	40.0	46.2	36.9	25.1	42.0	16.9	30.9	50.0	21.2
PHE	47.0	45.5	47.9	30.5	**62.0**	15.3	33.8	45.9	27.8
DP-BOA	**58.3**	**57.7**	**58.5**	**35.0**	57.3	**24.2**	**50.5**	**56.8**	**47.3**

Novel accuracy on all three datasets, with a particularly large gain on CUB ($+16.7/+19.5$) points in *All*/*Novel* accuracy over PHE, suggesting that its gains do not rely on a favorable random stream order.

5 Conclusion

We revisited on-the-fly category discovery from the perspective of probabilistic decision-making. Instead of relying on heuristic matching rules, DP-BOA formulates the assign-versus-birth choice as a Bayesian evidence comparison, combining a Dirichlet-process prior for open-ended category growth with NIW posterior predictives for geometry-aware and evidence-adaptive online decisions. By leveraging labeled known classes to initialize both the global prior and known-category posteriors, DP-BOA provides a new framework for streaming category discovery and achieves strong performance across standard OCD benchmarks.

Limitations and future work. DP-BOA is designed for the conventional OCD setting, where online decisions must be made efficiently without revisiting past samples. First, given the limited auxiliary information in this protocol, we use a small set of hyperparameters and simple fixed defaults, a common practice in Bayesian nonparametric mixture models [12, 15]; these defaults already yield strong empirical performance. In richer scenarios, extra signals (e.g., prior knowledge about class granularity or semantic side information) could potentially enable more adaptive hyperparameter calibration. Second, for efficiency, each category in our method is modeled by a single elliptical component, which provides a strong accuracy–efficiency trade-off on diverse benchmarks. When streams exhibit more complex intra-class structure, such as multi-modality or non-elliptical geometry induced by domain shift, more complex and expressive category models (e.g., multi-component or non-Gaussian likelihoods) may further improve robustness [4, 36]. However, such expressiveness typically increases latency and memory due to additional state maintenance, and may also introduce extra parameters. A future direction is to incorporate richer category models into our probabilistic framework while reducing these overheads.

Acknowledgements

This work was supported by the MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), and the Shanghai Key Technology R&D Program No. 25JC3200500.

References

1. Amari, S.i.: Backpropagation and stochastic gradient descent method. Neurocomputing **5**(4-5), 185–196 (1993)
2. An, W., Shi, W., Tian, F., Lin, H., Wang, Q., Wu, Y., Cai, M., Wang, L., Chen, Y., Zhu, H., et al.: Generalized category discovery with large language models in the loop. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 8653–8665 (2024)
3. Banerjee, A., Biswas, S.: Language-assisted feature representation and lightweight active learning for on-the-fly category discovery. Transactions on Machine Learning Research (2025)
4. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
5. Blackwell, D., MacQueen, J.B.: Ferguson distributions via pólya urn schemes. The annals of statistics **1**(2), 353–355 (1973)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Choi, S., Kang, D., Cho, M.: Contrastive mean-shift learning for generalized category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23094–23104 (2024)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Dinari, O., Freifeld, O.: Sampling in dirichlet process mixture models for clustering streaming data. In: International Conference on Artificial Intelligence and Statistics. pp. 818–835. PMLR (2022)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Du, R., Chang, D., Liang, K., Hospedales, T., Song, Y.Z., Ma, Z.: On-the-fly category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11691–11700 (2023)
12. Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association **90**(430), 577–588 (1995)
13. Ferguson, T.S.: A bayesian analysis of some nonparametric problems. The annals of statistics pp. 209–230 (1973)
14. Fini, E., Sangineto, E., Lathuilière, S., Zhong, Z., Nabi, M., Ricci, E.: A unified objective for novel class discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9284–9292 (2021)
15. Gershman, S.J., Blei, D.M.: A Tutorial on Bayesian Nonparametric Models. Journal of Mathematical Psychology **56**(1), 1–12 (2012)
16. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Autonovel: Automatically discovering and learning novel visual categories. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
17. Han, K., Vedaldi, A., Zisserman, A.: Learning to discover novel visual categories via deep transfer clustering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8401–8409 (2019)
18. Hartigan, J.A.: Clustering algorithms. John Wiley & Sons, Inc. (1975)

19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20. Hsu, Y.C., Lv, Z., Kira, Z.: Learning to cluster in order to transfer across domains and tasks. In: International Conference on Learning Representations (2018)
21. Hsu, Y.C., Lv, Z., Schlosser, J., Odom, P., Kira, Z.: Multi-class classification without multi-class labels. In: International Conference on Learning Representations (2018)
22. Jia, X., Han, K., Zhu, Y., Green, B.: Joint representation learning and novel category discovery on single-and multi-modal data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 610–619 (2021)
23. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
24. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
26. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
27. Liberty, E.: Simple and deterministic matrix sketching. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 581–588 (2013)
28. Liu, X., Pu, N., Zheng, H., Li, W., Sebe, N., Zhong, Z.: Generate, refine, and encode: Leveraging synthesized novel samples for on-the-fly fine-grained category discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1078–1087 (2025)
29. Liu, Y., Cai, Y., Jia, Q., Qiu, B., Wang, W., Pu, N.: Novel class discovery for ultra-fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17679–17688 (2024)
30. Liu, Y., Han, K.: Debgcd: Debiased learning with distribution guidance for generalized category discovery. arXiv preprint arXiv:2504.04804 (2025)
31. Liu, Y., He, Z., Han, K.: Hyperbolic category discovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9891–9900 (2025)
32. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
33. Ouldnoughi, R., Kuo, C.W., Kira, Z.: Clip-gcd: Simple language guided generalized category discovery. arXiv preprint arXiv:2305.10420 (2023)
34. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
35. Pu, N., Li, W., Ji, X., Qin, Y., Sebe, N., Zhong, Z.: Federated generalized category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28741–28750 (2024)
36. Rasmussen, C.E.: The infinite gaussian mixture model. In: Solla, S.A., Leen, T.K., Müller, K.R. (eds.) Advances in Neural Information Processing Systems 12. pp. 554–560. MIT Press, Cambridge, MA, USA (2000)
37. Ronen, M., Finder, S.E., Freifeld, O.: Deepdpm: Deep clustering with an unknown number of clusters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9861–9870 (2022)

38. Schaeffer, R., Liu, G.K.M., Du, Y., Linderman, S., Fiete, I.R.: Streaming inference for infinite non-stationary clustering. In: Conference on Lifelong Learning Agents. pp. 310–326. PMLR (2022)
39. Sethuraman, J.: A constructive definition of dirichlet priors. Statistica sinica pp. 639–650 (1994)
40. Su, Y., Zhou, R., Huang, S., Li, X., Wang, T., Wang, Z., Xu, M.: Multimodal generalized category discovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1634–1643 (2025)
41. Tan, K.C., Liu, Y., Ambrose, B., Tulig, M., Belongie, S.: The herbarium challenge 2019 dataset. arXiv preprint arXiv:1906.05372 (2019)
42. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
43. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7492–7501 (2022)
44. Vaze, S., Vedaldi, A., Zisserman, A.: No representation rules them all in category discovery. Advances in Neural Information Processing Systems **36**, 19962–19989 (2023)
45. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
46. Wang, C., Paisley, J., Blei, D.M.: Online variational inference for the hierarchical dirichlet process. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 752–760. JMLR Workshop and Conference Proceedings (2011)
47. Wang, H., Vaze, S., Han, K.: Hilo: A learning framework for generalized category discovery robust to domain shifts. arXiv preprint arXiv:2408.04591 (2024)
48. Wang, H., Vaze, S., Han, K.: Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. arXiv preprint arXiv:2403.13684 (2024)
49. Wang, Y., Chen, Z., Yang, D., Sun, Y., Qi, L.: Self-cooperation knowledge distillation for novel class discovery. arXiv preprint arXiv:2407.01930 (2024)
50. Welford, B.P.: Note on a method for calculating corrected sums of squares and products. Technometrics **4**(3), 419–420 (1962)
51. Wen, X., Zhao, B., Qi, X.: Parametric classification for generalized category discovery: A baseline study. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16590–16600 (2023)
52. Willes, J., Harrison, J., Harakeh, A., Finn, C., Pavone, M., Waslander, S.L.: Bayesian embeddings for few-shot open world recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(3), 1513–1529 (2022)
53. Xu, R., Zhang, C., Ren, H., He, X.: Dual-level adaptive self-labeling for novel class discovery in point cloud segmentation. arXiv preprint arXiv:2407.12489 (2024)
54. Zhang, C., Xu, R., He, X.: Novel class discovery for long-tailed recognition. Transactions on Machine Learning Research (2023)
55. Zhang, S., Khan, S., Shen, Z., Naseer, M., Chen, G., Khan, F.S.: Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3479–3488 (2023)

56. Zhao, B., Han, K.: Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. Advances in Neural Information Processing Systems **34** (2021)
57. Zheng, H., Pu, N., Li, W., Sebe, N., Zhong, Z.: Prototypical hash encoding for on-the-fly fine-grained category discovery. Advances in Neural Information Processing Systems **37**, 101428–101455 (2024)
58. Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., Sebe, N.: Neighborhood contrastive learning for novel class discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10867–10875 (2021)
59. Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., Sebe, N.: Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9462–9470 (2021)