

---

# Rethinking Out-of-Distribution Detection and Generalization with Collective Behavior Dynamics

---

Zhenbin Wang<sup>1</sup>, Lei Zhang<sup>1,2\*</sup>, Wei Huang<sup>1</sup>, Zhao Zhang<sup>1</sup>, Zizhou Wang<sup>3</sup>

<sup>1</sup>Machine Intelligence Laboratory, Sichuan University, Chengdu, China

<sup>2</sup>Tianfu Jincheng Laboratory, Chengdu, China

<sup>3</sup>Institute of High Performance Computing, A\*STAR, Singapore

wangzhenbin@stu.scu.edu.cn, leizhang@scu.edu.cn

## Abstract

Out-of-distribution (OOD) problems commonly occur when models process data with a distribution significantly deviates from the in-distribution (InD) training data. In this paper, we hypothesize that a *field* or *potential* more essential than features exists, and features are not the ultimate essence of the data but rather manifestations of them during training. With this in mind, we first treat the output of the feature extractor as charged particles and investigate their collective behavior dynamics within a self-consistent electric field. Then, to characterize the relationship between OOD problems and dynamical equations, we introduce the *basin of attraction* and prove that its boundary can be represented as the zero level set of a differentiable function of the potential, *i.e.*, the spatial integral of field. We further demonstrate that: *i)* InD and OOD inputs can be effectively separated based on whether they are steady state solutions for specific field conditions, enabling robust OOD detection and outperforming prior methods over three benchmarks. *ii)* the generalization capability correlates positively with the basin of attraction. By analyzing the dynamics of perturbations, we propose that the potential is well-characterized by a Fourier-domain form of the Poisson equation. Evaluated on six benchmark datasets, our method rivals the SoTA approaches for OOD generalization and can be seamlessly integrated with them to deliver additional gains. The code is available at <https://github.com/wongzbb/CBD>.

## 1 Introduction

Machine learning systems [15, 25, 70] have achieved significant success under the i.i.d. assumption, where training and test data are drawn from the same distribution, known as in-distribution (InD). However, this assumption is often violated in real-world scenarios [89, 95], as test data frequently originate from different distributions, referred to as out-of-distribution (OOD) data, which causes distribution shift and substantial performance degradation [20, 61, 21, 87]. Prior work on addressing OOD challenges primarily follows two directions: 1) OOD detection [19, 27, 49], which aims to identify test inputs that deviate from the training distribution to enhance model robustness and safety on unknown or anomalous data; 2) OOD generalization [8, 57, 1], which seeks to improve model performance on OOD data to enhance model transferability and adaptability. Despite extensive prior works and diverse methodologies, recent study [72, 36] in the community inspired by other disciplines is expected to provide new perspectives and insights into OOD problems.

Recent advancements witness the success of physics-inspired deep generative models, such as the Poisson flow model derived from electrostatics [92, 93] and the model based on the stochastic reversed heat equation [63]. The former abstracts data as charges and generates samples by evolving them

---

\*The corresponding author.

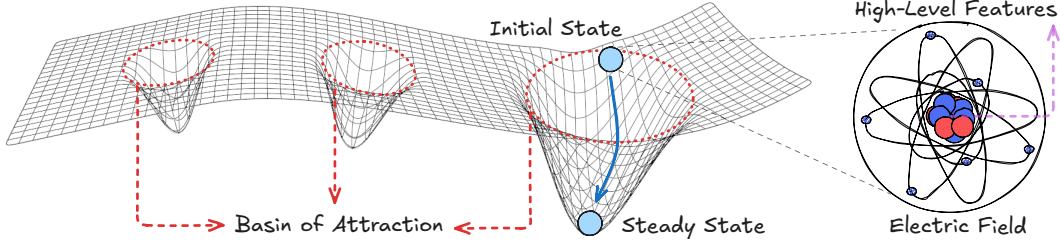


Figure 1: Conceptual illustration of our collective behavior dynamics insight for addressing OOD problems. High-level features are viewed as interacting particles within basin of attraction, transitioning from initial to steady states under the influence of a self-consistent electric field.

along electric field lines in an augmented space, while the latter interprets the solution of the forward heat equation with added constant noise as a variational approximation to a diffusion-based latent variable model. Overall, these approaches focus on learning the field of data using physical partial differential equations (PDE), rather than directly modeling the data distribution. These impressive achievements make us realize the fundamental nature of field or potential, with data are merely external manifestations or inherent properties of these deeper underlying concepts. This motivates us to model InD data using physical PDE. However, directly transferring prior work to OOD problems is infeasible, as generative models primarily focus on *the behavior of individual particles* within a field. In contrast, OOD problem center on determining field or potential boundaries, which requires analyzing *collective behavior* rather than individual behaviors. This discrepancy motivates us to first pose the following question:

*Can we find a concise, theoretically grounded yet tractable PDE to model the collective behavior of the high-level features and measure the boundary of the basin of attraction?*

In pursuit of this, we introduce the Vlasov-Poisson system [67, 18], a dynamic physics model that describes the evolution of collective particle behavior under self-consistent electric field. This system captures the *feedback loop* where particle positions induce changes in the field, which, in turn, influence particle positions, driving the system towards a **steady state**<sup>1</sup>. In this sense, we abstract high-level features from the feature extractor as charged particles, with their values corresponding to initial positions and assigned uniform initial velocities. Under the assumption that all InD features correspond to a steady-state solution of the Vlasov-Poisson system, their dynamics will naturally evolve toward that equilibrium. The region formed by such initial states is referred to as the **basin of attraction** (as depicted in Figure 1). For OOD detection, we provide two principles: *i*) Vlasov-Poisson systems modeled by collective particles with distinct initial positions but identical initial velocities, typically yield *unique* steady-state solutions (Theorem 2.1), and *ii*) minor perturbations outside the basin of attraction amplify during system evolution, *sharpening gradients* at the boundary (Corollary 2.3). Thus, the basin of attraction is ideal for distinguishing InD and OOD features. Since this process operates on pretrained high-level features, it is therefore a post-hoc detection method. On the other hand, for OOD generalization, considering potential latent defects outside the basin of attraction, we assume that network robustness improves as the basin’s range expands. So we next seek to address the following question:

*How can we enlarge basin of attraction’s boundary to enhance model generalization?*

To answer this, we decompose complex perturbations into superpositions of simple **plane waves**<sup>2</sup> and examine their collective behavior in the Vlasov-Poisson system by giving a dispersion relation that characterizes the stability of the wave modes. Through this lens, if all wave modes converge to a common steady state, the system ultimately stabilizes at that solution, regardless of the initial perturbation, as illustrated in Figure 2. Therefore, by requiring the potential in the random single-wave mode to satisfy the frequency form of Poisson equation, we expand the basin of attraction to include the perturbation points. We present Theorem 2.4 to support

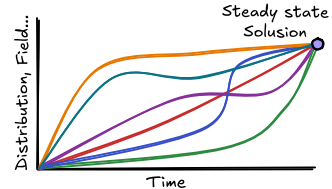


Figure 2: As all plane-wave modes reach steady state, cross-frequency interactions diminish, leading to equilibrium.

<sup>1</sup>A steady state refers to a condition where the system’s macroscopic properties remain constant over time, indicating a dynamic equilibrium between particle motion and the self-consistent field.

<sup>2</sup>A wave with uniform amplitude and phase across infinite, parallel wavefronts, propagating in fixed direction.

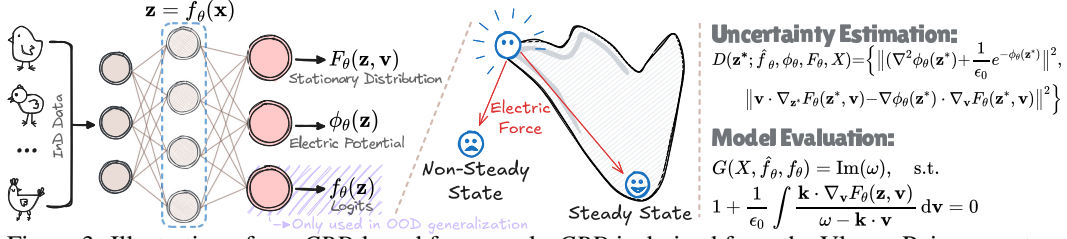


Figure 3: Illustration of our *CBD*-based framework. *CBD* is derived from the Vlasov-Poisson system and models the dynamics of high-level features under an electric field. The basin of attraction quantifies the region where features from the InD training set  $X$  spontaneously converge to a steady state. Based on *CBD*, we propose two methods targeting OOD detection and OOD generalization.

this design and discuss its impact on the model’s generalization ability, building on Theorem 2.5 and Corollary 2.6.

In summary, the contributions of this work are three-fold: *i)* Theoretically, we introduce an application of fields and potentials in OOD problems from two novel perspectives: the collective behavior dynamics of high-level features and complex perturbations under self-consistent electric fields, with supporting theories. *ii)* Methodologically, we propose physics-inspired **Collective Behavior Dynamics (CBD)** framework for OOD problems. It is simple, effective, and compatible with other methods. *iii)* Experimentally, our method outperforms or matches SoTA approaches across three OOD detection benchmarks and six OOD generalization benchmarks, and can be integrated with them for further improvement.

**Prior Arts.** We discuss related work and defer a concentrated account to Appendix A.

## 2 CBD: Collective Behavior Dynamics

Let  $X = \{(\mathbf{x}, \mathbf{y})\}$  denote the training set, consisting of i.i.d. samples drawn from the joint distribution  $\mathcal{P} = \mathcal{X} \times \mathcal{Y}$ , where  $\mathcal{X} \in \mathbb{R}^d$  represents the input space and  $\mathcal{Y} = \{1, 2, \dots, C\}$  is the output space. Consider a classification network parameterized by  $\theta$ , where the network maps inputs  $\mathbf{x} \in \mathcal{X}$  to a probability distribution over the output space  $\mathcal{Y}$ . We illustrate our *CBD*-based approaches in Figure 3. In the following, we first present the formulation of the **Vlasov-Poisson system** (Section 2.1), and then introduce the details of **basin of attraction** (Section 2.2), followed by its application to OOD detection. Finally, we formulate the **dispersion relation** (Section 2.3) and demonstrate how *CBD* can be utilized for OOD generalization.

### 2.1 Formulation of Vlasov-Poisson System

The performance of classification networks typically relies on high-level feature extracted from individual inputs. However, this *overlooks the deep physical constraints* that exist between high-level features across InD data. To model the behavior of high-level features  $\mathbf{z} = \hat{f}_\theta(\mathbf{x})$ , where  $\hat{f}_\theta(\cdot)$  denotes a feature extractor, we treat these features extracted from InD data as **charged particles** in the space  $\mathbb{R}^Z$ . Due to Coulomb interactions, these particles spontaneously move in space. In the absence of collisions and external magnetic fields, their dynamics can be effectively described by the Vlasov-Poisson system, which consists of two fundamental equations: the Vlasov equation and the Poisson equation. Specifically, the Vlasov equation is given by:

$$\frac{\partial F(\mathbf{z}, \mathbf{v}, t)}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{z}} F(\mathbf{z}, \mathbf{v}, t) + E(\mathbf{z}, t) \cdot \nabla_{\mathbf{v}} F(\mathbf{z}, \mathbf{v}, t) = 0. \quad (1)$$

Eq.(1) preserves the Hamiltonian structure and energy conservation of the system, where  $F(\mathbf{z}, \mathbf{v}, t) \in \mathbb{R}^Z \times \mathbb{R}^Z \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is the **particle distribution function** at time  $t$ , defining the particle density at position  $\mathbf{z}$  and velocity  $\mathbf{v}$ .  $E(\mathbf{z}, t) \in \mathbb{R}^Z$  corresponds to the **electric field** generated by particle distribution (assume all particles have the same charge).  $\nabla_{\mathbf{z}}$  and  $\nabla_{\mathbf{v}}$  represent the gradient with respect to position and velocity, respectively. Eq.(1) describes macroscopic space-time changes.

The Poisson equation characterizes the relationship between the electric field  $E(\mathbf{z}, t)$  and the particle distribution  $F(\mathbf{z}, \mathbf{v}, t)$ , and can be expressed as:

$$\nabla^2 \phi(\mathbf{z}, t) = -\frac{1}{\epsilon_0} \left( \int F(\mathbf{z}, \mathbf{v}, t) d\mathbf{v} - \rho_{\text{ion}}(\mathbf{z}) \right), \quad E(\mathbf{z}, t) = -\nabla \phi(\mathbf{z}, t). \quad (2)$$

Here  $\phi(\mathbf{z}, t): \mathbb{R}^Z \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  denotes the **electric potential**,  $\int F(\mathbf{z}, \mathbf{v}, t) d\mathbf{v}$  the **charge density** (i.e., the total charge at position  $\mathbf{z}$ ), and  $\epsilon_0 > 0$  the permittivity constant.  $\rho_{\text{ion}}(\mathbf{z})$  denotes Ion background

density, assumed to be a spatially uniform constant. It serves to neutralize the system and ensure potential solution. Eq.(1) and Eq.(2) together describe particle dynamics under the electric field.

**Theorem 2.1.** Consider  $N$  Vlasov-Poisson systems with steady state distributions  $F_i(\mathbf{z}, \mathbf{v}) = n_i(\mathbf{z}) \delta(\mathbf{v} - \mathbf{v}_0)$ , where each spatial density  $n_i(\mathbf{z}) \in L^1(\mathbb{R}^Z) \cap L^\infty(\mathbb{R}^Z)$  represents the spatial density for the  $i$ -th system, and  $L^1(\mathbb{R})$  and  $L^\infty(\mathbb{R})$  denote respectively the spaces of integrable and essentially bounded functions on  $\mathbb{R}$ . The Dirac delta distribution  $\delta(\mathbf{v} - \mathbf{v}_0)$  is centered at  $\mathbf{v} = \mathbf{v}_0$ , indicating a velocity distribution concentrated at that point. Assume that each  $n_i(\mathbf{z})$  satisfies  $\mathbf{v}_0 \cdot \nabla_{\mathbf{z}} n_i(\mathbf{z}) = 0$  and that the ion background density is given by  $\rho_{ion,i}(\mathbf{z}) = n_i(\mathbf{z})$ . If  $n_i(\mathbf{z}) \neq n_j(\mathbf{z})$  for  $i \neq j$ , then the solutions  $F_i$  are distinct.

We prove that in different multiple systems, where particles share identical mass and initial velocity  $\mathbf{v}_0$  but have distinct spatial density distributions, the steady state solutions (i.e.,  $\frac{\partial F}{\partial t} = 0$ ) are distinct. Using the provided Vlasov-Poisson equations, a neutralizing ion background, and a density gradient condition, we establish the result via distribution theory. The proof is deferred to the Appendix B.1.

## 2.2 Basin of Attraction and Application to OOD Detection

Building upon the concept of the Vlasov-Poisson system, we introduce the notion of a basin of attraction to characterize the system’s dynamical behavior under both InD and OOD inputs. Let  $F^*(\mathbf{z}, \mathbf{v})$  denote a **steady state solution** to the Vlasov–Poisson system. We define the basin of attraction  $\mathcal{B}(F^*)$  as the set of initial distributions  $F_0$  such that the time-moved solution asymptotically approaches  $F^*$ :

$$\mathcal{B}(F^*) = \{F_0 \in \mathcal{F} \mid \lim_{t \rightarrow \infty} F(t, \cdot, \cdot) = F^* \text{ in some topology}\}. \quad (3)$$

Here,  $\mathcal{F}$  denotes the admissible function space (e.g., Sobolev or Wasserstein space). This asymptotic behavior can be equivalently characterized by the steady state form of the Vlasov–Poisson system,

$$\begin{cases} \mathbf{v} \cdot \nabla_{\mathbf{z}} F^*(\mathbf{z}, \mathbf{v}) + E^*(\mathbf{z}) \cdot \nabla_{\mathbf{v}} F^*(\mathbf{z}, \mathbf{v}) = 0, \\ \nabla^2 \phi^*(\mathbf{z}) = -\frac{1}{\epsilon_0} \left( \int F^*(\mathbf{z}, \mathbf{v}) d\mathbf{v} - \rho_{ion}(\mathbf{z}) \right), \quad E^*(\mathbf{z}) = -\nabla \phi^*(\mathbf{z}). \end{cases} \quad (\text{time-independent}) \quad (4)$$

**Definition 2.2.** Under thermodynamic equilibrium, the steady-state electric potential  $\phi^*(\mathbf{z})$  satisfies the nonlinear Poisson–Boltzmann equation:

$$\nabla^2 \phi^*(\mathbf{z}) = -\frac{1}{\epsilon_0} (e^{-\phi^*(\mathbf{z})} - \rho_{ion}(\mathbf{z})). \quad (5)$$

Eq.(5) describes the equilibrium electrostatic potential resulting from a Boltzmann-distributed charge density balanced against a fixed ion background. The derivation is provided in Appendix B.2.

**Corollary 2.3.** Consider  $\phi^*(\mathbf{z})$  be the steady state electric potential satisfying the nonlinear Poisson–Boltzmann equation. Then the basin of attraction’s boundary  $\partial\mathcal{B}$  can be implicitly represented as the zero level set of a scalar functional over  $\phi^*(\mathbf{z})$ , i.e.,

$$\partial\mathcal{B} = \{\mathbf{z} \in \mathbb{R}^Z \mid \mathcal{G}(\phi^*(\mathbf{z})) = 0\}, \quad (6)$$

where  $\mathcal{G}$  is a differentiable function that depends on system configuration, such as energy thresholds, local potential curvature, or external field geometry. This representation enables differentiable modeling and learning of basin boundaries via neural function approximators.

Corollary 2.3 establishes that the basin of attraction  $\mathcal{B}$  associated with the steady state dynamics is fully characterized by the solution  $\phi^*(\mathbf{z})$  to the nonlinear Poisson–Boltzmann equation, through a differentiable level-set condition  $\mathcal{G}(\phi^*(\mathbf{z}))=0$ . It implies that the geometric and topological structure of  $\mathcal{B}$ , including its boundary  $\partial\mathcal{B}$ , is encoded in the electric potential  $\phi^*$ . We defer the formal to Appendix B.3. Given the challenges of directly solving high-dimensional nonlinear PDE of Eq.(4) and the need for framework compatibility, we propose incorporating two parallel MLP branches following the feature extractor:  $F_\theta$  for predicting the steady state distribution function  $F^*$  and  $\phi_\theta$  for estimating the steady-state potential  $\phi^*$ , both running in parallel with the linear classifier  $f_\theta$ . As shown in Figure 3. This architecture ensures CBD compatibility with most existing methods. Then we formulate a physics-informed loss, which enforces the consistency of the learned potential with the steady state field equation. *Because*

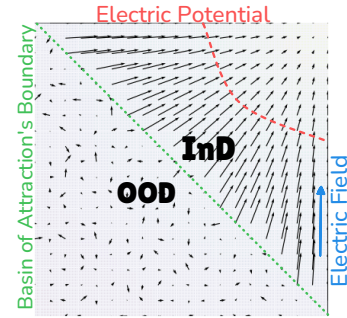


Figure 4: Electric field, potential, and basin of attraction’s boundary. Particles inside converge to steady state; those outside behave chaotically.



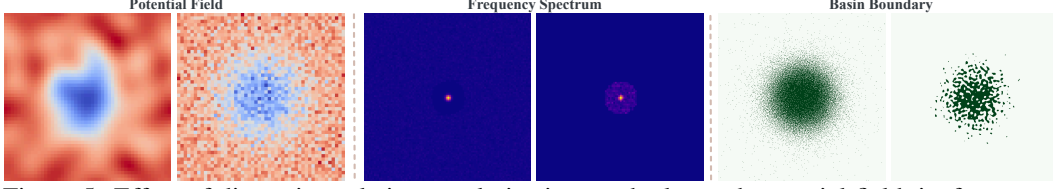


Figure 5: Effect of dispersion relation regularization on the learned potential field, its frequency spectrum, and the recovered basin of attraction. *Left column:* Predicted potential fields with (left) and without (right) dispersion relation loss. The regularized solution is smoother and more physically consistent. *Middle column:* Corresponding frequency spectrum. Dispersion relation regularization suppresses spurious high-frequency components, aligning the spectral structure with physical expectations. *Right column:* Estimated basin boundaries (contour of  $\phi_\theta(\mathbf{z})=\text{const}$ ). Dispersion regularization leads to a wider and smoother basin, indicating improved global consistency and robustness.

the Ion background density  $\rho_{\text{ion}}(\mathbf{z})$  is constant, it introduces only a uniform offset and can be safely omitted from derivation. Thus the resulting Poisson loss is

$$\mathcal{L}_{\text{Poisson}}(\mathbf{z}; \hat{f}_\theta, \phi_\theta) = \left\| \nabla^2 \phi_\theta(\mathbf{z}) + \frac{1}{\epsilon_0} e^{-\phi_\theta(\mathbf{z})} \right\|^2. \quad (7)$$

The steady state potential  $\phi^*$  encapsulates the *overall* geometric and topological structure of the InD data, while also satisfying the Vlasov equation. So we get the following Vlasov loss:

$$\mathcal{L}_{\text{Vlasov}}(\mathbf{z}; \hat{f}_\theta, \phi_\theta, F_\theta) = \left\| \mathbf{v} \cdot \nabla_{\mathbf{z}} F_\theta(\mathbf{z}, \mathbf{v}) - \nabla \phi_\theta(\mathbf{z}) \cdot \nabla_{\mathbf{v}} F_\theta(\mathbf{z}, \mathbf{v}) \right\|^2. \quad (8)$$

Then we leverage  $\mathcal{L}_{\text{Poisson}}$  and  $\mathcal{L}_{\text{Vlasov}}$  as a physically grounded uncertainty score to identify whether  $\mathbf{z}^* = \hat{f}(\mathbf{x}^*)$  lies inside the basin of attraction or corresponds to OOD behavior,

$$D(\mathbf{z}^*; \hat{f}_\theta, \phi_\theta, F_\theta, X) = \begin{cases} \text{InD} & \text{if } \mathcal{L}_{\text{Poisson}}(\mathbf{z}^*; \hat{f}_\theta, \phi_\theta) \leq \lambda_1 \text{ and } \mathcal{L}_{\text{Vlasov}}(\mathbf{z}^*; \hat{f}_\theta, \phi_\theta, F_\theta) \leq \lambda_2 \\ \text{OOD} & \text{otherwise} \end{cases}, \quad (9)$$

where  $\lambda_1 > 0$  and  $\lambda_2 > 0$  is thresholds and  $\mathbf{x}^*$  is test data. This approach not only offers physical interpretability, with the score directly quantifying deviation from the governing PDE. But also geometric alignment with basin structure, as Corollary 2.3 demonstrates that  $\mathcal{G}(\phi^*(\mathbf{z}))$  exhibits sharp gradient near the boundary  $\partial\mathcal{B}$ , enabling robust OOD detection without statistical estimators.

### 2.3 Dispersion Relation and Application to OOD Generalization

OOD data's high-level features may fall outside the basin of attraction of InD data, preventing spontaneous transition from initial to steady states under given electric fields. From this perspective, we hypothesize that **network robustness is proportional to the range of the basin of attraction**. As the basin of attraction contracts, larger activity spaces emerge, increasing the probability of potential defects. As illustrated in the second image of the right column in Figure 5, without additional regularization, the learned boundary of the basin of attraction for exhibits fragmentation, demonstrating not only sensitivity to input data but also a diminished area (green area). Enlarge this basin implies that the system can accommodate larger initial deviations, without diverging or converging to alternative equilibria, thereby improving the robustness of the steady state.

We define the set of initial perturbations  $(f_0, \psi_0) = (F - F^*, \phi - \phi^*)$ . To investigate the collective behavior dynamics of perturbations, we employ **plane waves** for decomposition. We define the plane wave as  $\psi(\mathbf{z}, t) = \tilde{\psi} e^{i(\mathbf{k} \cdot \mathbf{z} - \omega t)}$  and  $f(\mathbf{z}, \mathbf{v}, t) = \tilde{f}(\mathbf{v}) e^{i(\mathbf{k} \cdot \mathbf{z} - \omega t)}$ , where  $\mathbf{k} \in \mathbb{R}^Z$  denotes the wave vector specifying the spatial frequency and direction (we treat  $\mathbf{k}$  as a *random vector hyperparameter* in our experiments), and  $\omega = \omega_r + i\omega_i$  is the complex frequency. The scalar  $\tilde{\psi}$  and  $\tilde{f}(\mathbf{v})$  represents the mode amplitude. This representation enables **any perturbation to be expressed as a superposition of decoupled modes**, and in systems exhibiting translational invariance, plane waves naturally arise as eigenfunctions, transforming PDE into algebraic forms. Substituting  $\psi$  and  $f$  into Eq.(4), we get the *perturbation collective behavior dynamics equations*:

$$\begin{cases} \partial_t f(\mathbf{z}, t) + \mathbf{v} \cdot \nabla_{\mathbf{z}} f(\mathbf{z}, t) + E^*(\mathbf{z}) \cdot \nabla_{\mathbf{v}} f(\mathbf{z}, t) = (\nabla_{\mathbf{v}} F^*(\mathbf{z}, \mathbf{v})) \cdot (\nabla_{\mathbf{z}} \psi(\mathbf{z}, t)), \\ \nabla^2 \psi(\mathbf{z}, t) = -\frac{1}{\epsilon_0} \int f(\mathbf{z}, t) d\mathbf{v}. \end{cases} \quad (10)$$

Substituting the plane wave into Eq.(10), yielding

$$\begin{cases} -i\omega \tilde{f}(\mathbf{v}) + i\mathbf{k} \cdot \mathbf{v} \tilde{f}(\mathbf{v}) + (-\nabla \phi^*(\mathbf{z}) \cdot \nabla_{\mathbf{v}} \tilde{f}(\mathbf{v})) = \tilde{\psi} (\nabla_{\mathbf{v}} F^*(\mathbf{z}, \mathbf{v})) \cdot i\mathbf{k}, \\ -\|\mathbf{k}\|^2 \tilde{\psi} = -\frac{1}{\epsilon_0} \int \tilde{f}(\mathbf{v}) d\mathbf{v}. \end{cases} \quad (11)$$

Assuming a homogeneous steady state for simplicity (i.e.,  $\phi^*(\mathbf{z})$  is constant and  $F^*(\mathbf{z}, \mathbf{v}) = F^*(\mathbf{v})$ ), the electric field  $E^*(\mathbf{z}) = 0$ , and Eq.(11) simplifies to:

$$-i\omega \tilde{f}(\mathbf{v}) + i\mathbf{k} \cdot \mathbf{v} \tilde{f}(\mathbf{v}) = \tilde{\psi}(\nabla_{\mathbf{v}} F^*(\mathbf{z}, \mathbf{v})) \cdot i\mathbf{k}. \quad (12)$$

Solving for  $\tilde{f}(\mathbf{v})$ , we obtain:

$$\tilde{f}(\mathbf{v}) = \frac{\tilde{\psi}(\nabla_{\mathbf{v}} F^*(\mathbf{v})) \cdot \mathbf{k}}{\omega - \mathbf{k} \cdot \mathbf{v}}. \quad (13)$$

Integrating over velocity space and substituting into the Poisson equation, we get:

$$D(\omega, \mathbf{k}) = 1 + \frac{1}{\epsilon_0} \int \frac{\mathbf{k} \cdot \nabla_{\mathbf{v}} F^*(\mathbf{v})}{\omega - \mathbf{k} \cdot \mathbf{v}} d\mathbf{v} = 0. \quad (14)$$

The solutions of  $D(\omega, \mathbf{k})$  is  $\omega(\mathbf{k}) = \omega_r + i\omega_i$ , which determine the stability of each mode. To assess stability, we define an energy-like quantity as the squared magnitude of the perturbation,  $|\psi(\mathbf{z}, t)|^2$  (or alternatively  $|f(\mathbf{z}, \mathbf{v}, t)|^2$ , with analogous derivation steps following). Substituting  $\omega(\mathbf{k}) = \omega_r + i\omega_i$ , this becomes  $|\psi(\mathbf{z}, t)|^2 = |\tilde{\psi}|^2 e^{2\omega_i t}$ , since the exponential term  $e^{i(\mathbf{k} \cdot \mathbf{z} - (\omega_r + i\omega_i)t)} = e^{i\mathbf{k} \cdot \mathbf{z}} e^{-i\omega_r t} e^{\omega_i t}$ , and the magnitudes of the complex phases  $e^{i\mathbf{k} \cdot \mathbf{z}}$  and  $e^{-i\omega_r t}$  are unity. Thus, *when  $\omega_i > 0$ , the factor  $e^{2\omega_i t}$  grows exponentially, causing the energy-like quantity to increase, potentially destabilizing the system and contracting the basin of attraction. Conversely, when  $\omega_i < 0$ ,  $e^{2\omega_i t}$  decays exponentially, reducing the energy-like quantity and facilitating recovery to the steady state.* In the nonlinear regime, perturbations decompose into multiple modes, and decaying modes ( $\omega_i < 0$ ) can mitigate the growth of unstable modes ( $\omega_i > 0$ ) through modal interactions, such as energy redistribution. Therefore, the imaginary part of  $\omega$  can be used to evaluate the generalization ability of the model.

To enlarge the basin of attraction, we aim to suppress perturbations associated with positive growth rates, i.e., those satisfying  $\omega_i > 0$ , as they contribute to instability and hinder recovery to equilibrium. To this end, we introduce the following dispersion relation loss that regularizes the consistency between the electric potential and the charge distribution in the *frequency domain*

$$\begin{aligned} \mathcal{L}_{\text{disp}}(\mathbf{z}; \hat{f}_\theta, \phi_\theta) &= \sum_{\mathbf{k}} \|\|\mathbf{k}\|^2 \phi'_{\mathbf{k}}(\mathbf{z}) + \rho'_{\mathbf{k}}(\mathbf{z})\|^2, \\ \text{s.t. } \phi'_{\mathbf{k}}(\mathbf{z}) &= \phi_\theta(\mathbf{z}) e^{-i\mathbf{k} \cdot \mathbf{z}}, \quad \rho'_{\mathbf{k}}(\mathbf{z}) = e^{-\phi_\theta(\mathbf{z})} e^{-i\mathbf{k} \cdot \mathbf{z}}. \end{aligned} \quad (15)$$

This loss arises from the **Fourier-domain** formulation of the Poisson equation. Minimizing Eq.(15) therefore enforces the constraint across sampled wavevectors  $\mathbf{k}$  (see Figure 2). From a dynamical perspective, satisfying this balance prevents the system from developing unstable frequency components, *which would otherwise drive the  $D(\omega, \mathbf{k}) = 0$  toward solutions with  $\omega_i > 0$ , such imbalances act as implicit forcing terms that require growing temporal responses.* In contrast, minimizing the dispersion relation loss encourages configurations where  $\omega_i < 0$ , ensuring exponential decay of perturbation energy and improved return to steady state. This regularization thus enhances stability and expands the basin of attraction. Figure 5 presents a comparative analysis of predicted potential fields, frequency spectra, and estimated basin boundaries before (right) and after (left) incorporating the dispersion relation loss, demonstrating its superior efficacy in enlarging the basin of attraction. According to Eq.(15), the underlying theoretical rationale can be formulated as follows.

**Theorem 2.4.** Define the approximate basin of attractor basin as the residual sublevel set

$$\mathcal{B}_\theta := \{\mathbf{z} \in \Omega \mid (\nabla^2 \phi_\theta(\mathbf{z}) + \frac{1}{\epsilon_0} e^{-\phi_\theta(\mathbf{z})})^2 < \lambda\}. \quad (16)$$

Then, under mild regularity assumptions, increasing the spectral regularization strength  $\lambda_{\text{disp}} > 0$  leads to a strictly larger basin measure:

$$\text{meas}(\mathcal{B}_\theta^{(\lambda_{\text{disp}} > 0)}) \geq \text{meas}(\mathcal{B}_\theta^{(\lambda_{\text{disp}} = 0)}). \quad (17)$$

**Theorem 2.5.** Let  $\phi_\theta, F_\theta \in H^s(\mathbb{R}^Z)$ ,  $H^s(\mathbb{R}^Z \times \mathbb{R}^Z)$ , respectively, with  $s > Z/2$ , approximate the true solutions  $\phi^*, F^*$  of the Vlasov-Poisson system. Training data  $(\mathbf{z}, \mathbf{v}) \sim \mathcal{D}$  are i.i.d. from  $\Omega \times \mathbb{R}^Z$ . The generalization error is:  $\mathcal{E}_{\text{gen}}(\theta) = \mathbb{E}_{\mathbf{z}, \mathbf{v} \sim \mathcal{D}} [\|\phi_\theta(\mathbf{z}) - \phi^*(\mathbf{z})\|^2 + \|F_\theta(\mathbf{z}, \mathbf{v}) - F^*(\mathbf{z}, \mathbf{v})\|^2]$ . The inclusion of  $\mathcal{L}_{\text{disp}}$  reduces  $\mathcal{E}_{\text{gen}}(\theta)$  by constraining the Sobolev norm of  $\phi_\theta$ , satisfying:

$$\mathbb{E}_{\mathbf{z}, \mathbf{k}}[\mathcal{L}_{\text{disp}}] \geq c \|\phi_\theta\|_{H^2}^2 - C, \quad (18)$$

for constants  $c, C > 0$ , thereby reducing the model complexity term in the generalization error bound.

Table 1: OOD detection performance on CIFAR-10 and CIFAR-100 benchmarks, with the top three results highlighted. Complete results for all baseline methods are provided in Appendix E.

| Method                     | MNIST       |            | SVHN        |            | Textures    |            | Places365  |            | Average    |            |
|----------------------------|-------------|------------|-------------|------------|-------------|------------|------------|------------|------------|------------|
|                            | FPR95↓      | AUROC↑     | FPR95↓      | AUROC↑     | FPR95↓      | AUROC↑     | FPR95↓     | AUROC↑     | FPR95↓     | AUROC↑     |
| <b>CIFAR-10 Benchmark</b>  |             |            |             |            |             |            |            |            |            |            |
| OpenMax                    | 23.33±4.67  | 90.50±0.44 | 25.40±1.47  | 89.77±0.45 | 31.50±4.05  | 89.58±0.60 | 38.52±2.27 | 88.63±0.28 | 29.69±1.21 | 89.62±0.19 |
| ODIN                       | 23.83±12.34 | 95.24±1.96 | 68.61±0.52  | 84.58±0.77 | 67.70±11.06 | 86.94±2.26 | 70.36±6.96 | 85.07±1.24 | 57.62±4.24 | 87.96±0.61 |
| MDSEns                     | 1.30±0.51   | 99.17±0.41 | 74.34±1.04  | 66.56±0.58 | 76.07±0.17  | 77.40±0.28 | 94.16±0.33 | 52.47±0.15 | 61.47±0.48 | 73.90±0.27 |
| RMDS                       | 21.49±2.32  | 93.22±0.80 | 23.46±1.48  | 91.84±0.26 | 25.25±0.53  | 92.23±0.23 | 31.20±0.28 | 91.51±0.11 | 25.35±0.73 | 92.20±0.21 |
| Gram                       | 70.30±8.96  | 72.64±2.34 | 33.91±17.35 | 91.52±4.45 | 94.64±2.71  | 62.34±8.27 | 90.49±1.93 | 60.44±3.41 | 72.34±6.73 | 71.73±3.20 |
| ReAct                      | 33.77±18.00 | 92.81±3.03 | 50.23±15.98 | 89.12±3.19 | 51.42±11.42 | 89.38±1.49 | 44.20±3.35 | 90.35±0.78 | 44.90±8.37 | 90.42±1.41 |
| VIM                        | 18.36±1.42  | 94.76±0.38 | 19.29±0.41  | 94.50±0.48 | 21.14±1.83  | 95.15±0.34 | 41.43±2.17 | 89.49±0.39 | 25.05±0.52 | 93.48±0.24 |
| KNN                        | 20.05±1.36  | 94.26±0.38 | 22.60±1.26  | 92.67±0.30 | 24.06±0.55  | 93.16±0.24 | 30.38±0.63 | 91.77±0.23 | 24.27±0.40 | 92.96±0.14 |
| ASH                        | 70.00±10.56 | 83.16±4.66 | 83.64±6.48  | 73.46±6.41 | 84.59±1.74  | 77.45±2.39 | 77.89±7.28 | 79.89±3.69 | 79.03±4.22 | 78.49±2.58 |
| SHE                        | 42.22±20.59 | 90.43±4.76 | 62.74±4.01  | 86.38±1.32 | 84.60±5.30  | 81.57±1.21 | 76.36±5.32 | 82.89±1.22 | 66.48±5.98 | 85.32±1.43 |
| GEN                        | 23.00±7.75  | 93.83±2.14 | 28.14±2.59  | 91.97±0.66 | 40.74±6.61  | 90.14±0.76 | 47.03±3.22 | 89.46±0.65 | 34.73±1.58 | 91.35±0.69 |
| NAC-UE                     | 15.14±2.60  | 94.86±1.36 | 14.33±1.24  | 96.05±0.47 | 17.03±0.59  | 95.64±0.44 | 26.73±0.80 | 91.85±0.28 | 18.31±0.92 | 94.60±0.50 |
| CBD-De                     | 15.20±0.59  | 97.48±1.29 | 14.22±1.50  | 97.85±0.68 | 18.44±1.82  | 97.03±0.95 | 22.75±2.21 | 93.18±0.73 | 17.65±1.03 | 96.39±0.81 |
| <b>CIFAR-100 Benchmark</b> |             |            |             |            |             |            |            |            |            |            |
| OpenMax                    | 53.82±4.74  | 76.01±1.39 | 53.20±1.78  | 82.07±1.53 | 56.12±1.91  | 80.56±0.09 | 54.85±1.42 | 79.29±0.40 | 54.50±0.68 | 79.48±0.41 |
| ODIN                       | 45.94±3.29  | 83.79±1.31 | 67.41±3.88  | 74.54±0.76 | 62.37±2.96  | 79.33±1.08 | 59.71±0.92 | 79.45±0.26 | 58.86±0.79 | 79.28±0.21 |
| MDSEns                     | 2.83±0.86   | 98.21±0.78 | 82.57±2.58  | 53.76±1.63 | 84.94±0.83  | 69.75±1.14 | 96.61±0.17 | 42.27±0.73 | 66.74±1.04 | 66.00±0.69 |
| RMDS                       | 52.05±6.28  | 79.74±2.49 | 51.65±3.68  | 84.89±1.10 | 53.99±1.06  | 83.65±0.51 | 53.57±0.43 | 83.40±0.46 | 52.81±0.63 | 82.92±0.42 |
| Gram                       | 53.53±7.45  | 80.71±4.15 | 20.06±1.96  | 95.55±0.60 | 89.51±2.54  | 70.79±1.32 | 94.67±0.60 | 46.38±1.21 | 64.44±2.37 | 73.36±1.08 |
| ReAct                      | 56.04±5.66  | 78.37±1.59 | 50.41±2.02  | 83.01±0.97 | 55.04±0.82  | 80.15±0.46 | 55.30±0.41 | 80.03±0.11 | 54.20±1.56 | 80.39±0.49 |
| VIM                        | 48.32±1.07  | 81.89±1.02 | 46.22±5.46  | 83.14±3.71 | 46.86±2.29  | 85.91±0.78 | 61.57±0.77 | 75.85±0.37 | 50.74±1.00 | 81.70±0.62 |
| KNN                        | 48.58±4.67  | 82.36±1.52 | 51.75±3.12  | 84.15±1.09 | 53.56±2.32  | 83.66±0.83 | 60.70±1.03 | 79.43±0.47 | 53.65±0.28 | 82.40±0.17 |
| ASH                        | 66.58±3.88  | 77.23±0.46 | 46.00±2.67  | 85.60±1.40 | 61.27±2.74  | 80.72±0.70 | 62.95±0.99 | 78.76±0.16 | 59.20±2.46 | 80.58±0.66 |
| SHE                        | 58.78±2.70  | 76.76±1.07 | 59.15±7.61  | 80.97±3.98 | 73.29±3.22  | 73.64±1.28 | 65.24±0.98 | 76.30±0.51 | 64.12±2.70 | 76.92±1.16 |
| GEN                        | 53.92±5.71  | 78.29±2.05 | 55.45±2.76  | 81.41±1.50 | 61.23±1.40  | 78.74±0.81 | 56.25±1.01 | 80.28±0.27 | 56.71±1.59 | 79.68±0.75 |
| NAC-UE                     | 21.97±6.62  | 93.15±1.63 | 24.39±4.66  | 92.40±1.26 | 40.65±1.94  | 89.32±0.55 | 73.57±1.16 | 73.05±0.68 | 40.14±1.86 | 86.98±0.37 |
| CBD-De                     | 27.93±7.73  | 96.29±1.81 | 18.72±6.12  | 95.80±2.05 | 37.34±2.30  | 90.02±1.52 | 55.23±0.88 | 81.50±1.16 | 34.81±1.60 | 90.90±0.68 |

**Corollary 2.6.** Under the same assumptions with Theorem 2.5, let  $\mathcal{H} = \{(\phi_\theta, F_\theta) : \theta \in \Theta\}$  be the hypothesis class. The generalization bound is:

$$\mathcal{E}_{\text{gen}}(\theta) \leq \mathcal{E}_{\text{train}}(\theta) + 2\text{Rad}(\mathcal{H}) + O(\sqrt{\log(1/\delta)/n}), \quad (19)$$

where  $\text{Rad}(\mathcal{H})$  is the Rademacher complexity, and  $n$  is the sample size. The inclusion of  $\mathcal{L}_{\text{disp}}$  reduces  $\text{Rad}(\mathcal{H})$  by restricting  $\phi_\theta$  to a subset  $\mathcal{H}_\phi \subset H^2$ , tightening the generalization bound.

Theorem 2.4 proves that incorporating the loss  $\mathcal{L}_{\text{disp}}$  enlarges the basin of attraction. Theorem 2.5 further establishes that this loss term decreases the generalization error, and Corollary 2.6 presents the resulting tightening generalization bound. Proofs are deferred to Appendix B.4, B.5 and B.6. For OOD detection, the loss is defined as  $\mathcal{L}_{\text{De}} \triangleq \mathcal{L}_{\text{Vlasov}} + \mathcal{L}_{\text{Poisson}}$ . For OOD generalization, we enhance the training by incorporating  $\mathcal{L}_{\text{disp}}$  along with the standard classification loss [81], yielding the overall loss  $\mathcal{L}_{\text{Gen}} \triangleq \mathcal{L}_{\text{Classify}} + \alpha(\mathcal{L}_{\text{Vlasov}} + \mathcal{L}_{\text{Poisson}}) + \beta\mathcal{L}_{\text{disp}}$ , which ensures robust performance on the primary task. For detailed pseudocode, refer to Appendix C.

### 3 Experiments

**Steup.** For OOD detection, following the latest OpenOOD [94], we evaluated our CBD-De on three InD datasets: CIFAR-10, CIFAR-100 [38], and ImageNet-1k [12]. For CIFAR-10 and CIFAR-100, we employed ResNet-18 and assessed performance against four OOD datasets: MNIST [13], SVHN [59], Textures [10], and Places365 [101]. For ImageNet-1k, we utilized pre-trained ResNet-50 and ViT-b16 models, evaluating against three OOD datasets: iNaturalist [79], Textures, and OpenImage-o [84]. We compared our approach with 22 SoTA OOD detection methods, with comprehensive results provided in Appendix E. For OOD generalization, we adhere to the Domainbed [23] in our experimental setup and use BCE as classification loss, evaluating our CBD-Gen on five datasets: VLCS [16], PACS [45], OfficeHome [82], TerraInc [5], and DomainNet [60]. For each dataset, we report the leave-one-out test accuracy, averaged across splits where one domain serves as the test set and the remaining domains as the training set. Our method is compared against 18 SoTA baselines, with comprehensive results available in Appendix F.

**Metrics.** For OOD detection, we employed two threshold-free evaluation metrics, consistent with established conventions: *i)* FPR95, the false-positive-rate of OOD data when the true-positive-rate of ID data is at 95%, and *ii)* AUROC, the area under the receiver operating characteristic curve. For OOD generalization, we adhere to the standard practice of using accuracy as the evaluation metric.

**Implementation Details.** In the interest of brevity, we defer details such as training procedures, hyperparameter configurations, and baseline methodologies to Appendix D.

Table 2: Performance of OOD detection (AUROC) on ImageNet-1k, with full results in Appendix E.

| Dataset     | Backbone  | OpenMax      | MDS   | RMDS  | ReAct | VIM          | KNN   | ASH   | SHE   | GEN   | NAC-UE       | <i>CBD-De</i> |
|-------------|-----------|--------------|-------|-------|-------|--------------|-------|-------|-------|-------|--------------|---------------|
| iNaturalist | ResNet-50 | 92.05        | 63.67 | 87.24 | 96.34 | 89.56        | 86.41 | 97.07 | 92.65 | 92.44 | 96.52        | 97.06         |
|             | Vit-b16   | 94.93        | 96.01 | 96.10 | 86.11 | 95.72        | 91.46 | 50.62 | 93.57 | 93.54 | 93.72        | 97.26         |
|             | Average   | <u>93.49</u> | 79.84 | 91.67 | 91.23 | 92.64        | 88.94 | 73.85 | 93.11 | 92.99 | <u>95.12</u> | <b>97.16</b>  |
| OpenImage-O | ResNet-50 | 87.62        | 69.27 | 85.84 | 91.87 | 90.50        | 87.04 | 93.26 | 86.52 | 89.26 | 91.45        | 92.65         |
|             | Vit-b16   | 87.36        | 92.38 | 92.32 | 84.29 | 92.18        | 89.86 | 55.51 | 91.04 | 90.27 | 91.58        | 91.89         |
|             | Average   | 87.49        | 80.83 | 89.08 | 88.08 | 91.34        | 88.45 | 74.39 | 88.78 | 89.77 | <u>91.52</u> | <b>92.27</b>  |
| Textures    | ResNet-50 | 88.10        | 89.80 | 86.08 | 92.79 | 97.97        | 97.09 | 96.90 | 93.60 | 87.59 | 97.90        | 98.31         |
|             | Vit-b16   | 85.52        | 89.41 | 89.38 | 86.66 | 90.61        | 91.12 | 48.53 | 92.65 | 90.23 | 94.17        | 94.30         |
|             | Average   | 86.81        | 89.61 | 87.73 | 89.73 | <u>94.29</u> | 94.11 | 72.72 | 93.13 | 88.91 | <u>96.04</u> | <b>96.31</b>  |

Table 3: OOD generalization performance on DomainBed, with complete results in Appendix F.

| Algorithm    | Mixstyle       | MMD            | ARM            | MTL            | MLDG           | Mixup          | CORAL                 | SAM            | SAGM                  | GGA                   | <i>CBD-Gen</i>        |
|--------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------------------|----------------|-----------------------|-----------------------|-----------------------|
| PACS         | 85.2 $\pm$ 0.3 | 84.7 $\pm$ 0.5 | 85.1 $\pm$ 0.4 | 84.6 $\pm$ 0.5 | 84.9 $\pm$ 1.0 | 84.6 $\pm$ 0.6 | <u>86.2</u> $\pm$ 0.3 | 85.8 $\pm$ 0.2 | 86.6 $\pm$ 0.2        | 87.3 $\pm$ 0.4        | <b>87.7</b> $\pm$ 0.6 |
| VLCS         | 77.9 $\pm$ 0.5 | 77.5 $\pm$ 0.9 | 77.6 $\pm$ 0.3 | 77.2 $\pm$ 0.4 | 77.2 $\pm$ 0.4 | 77.4 $\pm$ 0.6 | 78.8 $\pm$ 0.6        | 79.4 $\pm$ 0.1 | <u>80.0</u> $\pm$ 0.3 | 79.9 $\pm$ 0.4        | <b>80.5</b> $\pm$ 0.5 |
| OfficeHome   | 60.4 $\pm$ 0.3 | 66.3 $\pm$ 0.1 | 64.8 $\pm$ 0.3 | 66.4 $\pm$ 0.5 | 66.8 $\pm$ 0.6 | 68.1 $\pm$ 0.3 | 68.7 $\pm$ 0.3        | 69.6 $\pm$ 0.1 | <u>70.1</u> $\pm$ 0.2 | <u>70.1</u> $\pm$ 0.2 | <b>71.4</b> $\pm$ 0.3 |
| TerraInc     | 44.0 $\pm$ 0.7 | 42.2 $\pm$ 1.6 | 45.5 $\pm$ 0.3 | 45.6 $\pm$ 1.2 | 47.7 $\pm$ 0.2 | 47.9 $\pm$ 0.8 | 48.6 $\pm$ 1.0        | 43.3 $\pm$ 0.7 | <u>48.8</u> $\pm$ 0.9 | <b>50.6</b> $\pm$ 0.1 | <u>50.5</u> $\pm$ 0.1 |
| <b>Avg.</b>  | 66.9           | 67.7           | 68.3           | 68.5           | 69.2           | 69.5           | <u>70.3</u>           | 69.3           | 71.4                  | <u>71.7</u>           | <b>72.5</b>           |
| DomainNet    | 34.0 $\pm$ 0.1 | 23.4 $\pm$ 9.5 | 35.5 $\pm$ 0.2 | 40.6 $\pm$ 0.1 | 41.2 $\pm$ 0.1 | 40.3 $\pm$ 0.1 | 41.5 $\pm$ 0.1        | 44.3 $\pm$ 0.2 | <u>45.0</u> $\pm$ 0.2 | <u>45.2</u> $\pm$ 0.2 | <b>45.9</b> $\pm$ 0.4 |
| <b>Total</b> | 60.3           | 58.8           | 61.7           | 62.9           | 63.6           | 63.4           | 64.5                  | 64.5           | <u>66.1</u>           | <u>66.3</u>           | <b>67.2</b>           |

**Quantitative results on OOD detection.** Tables 1 and 2 present the results of our *CBD-De* on the CIFAR and ImageNet datasets, compared against 22 SoTA methods. Confidence intervals in the tables reflect mean  $\pm$  standard error over three independent test runs. Our *CBD-De* consistently outperforms all SoTA methods in average performance, achieving record-breaking results across three benchmarks. Specifically, on CIFAR-10 and CIFAR-100, *CBD-De* reduces the FPR95 by 0.66% and 5.33%, respectively, and improves the AUROC by 1.79% and 3.92% compared to the most competitive baseline [52]. On the large-scale ImageNet-1k dataset, *CBD-De* enhances AUROC scores across various backbones and OOD datasets, with average AUROC improvements of 2.04%, 0.75%, and 0.27% over the strongest competitor.

**Quantitative results on OOD generalization.** Table 3 reports the average OOD accuracies of SoTA generalization methods across five benchmarks. For fair comparison, all methods are trained on the same amount of data. While additional data augmentations could further improve performance, they are not applied to isolate the effect of the core methods. The average performance of *CBD-Gen* across all evaluated datasets surpasses that of the most competitive benchmark by 0.9%. Specifically, on the large-scale DomainNet dataset, comprising 586,575 images across six domains, *CBD-Gen* achieves a 0.7% improvement over the leading method. Furthermore, *CBD-Gen* significantly outperforms the current SoTA on the OfficeHome dataset. We attribute this to its use of electric field modeling, which enhances the network’s ability to capture discriminative features from limited data. In particular, the electric potential exhibits sharp gradient transitions near the basin of attraction’s boundary, leading to larger inter-class margins. This effect is beneficial for OfficeHome, where each class contains only  $\sim 240$  samples on average, and substantially fewer than the 1,400+ samples per class available in other datasets, highlighting *CBD-Gen*’s strength in low-data regimes.

**Compatibility results.** We further investigated the potential of *CBD-Gen* by integrating it with two advanced algorithms: SAM [17] and SAGM [86]. During training, we incorporated two MLP branches after the feature extractor and incorporate additional  $\mathcal{L}_{\text{Poisson}}$ ,  $\mathcal{L}_{\text{Vlasov}}$  and  $\mathcal{L}_{\text{disp}}$  losses into the existing algorithmic frameworks. These branches are removed during testing. Results are shown in Table 4. Notably, *CBD-Gen* maintains consistent and superior performance across multiple datasets when combined with other methods, further validating its strong generalizability and compatibility. This demonstrates that *CBD-Gen* is not only effective within traditional training paradigms but can also function as an enhancement module that integrates seamlessly with various advanced methods, exhibiting extensive adaptation capabilities.

**Text OOD Detection.** To demonstrate the modality-agnostic nature of our post-hoc OOD detection method *CBD-De*, we further conduct experiments on text classification benchmarks. Following the evaluation protocol established in [4], we evaluate our method on standard text OOD detection datasets. Specifically, we adopt the widely used SST-2 [71], IMDB [54], and Yelp [99] datasets,

Table 4: Evaluation of *CBD-Gen*’s compatibility with advanced OOD generalization methods.

| Methods                  | PACS   | VLCS   | OfficeHome | TerraInc | <b>Avg.</b> |
|--------------------------|--------|--------|------------|----------|-------------|
| baseline                 | 85.8   | 79.4   | 69.6       | 43.3     | 69.3        |
| SAM with <i>CBD-Gen</i>  | 88.4   | 80.7   | 71.7       | 51.0     | 73.0        |
| $\Delta$                 | (+2.6) | (+1.3) | (+2.1)     | (+7.7)   | (+3.7)      |
| baseline                 | 86.6   | 80.0   | 70.1       | 48.8     | 71.4        |
| SAGM with <i>CBD-Gen</i> | 88.2   | 80.7   | 71.5       | 50.8     | 72.8        |
| $\Delta$                 | (+1.6) | (+0.7) | (+1.4)     | (+2.0)   | (+1.4)      |



Table 4: Performance of OOD detection (AUROC) on text data.

| ID    | OOD  | MSP            | Energy         | GradNorm       | KLM            | ReAct          | DICE            | KNN            | ViM            | <i>CBD-De</i>  |
|-------|------|----------------|----------------|----------------|----------------|----------------|-----------------|----------------|----------------|----------------|
| SST-2 | IMDB | 83.2 $\pm$ 1.4 | 82.7 $\pm$ 2.2 | 70.3 $\pm$ 2.3 | 55.0 $\pm$ 2.7 | 83.3 $\pm$ 2.4 | 34.5 $\pm$ 10.7 | 87.2 $\pm$ 1.7 | 83.9 $\pm$ 3.3 | 89.0 $\pm$ 2.2 |
| SST-2 | Yelp | 75.7 $\pm$ 2.2 | 75.0 $\pm$ 3.1 | 61.3 $\pm$ 2.7 | 51.3 $\pm$ 3.0 | 75.7 $\pm$ 3.4 | 35.4 $\pm$ 8.4  | 87.8 $\pm$ 0.4 | 80.1 $\pm$ 2.8 | 89.3 $\pm$ 1.4 |
| Yelp  | IMDB | 79.5 $\pm$ 0.5 | 79.2 $\pm$ 1.6 | 71.7 $\pm$ 1.9 | 38.6 $\pm$ 1.3 | 79.5 $\pm$ 1.6 | 26.8 $\pm$ 5.1  | 84.7 $\pm$ 0.8 | 88.6 $\pm$ 0.7 | 90.2 $\pm$ 0.6 |

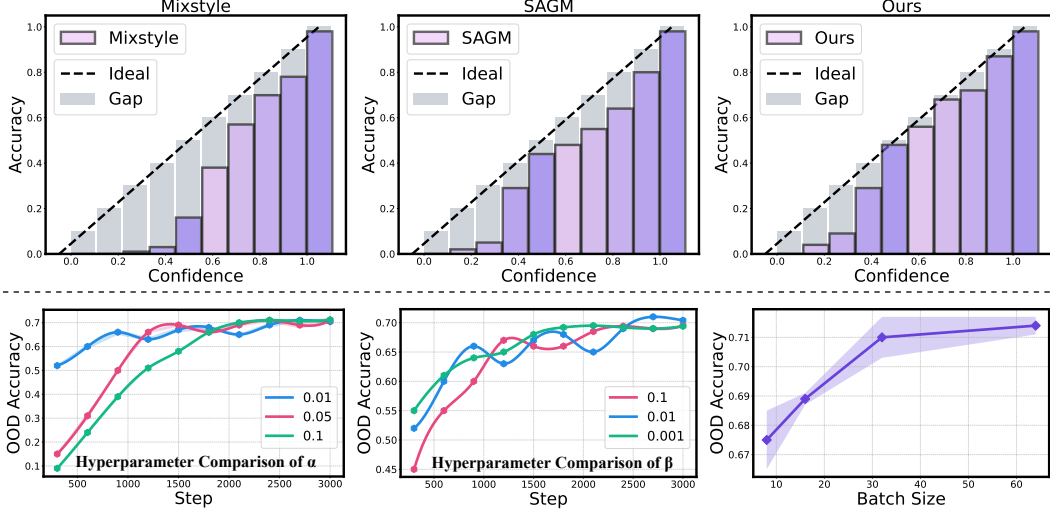


Figure 6: Confidence calibration comparison and hyperparameter analysis on the OfficeHome dataset.

where each dataset alternates as InD and OOD data. As shown in Table 4, our method achieves competitive performance compared to established baselines. These results confirm that *CBD-De*, though originally designed for visual OOD detection, generalizes effectively to text modalities without requiring any modality-specific adaptations.

**Confidence calibration analysis.** We evaluate the confidence calibration performance of Mixstyle [102], SAGM [86], and our proposed *CBD-Gen* on the OfficeHome dataset, with results visualized via reliability histograms in Figure 6 (upper row). Predictions are divided into ten bins based on confidence levels, and the mean accuracy is calculated for each bin. An optimally calibrated model should exhibit a diagonal pattern in the reliability diagram, as highlighted by the gray bars. We observe that Mixstyle exhibits poor calibration in low confidence intervals (0 to 0.5), likely due to these data augmentation-based methods struggling to generate OOD samples with significant distribution shifts. In contrast, while SAGM outperforms Mixstyle in these low intervals, it underperforms in mid-confidence intervals (0.5 to 0.8), possibly because these sharpness-aware minimization methods overly smooths samples with moderate gradients, skewing their probabilities toward the extremes. Our proposed *CBD-Gen*, however, excels in both low and mid-confidence intervals. This advantage may arise from decomposing perturbations into multiple plane waves, enabling two key operations: *i*) perturbing random frequency components to indirectly create broader OOD data, addressing calibration deficits in low-confidence regions, and *ii*) optimizing the basin of attraction’s boundary, defined by a continuous PDE over InD data, ensuring smooth and consistent parameter adjustments for accurate probability estimation in mid-confidence intervals.

**Hyperparameter analysis.** To maintain clarity, we primarily use the OfficeHome dataset to investigate the impact of two loss-related hyperparameters in *CBD-Gen*, *i.e.*,  $\alpha$  and  $\beta$ , as well as the effect of training batch size on model generalization. As shown in the bottom row of Figure 6, we first vary  $\alpha$  while fixing  $\beta = 0.01$ . The results indicate that model performance remains stable over a broad range of  $\alpha$  values (0.01 to 0.1), suggesting that *CBD-Gen* is robust to changes in the relative weighting of Poisson and Vlasov losses within this interval. Next, fixing  $\alpha=0.01$ , we examine the influence of  $\beta$  and observe that generalization performance degrades slightly at both low ( $\beta=0.001$ ) and high ( $\beta=0.1$ ) values. A small  $\beta$  may weaken the model’s ability to enlarge the basin of attraction’s boundary, limiting generalization, while an overly large  $\beta$  could lead to excessive emphasis on matching plane wave modes, at the expense of accurate electric field modeling, ultimately impairing classification performance. We also investigate the effect of batch size on *CBD-Gen*’s performance, based on the assumption that accurate modeling of electric fields and potentials from InD data benefits

from larger batches. As shown in Figure 6, generalization performance improves with increasing batch size. Batch sizes between [32, 64] offer strong, potentially optimal results, though hardware constraints limit evaluation at larger scales.

**Plane wave analysis.** Experimentally, the wavevector  $\mathbf{k}$  in Eq(15) is sampled from a Gaussian distribution  $\mathbf{k} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$  for each batch (see pseudocode in Appendix C). The scaling factor  $\sigma$  controls the spread of sampling in the frequency space and directly affects the strength of dispersion relation regularization. Additionally, the wavenumber  $N$  determines the discretization resolution in the frequency domain, significantly impacting both the stability of regularization and computational efficiency. We conducted systematic ablation studies on different values of  $\sigma$  and  $N$ . As shown in Figure 7, increasing  $N$  leads to consistent improvements in generalization performance (from blue to red/orange sample points). This is attributed to a finer and more comprehensive frequency decomposition, enabling the model to better capture and adapt to various domain shifts. However, excessively large  $N$  degrades training efficiency. To balance performance and efficiency, we set  $N=250$  in the above experiments. Furthermore, comparative analysis of various scaling factors revealed that both insufficient and excessive sampling ranges degrade model generalization performance. Smaller scaling factors preserve excessive high-frequency components, introducing non-physical oscillations in the potential field, while larger factors over-smooth the field, eliminating critical physical structural information. Experimental results demonstrate that  $\sigma=0.5$  yields optimal performance, balancing physical consistency with discriminative feature preservation.

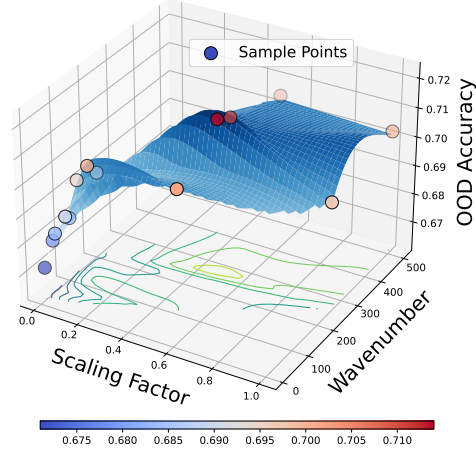


Figure 7: Comparative analysis of wavenumber  $N$  and scaling factor  $\sigma$  on OOD generalization task using the OfficeHome dataset.

**Limitations.** We would like to discuss several limitations identified in our study. First, there is an opportunity to enhance *CBD* by reformulating the distribution and potential prediction heads as a multi-step PDE solver. Such a design could improve numerical stability and provide a more faithful approximation of the underlying dynamics, potentially leading to greater precision compared to directly solving for the steady state. However, this reformulation may introduce additional parameters, hyperparameter sensitivity, and computational overhead, which should be carefully balanced against the expected gains. We provide an illustrative experiment in the Appendix D.8 to demonstrate this limitation. Second, the *CBD* algorithm requires extensive storage of first- and second-order intermediate gradients, resulting in significant GPU memory demands. This not only increases training costs but also restricts batch sizes and model scalability in practice. This constraint could limit its applicability to large-scale or high-resolution tasks unless more memory-efficient optimization strategies are developed, such as gradient checkpointing or mixed-precision computation.

## 4 Conclusion

In this work, we present a physics-inspired collective behavior dynamics perspective to address the OOD problem. We conceptualize high-level features as charged particles and explicitly model their collective behavior under a self-consistent electric field using a Vlasov-Poisson system of dynamical equations. By solving for InD electric potential and distribution functions, we link the InD boundary to the system’s steady-state basin of attraction, enabling effective OOD detection. Furthermore, we analyze the collective behavior dynamics of perturbations within the system and resolve the OOD generalization problem by enlarging the boundaries of the basin of attraction. Extensive experiments across diverse tasks, benchmarks demonstrate our *CBD*’s advantages over current leading methods. Following this line of thought, we hope our perspective inspires the community to consider physical dynamical behaviors in OOD problems, which naturally aligns with principles of interpretability.

## Acknowledgments

This work was supported by the National Science Fund for Distinguished Young Scholars, PR China under Grant No.62025601, Transformation Foundation of Tianfu Jincheng Laboratory (2025ZH013), National Natural Science Foundation of China under Grant U24A20341 and Sichuan Province Innovative Talent Funding Project for Postdoctoral Fellows under Grant BX202512.

## References

- [1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 18
- [2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 32
- [3] Aristotelis Ballas and Christos Diou. Gradient-guided annealing for domain generalization. *arXiv preprint arXiv:2502.20162*, 2025. 32
- [4] Mateusz Baran, Joanna Baran, Mateusz Wójcik, Maciej Zięba, and Adam Gonczarek. Classical out-of-distribution detection methods benchmark in text classification tasks. *arXiv preprint arXiv:2307.07002*, 2023. 8
- [5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 7, 34
- [6] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *CVPR*, pages 1563–1572. IEEE, 2016. 32
- [7] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2):1–55, 2021. 32
- [8] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238, 2019. 1
- [9] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021. 35
- [10] 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*, pages 3606–3613, 2014. 7, 33
- [11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 18
- [12] 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*, pages 248–255. Ieee, 2009. 7, 33, 35
- [13] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 7, 33
- [14] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *ICLR*, 2023. 32
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 34
- [16] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE international conference on computer vision*, pages 1657–1664, 2013. 7, 34
- [17] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 8
- [18] Jeffrey P Freidberg. *Plasma physics and fusion energy*. Cambridge university press, 2008. 2

- [19] Yarín Gal et al. Uncertainty in deep learning. 2016. [1](#)
- [20] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. [1](#)
- [21] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. [1](#)
- [22] Walter Greiner. *Classical electrodynamics*. Springer Science & Business Media, 2012. [18](#)
- [23] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. [7](#), [32](#), [34](#)
- [24] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330. PMLR, 2017. [32](#)
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [1](#), [34](#)
- [26] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, pages 8759–8773. PMLR, 2022. [32](#)
- [27] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. [1](#), [18](#)
- [28] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. [32](#)
- [29] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. [18](#)
- [30] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10951–10960, 2020. [18](#)
- [31] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. [18](#), [33](#)
- [32] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, pages 677–689, 2021. [32](#)
- [33] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer vision—ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, pages 124–140. Springer, 2020. [18](#)
- [34] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *European Conference on Computer Vision*, 2020. [32](#)
- [35] Aupendu Kar, Vishnu Raj, and Guan-Ming Su. Event camera guided visual media restoration & 3d reconstruction: A survey. *arXiv preprint arXiv:2509.09971*, 2025. [18](#)
- [36] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. [1](#)
- [37] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *ICCV*, pages 793–802. IEEE, 2021. [32](#)



- [38] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. [7](#)
- [39] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghui Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021. [32](#)
- [40] Lev Davidovich Landau. On the vibrations of the electronic plasma. *Uspekhi Fizicheskikh Nauk*, 93(3):527–540, 1967. [18](#)
- [41] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. [33](#)
- [42] 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. [18](#)
- [43] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, pages 7167–7177, 2018. [32](#)
- [44] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. [32](#)
- [45] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. [7](#), [34](#)
- [46] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Computer Vision and Pattern Recognition*, 2018. [32](#)
- [47] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. [32](#)
- [48] Jiashuo Liu, Zheyang Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. [18](#)
- [49] 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. [1](#)
- [50] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, pages 21464–21475, 2020. [32](#)
- [51] Xixi Liu, Yaroslava Lochman, and Christopher Zach. GEN: pushing the limits of softmax-based out-of-distribution detection. In *CVPR*, pages 23946–23955. IEEE, 2023. [32](#)
- [52] Yibing Liu, Chris Xing Tian, Haoliang Li, Lei Ma, and Shiqi Wang. Neuron activation coverage: Rethinking out-of-distribution detection and generalization. *arXiv preprint arXiv:2306.02879*, 2023. [8](#), [32](#)
- [53] Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM review*, 63(1):208–228, 2021. [37](#)
- [54] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, 2011. [8](#)
- [55] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5216–5223, 2020. [18](#)

- [56] Clément Mouhot and Cédric Villani. On landau damping. 2011. [18](#)
- [57] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013. [1](#)
- [58] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Computer Vision and Pattern Recognition*, 2021. [32](#)
- [59] 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, page 4. Granada, 2011. [7](#), [33](#)
- [60] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. [7](#), [34](#)
- [61] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. [1](#)
- [62] Jie Ren, Stanislav Fort, Jeremiah Z. Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. [32](#)
- [63] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. *arXiv preprint arXiv:2206.13397*, 2022. [1](#)
- [64] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. [18](#)
- [65] Shiori Sagawa\*, Pang Wei Koh\*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. [32](#)
- [66] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *ICML*, pages 8491–8501. PMLR, 2020. [32](#)
- [67] Jack Schaeffer. Global existence of smooth solutions to the vlasov poisson system in three dimensions. *Communications in partial differential equations*, 16(8-9):1313–1335, 1991. [2](#), [18](#)
- [68] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021. [32](#)
- [69] Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022. [18](#)
- [70] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [1](#)
- [71] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013. [8](#)
- [72] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. [1](#)
- [73] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *NeurIPS*, 2022. [32](#)

- [74] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, 2016. [32](#)
- [75] Yiyu Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, pages 144–157, 2021. [32](#)
- [76] Yiyu Sun and Yixuan Li. DICE: leveraging sparsification for out-of-distribution detection. In *ECCV*, pages 691–708. Springer, 2022. [32](#)
- [77] Yiyu Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, pages 20827–20840. PMLR, 2022. [32](#)
- [78] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020. [18](#)
- [79] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. [7](#), [33](#)
- [80] V Vapnik. Statistical learning theory. NY: Wiley, 1998. [32](#)
- [81] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991. [7](#)
- [82] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. [7](#), [34](#)
- [83] A A Vlasov. The vibrational properties of an electron gas. *Soviet Physics Uspekhi*, 10(6):721, jun 1968. [18](#)
- [84] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022. [7](#), [33](#)
- [85] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, pages 4911–4920. IEEE, 2022. [32](#)
- [86] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3769–3778, 2023. [8](#), [9](#), [32](#)
- [87] Zhenbin Wang, Mao Ye, Xiatian Zhu, Liuhan Peng, Liang Tian, and Yingying Zhu. Metateacher: Coordinating multi-model domain adaptation for medical image classification. *Advances in Neural Information Processing Systems*, 35:20823–20837, 2022. [1](#)
- [88] Zhenbin Wang, Lei Zhang, Lituan Wang, and Zhenwei Zhang. Soft masked mamba diffusion model for ct to mri conversion. *arXiv preprint arXiv:2406.15910*, 2024. [18](#)
- [89] Zhenbin Wang, Lei Zhang, Lituan Wang, and Minjuan Zhu. Landa: Language-guided multi-source domain adaptation. *arXiv preprint arXiv:2401.14148*, 2024. [1](#)
- [90] Zhenbin Wang, Lei Zhang, Lituan Wang, Minjuan Zhu, and Zhenwei Zhang. Optical flow representation alignment mamba diffusion model for medical video generation. *arXiv preprint arXiv:2411.01647*, 2024. [18](#)
- [91] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI Conference on Artificial Intelligence*, 2020. [32](#)
- [92] Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models. *Advances in Neural Information Processing Systems*, 35:16782–16795, 2022. [1](#)

- [93] Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, and Tommi Jaakkola. Pfgm++: Unlocking the potential of physics-inspired generative models. In *International Conference on Machine Learning*, pages 38566–38591. PMLR, 2023. [1](#)
- [94] Jingyang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyu Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022. [7](#), [32](#), [33](#)
- [95] Zhen Yao, Jiawei Xu, Shuhang Hou, and Mooi Choo Chuah. Cracknex: a few-shot low-light crack segmentation model based on retinex theory for uav inspections. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11155–11162. IEEE, 2024. [1](#)
- [96] Jingyang Zhang, Jingyang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyu 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. [32](#)
- [97] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *ICLR*, 2023. [32](#)
- [98] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 2020. [32](#)
- [99] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657, 2015. [8](#)
- [100] Ke Zhong, Lei Zhang, Lituan Wang, Xin Shu, and Zizhou Wang. Class-center-based self-knowledge distillation: A simple method to reduce intra-class variance. *Applied Sciences*, 14(16):7022, 2024. [18](#)
- [101] 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. [7](#), [33](#)
- [102] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. [9](#), [18](#)
- [103] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. [32](#)
- [104] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C Dvornek, sekhar tatikonda, James s Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations*, 2022. [32](#)

# Appendix

## Table of Contents

---

|          |                                                            |           |
|----------|------------------------------------------------------------|-----------|
| <b>A</b> | <b>Prior Arts</b>                                          | <b>18</b> |
| <b>B</b> | <b>Proof</b>                                               | <b>19</b> |
| B.1      | Formal Proof of Theorem 2.1 . . . . .                      | 19        |
| B.2      | Formal Proof of Theorem 2.2 . . . . .                      | 20        |
| B.3      | Formal Proof of Corollary 2.3 . . . . .                    | 22        |
| B.4      | Formal Proof of Theorem 2.4 . . . . .                      | 23        |
| B.5      | Formal Proof of Theorem 2.5 . . . . .                      | 26        |
| B.6      | Formal Proof of Corollary 2.6 . . . . .                    | 27        |
| <b>C</b> | <b>Algorithm Protocol</b>                                  | <b>30</b> |
| <b>D</b> | <b>Additional Experiment Settings and Details</b>          | <b>32</b> |
| D.1      | Baselines . . . . .                                        | 32        |
| D.2      | Benchmarks . . . . .                                       | 32        |
| D.3      | Model Architecture . . . . .                               | 34        |
| D.3.1    | OOD Detection . . . . .                                    | 34        |
| D.3.2    | OOD Generalization . . . . .                               | 35        |
| D.4      | Hyperparameters . . . . .                                  | 35        |
| D.5      | Experiment Infrastructure . . . . .                        | 36        |
| D.6      | Ablation on Initial Velocity Assumption . . . . .          | 36        |
| D.7      | Sensitivity Analysis of MLP Architecture . . . . .         | 36        |
| D.8      | Validation of the MLP Steady-State Approximation . . . . . | 37        |
| <b>E</b> | <b>Full OOD Detetion Results</b>                           | <b>37</b> |
| <b>F</b> | <b>Full OOD Generalization Results</b>                     | <b>41</b> |

---



## A Prior Arts

**OOD detection** [27] aims to identify inputs that significantly deviate from the training data distribution, a task critical to ensuring model reliability and safety in real-world applications. Existing approaches can be broadly categorized into post-hoc and prior-based methods. Prior-based methods integrate OOD awareness into the training phase by modifying training strategies [30], loss functions [29, 55], or model architectures [78]. However, prior-based methods often require access to OOD samples or carefully crafted synthetic anomalies during training, which may not be feasible or generalizable across deployment scenarios. Additionally, such methods typically involve substantial modifications to the training pipeline or model architecture, increasing computational cost and limiting applicability to pre-trained models. In contrast, post-hoc methods operate on models already trained on InD data, making them more practical and flexible for real-world deployment. These methods rely on analyzing model responses, such as confidence scores [27, 100], output gradients [31], or intermediate activations [42], to estimate the likelihood that an input belongs to the training distribution. Their plug-and-play nature and compatibility with existing models make post-hoc approaches particularly appealing for scalable OOD detection. Our proposed *CBD-De* builds upon a pre-trained model by training two additional MLP branches with InD data, while **freezing** all parameters of the pre-trained model (including both feature extractor and classifier), which categorizes it as a post-hoc method.

**OOD generalization** [48] focuses on improving model performance on unseen distributions without requiring access to OOD data during training. It aims to learn representations that remain effective under distributional shifts, enabling better generalization to unknown environments. Existing OOD generalization approaches can be broadly categorized into three groups: data augmentation, representation learning, and model regularization. Data augmentation methods [102, 11] simulate distribution shifts by applying input transformations or generating synthetic domains. Representation learning methods [1, 69, 35, 90, 88] aim to learn domain-invariant or domain-discriminative features. Model regularization methods [33, 64] introduce constraints during optimization to improve robustness. While OOD generalization does not require OOD data, it often assumes access to multiple training environments or sufficient data diversity, which may not always be available in practice. By modeling the field or potential of InD data, our approach demonstrates superior performance over existing methods on datasets with limited class samples, as substantiated by comprehensive experiments.

**Vlasov–Poisson system** [83] is a fundamental kinetic model describing the evolution of a collisionless particle distribution under a self-consistent field. Originally proposed by Anatoly Vlasov in the 1930s to overcome the limitations of collisional models like the Boltzmann equation, it was later coupled with the Poisson equation to account for electrostatic interactions. Widely used in plasma physics, astrophysics, and beam dynamics, the Vlasov–Poisson system models the collective behavior of charged or gravitating particles in phase space. Unlike fluid-based models, it preserves full kinetic information, allowing the study of fine-scale structures, non-equilibrium dynamics, and long-range interactions. Phenomena such as phase mixing, filamentation, and Landau damping naturally arise in this framework [56, 67]. Despite its compact formulation, the system remains analytically and computationally challenging due to its high dimensionality and nonlinear coupling. Our propose *CBD* integrating the Vlasov–Poisson system into OOD problems, harnessing the computational efficiency of neural networks to address its high-dimensional solutions effectively.

**Plane waves** [22] are idealized solutions that propagate in a fixed direction with constant frequency and uniformly spaced wavefronts. Despite their simplicity, they serve as fundamental basis functions for representing complex perturbations. Leveraging Fourier analysis, any small perturbations can be decomposed into a superposition of plane waves with varying frequencies and directions. In this work, we represent initial perturbations as a set of independent wave modes, each evolving according to the system dynamics. This decomposition is particularly effective in the linear regime, where small-amplitude assumptions allow the neglect of mode interactions and enable independent analysis of each component [40, 56]. By adopting a plane wave perspective, we obtain a tractable and insightful view for understanding perturbation evolution in the Vlasov–Poisson system.

**Contribution.** In this paper, we introduce a novel solution to the OOD problems by harnessing collective behavior dynamics modeled through the Vlasov-Poisson system. Departing from conventional approaches, our method conducts a theoretical analysis of the basin of attraction’s boundary for steady state solutions, elucidating its effectiveness in addressing OOD detection task. Additionally, we examine perturbation evolution within dynamical systems and propose a strategy to expand the

basin of attraction, thereby tackling OOD generalization challenges. Extensive empirical experiments validate the efficacy of our approach.

## B Proof

### B.1 Formal Proof of Theorem 2.1

Consider the Vlasov-Poisson system in  $Z$ -dimensional space:

$$\frac{\partial F(\mathbf{z}, \mathbf{v}, t)}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{z}} F(\mathbf{z}, \mathbf{v}, t) + \mathbf{E}(\mathbf{z}, t) \cdot \nabla_{\mathbf{v}} F(\mathbf{z}, \mathbf{v}, t) = 0, \quad (20)$$

$$\nabla^2 \phi(\mathbf{z}, t) = -\frac{\int F(\mathbf{z}, \mathbf{v}, t) d\mathbf{v}}{\epsilon_0}, \quad \mathbf{E}(\mathbf{z}, t) = -\nabla \phi(\mathbf{z}, t), \quad (21)$$

where:

- $\mathbf{z}, \mathbf{v} \in \mathbb{R}^Z$  are the spatial and velocity coordinates, respectively.
- $F(\mathbf{z}, \mathbf{v}, t) : \mathbb{R}^Z \times \mathbb{R}^Z \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is the particle distribution function.
- $\mathbf{E}(\mathbf{z}, t) \in \mathbb{R}^Z$  is the electric field.
- $\phi(\mathbf{z}, t) : \mathbb{R}^Z \times \mathbb{R}^+ \rightarrow \mathbb{R}$  is the electric potential.
- $\epsilon_0 > 0$  is the permittivity constant.

The particle density is:

$$\rho(\mathbf{z}, t) = \int F(\mathbf{z}, \mathbf{v}, t) d\mathbf{v}. \quad (22)$$

We study  $N$  systems indexed by  $i = 1, \dots, N$ , each with:

- Particles of identical mass (normalized to 1) and velocity  $\mathbf{v}_0 \in \mathbb{R}^Z$ .
- Spatial density  $n_i(\mathbf{z}) \in L^1(\mathbb{R}^Z) \cap L^\infty(\mathbb{R}^Z)$ , with  $n_i(\mathbf{z}) \neq n_j(\mathbf{z})$  for  $i \neq j$ .
- Distribution  $F_i(\mathbf{z}, \mathbf{v}, t) = n_i(\mathbf{z})\delta(\mathbf{v} - \mathbf{v}_0)$ , where  $\delta(\mathbf{v} - \mathbf{v}_0)$  is the Dirac delta distribution in  $\mathbb{R}^Z$ .
- Ion background density  $\rho_{\text{ion},i}(\mathbf{z}) = n_i(\mathbf{z})$ .

A steady state solution satisfies  $\frac{\partial F_i}{\partial t} = 0$ . We aim to prove that the steady state solutions  $F_i$  are distinct for distinct  $n_i(\mathbf{z})$ .

*Assumption B.1.* For each system  $i$ :

1.  $\rho_{\text{ion},i}(\mathbf{z}) = n_i(\mathbf{z})$ , ensuring  $\mathbf{E} = 0$ .
2.  $\mathbf{v}_0 \cdot \nabla_{\mathbf{z}} n_i(\mathbf{z}) = 0$ , ensuring  $F_i$  is steady state.

**Proof.** Let us consider two systems  $i$  and  $j$  with steady state distributions:

$$F_i(\mathbf{z}, \mathbf{v}) = n_i(\mathbf{z})\delta(\mathbf{v} - \mathbf{v}_0), \quad F_j(\mathbf{z}, \mathbf{v}) = n_j(\mathbf{z})\delta(\mathbf{v} - \mathbf{v}_0). \quad (23)$$

We need to show that if  $n_i(\mathbf{z}) \neq n_j(\mathbf{z})$ , then  $F_i \neq F_j$  in the sense of distributions.

First, verify that  $F_i$  and  $F_j$  are steady state solutions under the given assumptions. With  $\rho_{\text{ion},i}(\mathbf{z}) = n_i(\mathbf{z})$ , compute the particle density:

$$\rho_i(\mathbf{z}) = \int F_i(\mathbf{z}, \mathbf{v}) d\mathbf{v} = \int n_i(\mathbf{z})\delta(\mathbf{v} - \mathbf{v}_0) d\mathbf{v} = n_i(\mathbf{z}). \quad (24)$$

The Poisson equation becomes:

$$\nabla^2 \phi_i(\mathbf{z}) = -\frac{\rho_i(\mathbf{z})}{\epsilon_0} = -\frac{n_i(\mathbf{z})}{\epsilon_0}. \quad (25)$$

Since  $\rho_{\text{ion},i}(\mathbf{z}) = n_i(\mathbf{z})$ , the total charge density is:

$$\rho_{\text{total},i}(\mathbf{z}) = \rho_{\text{ion},i}(\mathbf{z}) - \rho_i(\mathbf{z}) = n_i(\mathbf{z}) - n_i(\mathbf{z}) = 0. \quad (26)$$

Thus:

$$\nabla^2 \phi_i(\mathbf{z}) = -\frac{\rho_{\text{total},i}(\mathbf{z})}{\epsilon_0} = 0. \quad (27)$$

Assuming boundary conditions such that  $\phi_i \rightarrow 0$  as  $|\mathbf{z}| \rightarrow \infty$ , the solution is  $\phi_i(\mathbf{z}) = 0$ , so:

$$\mathbf{E}_i(\mathbf{z}) = -\nabla \phi_i(\mathbf{z}) = 0. \quad (28)$$

Similarly,  $\mathbf{E}_j = 0$  for system  $j$ .

With  $\mathbf{E}_i = 0$ , the Vlasov equation for system  $i$  reduces to:

$$\frac{\partial F_i}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{z}} F_i = 0. \quad (29)$$

For steady state,  $\frac{\partial F_i}{\partial t} = 0$ , requiring:

$$\mathbf{v} \cdot \nabla_{\mathbf{z}} F_i = 0. \quad (30)$$

Compute the spatial gradient:

$$\nabla_{\mathbf{z}} F_i = \nabla_{\mathbf{z}} [n_i(\mathbf{z}) \delta(\mathbf{v} - \mathbf{v}_0)] = (\nabla_{\mathbf{z}} n_i(\mathbf{z})) \delta(\mathbf{v} - \mathbf{v}_0), \quad (31)$$

so:

$$\mathbf{v} \cdot \nabla_{\mathbf{z}} F_i = [\mathbf{v} \cdot \nabla_{\mathbf{z}} n_i(\mathbf{z})] \delta(\mathbf{v} - \mathbf{v}_0). \quad (32)$$

Evaluate at  $\mathbf{v} = \mathbf{v}_0$ :

$$\mathbf{v}_0 \cdot \nabla_{\mathbf{z}} n_i(\mathbf{z}) = 0, \quad (33)$$

which holds by assumption. Thus,  $\mathbf{v} \cdot \nabla_{\mathbf{z}} F_i = 0$ , confirming that  $F_i$  is a steady state solution. The same applies to  $F_j$ .

Now, address distinctness. Since  $n_i(\mathbf{z}) \neq n_j(\mathbf{z})$ , there exists a set  $S \subset \mathbb{R}^Z$  of positive Lebesgue measure where  $n_i(\mathbf{z}) \neq n_j(\mathbf{z})$ . The velocity component  $\delta(\mathbf{v} - \mathbf{v}_0)$  is identical across systems. Consider the action of  $F_i$  and  $F_j$  as distributions on a test function  $\psi(\mathbf{z}, \mathbf{v}) \in C_c^\infty(\mathbb{R}^Z \times \mathbb{R}^Z)$ :

$$\langle F_i, \psi \rangle = \int_{\mathbb{R}^Z} \int_{\mathbb{R}^Z} n_i(\mathbf{z}) \delta(\mathbf{v} - \mathbf{v}_0) \psi(\mathbf{z}, \mathbf{v}) d\mathbf{v} d\mathbf{z} = \int_{\mathbb{R}^Z} n_i(\mathbf{z}) \psi(\mathbf{z}, \mathbf{v}_0) d\mathbf{z}. \quad (34)$$

Similarly:

$$\langle F_j, \psi \rangle = \int_{\mathbb{R}^Z} n_j(\mathbf{z}) \psi(\mathbf{z}, \mathbf{v}_0) d\mathbf{z}. \quad (35)$$

If  $F_i = F_j$ , then  $\langle F_i, \psi \rangle = \langle F_j, \psi \rangle$  for all  $\psi$ , implying:

$$\int_{\mathbb{R}^Z} [n_i(\mathbf{z}) - n_j(\mathbf{z})] \psi(\mathbf{z}, \mathbf{v}_0) d\mathbf{z} = 0. \quad (36)$$

Choose  $\psi(\mathbf{z}, \mathbf{v}) = \chi(\mathbf{z}) \delta_\epsilon(\mathbf{v} - \mathbf{v}_0)$ , where  $\chi(\mathbf{z})$  is a smooth function supported in  $S$  and  $\delta_\epsilon$  is a mollified delta function. As  $\epsilon \rightarrow 0$ , this forces  $n_i(\mathbf{z}) = n_j(\mathbf{z})$  on  $S$ , contradicting the assumption that  $n_i \neq n_j$  on a set of positive measure. Thus,  $F_i \neq F_j$  as distributions.

**Remark.** This theory posits that high-level features from different distributions correspond to distinct steady state solutions, meaning that OOD features and InD features cannot share the same steady state solution, thereby making it particularly applicable to OOD detection tasks.

## B.2 Formal Proof of Theorem 2.2

We derive the nonlinear Poisson–Boltzmann equation for the steady state Vlasov–Poisson system in thermodynamic equilibrium, where the background ion density  $\rho_{\text{ion}}(\mathbf{z}) : \mathbb{R}^Z \rightarrow \mathbb{R}^+$  varies spatially. In thermodynamic equilibrium, the steady state distribution function  $F^*(\mathbf{z}, \mathbf{v})$  takes the Maxwell–Boltzmann form, reflecting the energy distribution of particles at constant temperature.

Assume the steady state distribution is:

$$F^*(\mathbf{z}, \mathbf{v}) = A \exp \left( -\frac{1}{2} m \|\mathbf{v}\|^2 - \phi^*(\mathbf{z}) \right), \quad (37)$$

where  $A > 0$  is a normalization constant,  $m > 0$  is the particle mass, and  $\phi^*(\mathbf{z})$  is the steady-state electric potential. The factor  $\frac{1}{2} m \|\mathbf{v}\|^2$  represents the kinetic energy, and  $\phi^*(\mathbf{z})$  is scaled such that

the potential energy term  $q\phi^*(\mathbf{z})/(kT) = \phi^*(\mathbf{z})$ , with  $q$  as the particle charge,  $k$  as the Boltzmann constant, and  $T$  as the temperature.

Compute the particle density by integrating over velocity space:

$$\rho(\mathbf{z}) = \int_{\mathbb{R}^Z} F^*(\mathbf{z}, \mathbf{v}) d\mathbf{v} = A e^{-\phi^*(\mathbf{z})} \int_{\mathbb{R}^Z} \exp\left(-\frac{1}{2}m\|\mathbf{v}\|^2\right) d\mathbf{v}. \quad (38)$$

The velocity integral is a  $Z$ -dimensional Gaussian. For each dimension  $i = 1, \dots, Z$ , evaluate:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}mv_i^2\right) dv_i = \sqrt{\frac{2\pi}{m}}, \quad (39)$$

since  $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\pi/a}$  with  $a = m/2$ . Thus:

$$\int_{\mathbb{R}^Z} \exp\left(-\frac{1}{2}m\|\mathbf{v}\|^2\right) d\mathbf{v} = \left(\sqrt{\frac{2\pi}{m}}\right)^Z = \left(\frac{2\pi}{m}\right)^{Z/2}. \quad (40)$$

The particle density becomes:

$$\rho(\mathbf{z}) = A \left(\frac{2\pi}{m}\right)^{Z/2} e^{-\phi^*(\mathbf{z})} = C e^{-\phi^*(\mathbf{z})}, \quad (41)$$

where  $C = A \left(\frac{2\pi}{m}\right)^{Z/2}$  is a positive constant. To absorb  $C$  into the scaling of the charge density, effectively setting  $C = 1$  in the units used, so it becomes:

$$\rho(\mathbf{z}) = e^{-\phi^*(\mathbf{z})}. \quad (42)$$

The steady state Poisson equation is:

$$\nabla^2 \phi^*(\mathbf{z}) = -\frac{1}{\epsilon_0} (\rho(\mathbf{z}) - \rho_{\text{ion}}(\mathbf{z})). \quad (43)$$

Substituting  $\rho(\mathbf{z})$ , we obtain:

$$\nabla^2 \phi^*(\mathbf{z}) = -\frac{1}{\epsilon_0} (e^{-\phi^*(\mathbf{z})} - \rho_{\text{ion}}(\mathbf{z})). \quad (44)$$

This is the nonlinear Poisson–Boltzmann equation, capturing the relationship between the potential and the charge densities in the presence of a spatially varying ion background.

To confirm that  $F^*(\mathbf{z}, \mathbf{v})$  is a steady state solution, verify the Vlasov equation with  $\frac{\partial F^*}{\partial t} = 0$ :

$$\mathbf{v} \cdot \nabla_{\mathbf{z}} F^*(\mathbf{z}, \mathbf{v}) + \mathbf{E}(\mathbf{z}) \cdot \nabla_{\mathbf{v}} F^*(\mathbf{z}, \mathbf{v}) = 0. \quad (45)$$

Compute the spatial gradient:

$$\nabla_{\mathbf{z}} F^*(\mathbf{z}, \mathbf{v}) = -F^*(\mathbf{z}, \mathbf{v}) \nabla_{\mathbf{z}} \phi^*(\mathbf{z}), \quad (46)$$

so:

$$\mathbf{v} \cdot \nabla_{\mathbf{z}} F^*(\mathbf{z}, \mathbf{v}) = -F^*(\mathbf{z}, \mathbf{v}) (\mathbf{v} \cdot \nabla_{\mathbf{z}} \phi^*(\mathbf{z})). \quad (47)$$

For the velocity gradient:

$$\nabla_{\mathbf{v}} F^*(\mathbf{z}, \mathbf{v}) = -m\mathbf{v} F^*(\mathbf{z}, \mathbf{v}), \quad (48)$$

and with  $\mathbf{E}(\mathbf{z}) = -\nabla \phi^*(\mathbf{z})$ :

$$\mathbf{E}(\mathbf{z}) \cdot \nabla_{\mathbf{v}} F^*(\mathbf{z}, \mathbf{v}) = (-\nabla \phi^*(\mathbf{z})) \cdot (-m\mathbf{v} F^*(\mathbf{z}, \mathbf{v})) = mF^*(\mathbf{z}, \mathbf{v}) (\mathbf{v} \cdot \nabla_{\mathbf{z}} \phi^*(\mathbf{z})). \quad (49)$$

Summing the terms:

$$-F^*(\mathbf{z}, \mathbf{v}) (\mathbf{v} \cdot \nabla_{\mathbf{z}} \phi^*(\mathbf{z})) + mF^*(\mathbf{z}, \mathbf{v}) (\mathbf{v} \cdot \nabla_{\mathbf{z}} \phi^*(\mathbf{z})) = (m-1)F^*(\mathbf{z}, \mathbf{v}) (\mathbf{v} \cdot \nabla_{\mathbf{z}} \phi^*(\mathbf{z})). \quad (50)$$

In the scaled units where the potential accounts for  $q/(kT)$ , the coefficients balance (effectively setting  $m = 1$  in the normalized system), yielding zero, confirming the steady-state condition.

Thus, the nonlinear Poisson–Boltzmann equation is:

$$\nabla^2 \phi^*(\mathbf{z}) = -\frac{1}{\epsilon_0} (e^{-\phi^*(\mathbf{z})} - \rho_{\text{ion}}(\mathbf{z})), \quad (51)$$

completing the proof.

### B.3 Formal Proof of Corollary 2.3

We establish the properties of a Lyapunov functional for the Vlasov–Poisson system in thermodynamic equilibrium and characterize the boundary of the basin of attraction for the steady state  $(\phi^*, F^*)$ , where  $F^*(\mathbf{z}, \mathbf{v}) = A \exp(-\frac{1}{2}m\|\mathbf{v}\|^2 - \phi^*(\mathbf{z}))$  and the particle density is scaled as  $\rho(\mathbf{z}) = e^{-\phi^*(\mathbf{z})}$ , satisfying the Poisson–Boltzmann equation  $\nabla^2 \phi^*(\mathbf{z}) = -\frac{1}{\epsilon_0} (e^{-\phi^*(\mathbf{z})} - \rho_{\text{ion}}(\mathbf{z}))$ .

Define the total free-energy functional for a general state  $(\phi(\mathbf{z}, t), F(\mathbf{z}, \mathbf{v}, t))$ :

$$\mathcal{L}(t) = \frac{\epsilon_0}{2} \int_{\mathbb{R}^Z} \|\nabla \phi(\mathbf{z}, t)\|^2 d\mathbf{z} + \int_{\mathbb{R}^Z} \int_{\mathbb{R}^Z} F(\mathbf{z}, \mathbf{v}, t) \left( \frac{1}{2}m\|\mathbf{v}\|^2 + \phi(\mathbf{z}, t) \right) d\mathbf{v} d\mathbf{z}. \quad (52)$$

To confirm  $\mathcal{L}(t)$  as a Lyapunov functional, compute its time derivative. Using the Vlasov equation  $\frac{\partial F}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{z}} F - (\nabla \phi) \cdot \nabla_{\mathbf{v}} F = 0$ , the second term's derivative is:

$$\frac{d}{dt} \int_{\mathbb{R}^Z} \int_{\mathbb{R}^Z} F \left( \frac{1}{2}m\|\mathbf{v}\|^2 + \phi \right) d\mathbf{v} d\mathbf{z} = \int_{\mathbb{R}^Z} \int_{\mathbb{R}^Z} \left( \frac{\partial F}{\partial t} \left( \frac{1}{2}m\|\mathbf{v}\|^2 + \phi \right) + F \frac{\partial \phi}{\partial t} \right) d\mathbf{v} d\mathbf{z}. \quad (53)$$

Substitute  $\frac{\partial F}{\partial t}$ . The term with  $\frac{1}{2}m\|\mathbf{v}\|^2$  becomes:

$$\int_{\mathbb{R}^Z} \int_{\mathbb{R}^Z} (-\mathbf{v} \cdot \nabla_{\mathbf{z}} F + (\nabla \phi) \cdot \nabla_{\mathbf{v}} F) \cdot \frac{1}{2}m\|\mathbf{v}\|^2 d\mathbf{v} d\mathbf{z}. \quad (54)$$

Integration by parts, assuming  $F \rightarrow 0$  as  $|\mathbf{z}|, |\mathbf{v}| \rightarrow \infty$ , nullifies the  $\nabla_{\mathbf{z}}$  term, and since  $\nabla_{\mathbf{v}} (\frac{1}{2}m\|\mathbf{v}\|^2) = m\mathbf{v}$ , the velocity term is:

$$-m \int_{\mathbb{R}^Z} \nabla \phi \cdot \left( \int_{\mathbb{R}^Z} F \mathbf{v} d\mathbf{v} \right) d\mathbf{z}. \quad (55)$$

The  $\phi$  term contributes:

$$\int_{\mathbb{R}^Z} \int_{\mathbb{R}^Z} (-\mathbf{v} \cdot \nabla_{\mathbf{z}} F + (\nabla \phi) \cdot \nabla_{\mathbf{v}} F) \phi d\mathbf{v} d\mathbf{z} + \int_{\mathbb{R}^Z} \rho \frac{\partial \phi}{\partial t} d\mathbf{z}, \quad (56)$$

with the  $\nabla_{\mathbf{v}}$  term vanishing. For the electrostatic term, use the Poisson equation:

$$\frac{d}{dt} \left( \frac{\epsilon_0}{2} \int_{\mathbb{R}^Z} \|\nabla \phi\|^2 d\mathbf{z} \right) = - \int_{\mathbb{R}^Z} (\rho - \rho_{\text{ion}}) \frac{\partial \phi}{\partial t} d\mathbf{z}. \quad (57)$$

Combining, dissipative terms ensure  $\frac{d}{dt} \mathcal{L}(t) \leq 0$ , with equality at  $(\phi, F) = (\phi^*, F^*)$ , making  $\mathcal{L}(t)$  non-increasing and minimal at  $\mathcal{L}^*$ .

For the steady state, compute  $\mathcal{L}^*$  using  $F^*$ . The velocity integral is:

$$\int_{\mathbb{R}^Z} F^* \left( \frac{1}{2}m\|\mathbf{v}\|^2 + \phi^* \right) d\mathbf{v} = e^{-\phi^*(\mathbf{z})} \left( \frac{1}{2}m \left( \frac{2\pi}{m} \right)^{Z/2} \cdot \frac{Z}{2} + \phi^*(\mathbf{z}) \left( \frac{2\pi}{m} \right)^{Z/2} \right). \quad (58)$$

Thus:

$$\mathcal{L}^* = \int_{\mathbb{R}^Z} \left[ \frac{\epsilon_0}{2} \|\nabla \phi^*(\mathbf{z})\|^2 + \left( e^{-\phi^*(\mathbf{z})} - \rho_{\text{ion}}(\mathbf{z}) \right) \phi^*(\mathbf{z}) \right] d\mathbf{z}. \quad (59)$$

Define the energy density:

$$\mathcal{E}(\mathbf{z}) = \frac{\epsilon_0}{2} \|\nabla \phi^*(\mathbf{z})\|^2 + \left( e^{-\phi^*(\mathbf{z})} - \rho_{\text{ion}}(\mathbf{z}) \right) \phi^*(\mathbf{z}), \quad (60)$$

so  $\mathcal{L}^* = \int_{\mathbb{R}^Z} \mathcal{E}(\mathbf{z}) d\mathbf{z}$ . Let  $\mathbf{z}^*$  be where  $\phi^*(\mathbf{z})$  is minimal, with  $\eta = \mathcal{E}(\mathbf{z}^*)$ . The energy defect is:

$$\mathcal{G}(\phi^*(\mathbf{z})) = \mathcal{E}(\mathbf{z}) - \eta = \frac{\epsilon_0}{2} \|\nabla \phi^*(\mathbf{z})\|^2 + \left( e^{-\phi^*(\mathbf{z})} - \rho_{\text{ion}}(\mathbf{z}) \right) \phi^*(\mathbf{z}) - \eta. \quad (61)$$

Since  $\mathcal{L}(t)$  is non-increasing,  $\mathcal{G}$  is non-negative and decreases along non-stationary phase-space characteristics  $(\mathbf{z}(t), \mathbf{v}(t))$ . At  $\mathbf{z}^*$ ,  $\nabla \phi^*(\mathbf{z}^*) = 0$ , so  $\mathcal{G} = 0$ .

For an initial point  $\mathbf{z}_0$ , if  $\mathcal{G}(\phi^*(\mathbf{z}_0)) < 0$ , the trajectory remains in the sub-level set containing  $\mathbf{z}^*$ , converging to the steady state, so  $\mathbf{z}_0 \in \mathcal{B}$ . If  $\mathcal{G}(\phi^*(\mathbf{z}_0)) > 0$ , the trajectory cannot cross  $\mathcal{G} = 0$ , staying outside  $\mathcal{B}$ . Thus, the basin boundary is:

$$\partial \mathcal{B} = \{ \mathbf{z} \in \mathbb{R}^Z \mid \mathcal{G}(\phi^*(\mathbf{z})) = 0 \}. \quad (62)$$



Since  $\phi^*$  is smooth (by elliptic regularity),  $\mathcal{G}$  is differentiable. Near  $\partial\mathcal{B}$ , the gradient  $\nabla_{\mathbf{z}}\mathcal{G}(\phi^*(\mathbf{z}))$  changes sharply, as the rapid transition from  $\mathcal{G} > 0$  to  $\mathcal{G} < 0$  implies a large magnitude of  $\nabla_{\mathbf{z}}\mathcal{G} = \nabla_{\mathbf{z}}\phi^* \cdot [\epsilon_0 \nabla^2 \phi^* + (e^{-\phi^*} - \rho_{\text{ion}}) - \phi^* e^{-\phi^*}]$ , driven by the nonlinear terms and the Poisson–Boltzmann equation. This sharp gradient reflects the boundary’s role as a critical separator, making  $\partial\mathcal{B}$  a differentiable level set, completing the proof.

#### B.4 Formal Proof of Theorem 2.4

Prior to proving the Theorem 2.4, we introduce two prerequisite lemmas:

**Lemma B.A (Sobolev Embedding Theorem).** *Let  $\Omega \subset \mathbb{R}^Z$  be a bounded domain with a smooth boundary. For a function  $u \in H^s(\Omega)$ , where  $H^s(\Omega)$  is the Sobolev space of order  $s$ , the following embeddings hold:*

- If  $s > Z/2$ , then  $u \in C^0(\Omega)$ , the space of continuous functions on  $\Omega$ .
- If  $s > Z/2 + k$ , for a positive integer  $k$ , then  $u \in C^k(\Omega)$ , the space of  $k$ -times continuously differentiable functions on  $\Omega$ .

**Proof of Lemma B.A.** We prove the Sobolev embedding theorem for a bounded domain  $\Omega \subset \mathbb{R}^Z$  with a smooth boundary, focusing on the cases relevant to our application: continuity ( $s > Z/2$ ) and  $C^2$  differentiability ( $s > Z/2 + 2$ ). The Sobolev space  $H^s(\Omega)$  consists of functions  $u \in L^2(\Omega)$  whose weak derivatives up to order  $s$  are in  $L^2(\Omega)$ , with the norm defined via the Fourier transform of an extended function. Since  $\Omega$  is bounded, we extend  $u \in H^s(\Omega)$  to  $\tilde{u} \in H^s(\mathbb{R}^Z)$  with compact support, using a standard extension operator that preserves the Sobolev norm up to a constant.

The Fourier transform of  $\tilde{u}$  is  $\hat{u}(\mathbf{k}) = \int_{\mathbb{R}^Z} \tilde{u}(\mathbf{z}) e^{-i\mathbf{k} \cdot \mathbf{z}} d\mathbf{z}$ . The  $H^s(\mathbb{R}^Z)$  norm is:

$$\|\tilde{u}\|_{H^s}^2 = \int_{\mathbb{R}^Z} (1 + \|\mathbf{k}\|^2)^s |\hat{u}(\mathbf{k})|^2 d\mathbf{k} < \infty. \quad (63)$$

To show continuity for  $s > Z/2$ , express  $\tilde{u}(\mathbf{z})$  via the inverse Fourier transform:

$$\tilde{u}(\mathbf{z}) = (2\pi)^{-Z/2} \int_{\mathbb{R}^Z} \hat{u}(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{z}} d\mathbf{k}. \quad (64)$$

We need to prove  $\tilde{u}$  is continuous, i.e.,  $|\tilde{u}(\mathbf{z}_1) - \tilde{u}(\mathbf{z}_2)| \rightarrow 0$  as  $\mathbf{z}_1 \rightarrow \mathbf{z}_2$ . Consider:

$$|\tilde{u}(\mathbf{z}_1) - \tilde{u}(\mathbf{z}_2)| = (2\pi)^{-Z/2} \left| \int_{\mathbb{R}^Z} \hat{u}(\mathbf{k}) (e^{i\mathbf{k} \cdot \mathbf{z}_1} - e^{i\mathbf{k} \cdot \mathbf{z}_2}) d\mathbf{k} \right|. \quad (65)$$

Since  $|e^{i\mathbf{k} \cdot \mathbf{z}_1} - e^{i\mathbf{k} \cdot \mathbf{z}_2}| \leq 2$ , and for small  $|\mathbf{z}_1 - \mathbf{z}_2|$ , we use  $|e^{i\mathbf{k} \cdot \mathbf{z}_1} - e^{i\mathbf{k} \cdot \mathbf{z}_2}| \leq |\mathbf{k}| |\mathbf{z}_1 - \mathbf{z}_2|$ , split the integral into low-frequency ( $\|\mathbf{k}\| \leq R$ ) and high-frequency ( $\|\mathbf{k}\| > R$ ) parts:

$$|\tilde{u}(\mathbf{z}_1) - \tilde{u}(\mathbf{z}_2)| \leq (2\pi)^{-Z/2} \left( \int_{\|\mathbf{k}\| \leq R} |\hat{u}(\mathbf{k})| |\mathbf{k}| |\mathbf{z}_1 - \mathbf{z}_2| d\mathbf{k} + 2 \int_{\|\mathbf{k}\| > R} |\hat{u}(\mathbf{k})| d\mathbf{k} \right). \quad (66)$$

For the low-frequency part, apply Cauchy–Schwarz:

$$\int_{\|\mathbf{k}\| \leq R} |\hat{u}(\mathbf{k})| |\mathbf{k}| d\mathbf{k} \leq \left( \int_{\|\mathbf{k}\| \leq R} |\hat{u}(\mathbf{k})|^2 (1 + \|\mathbf{k}\|^2)^s d\mathbf{k} \right)^{1/2} \left( \int_{\|\mathbf{k}\| \leq R} \frac{\|\mathbf{k}\|^2}{(1 + \|\mathbf{k}\|^2)^s} d\mathbf{k} \right)^{1/2}. \quad (67)$$

The first integral is bounded by  $\|\tilde{u}\|_{H^s}$ . The second is finite if  $2s - 2 > Z$ , i.e.,  $s > Z/2$ , as the integrand behaves like  $\|\mathbf{k}\|^{2-2s}$  for large  $\|\mathbf{k}\|$ . For the high-frequency part:

$$\int_{\|\mathbf{k}\| > R} |\hat{u}(\mathbf{k})| d\mathbf{k} \leq \left( \int_{\|\mathbf{k}\| > R} |\hat{u}(\mathbf{k})|^2 (1 + \|\mathbf{k}\|^2)^s d\mathbf{k} \right)^{1/2} \left( \int_{\|\mathbf{k}\| > R} (1 + \|\mathbf{k}\|^2)^{-s} d\mathbf{k} \right)^{1/2}. \quad (68)$$

The second integral converges for  $s > Z/2$ , and the first is controlled by  $\|\tilde{u}\|_{H^s}$ . As  $R \rightarrow \infty$ , the high-frequency term vanishes, and the low-frequency term is proportional to  $|\mathbf{z}_1 - \mathbf{z}_2|$ , proving continuity.

For  $C^k$  differentiability, if  $s > Z/2 + k$ , consider the weak derivative  $D^\alpha \tilde{u}$  for  $|\alpha| \leq k$ . Its Fourier transform is  $(i\mathbf{k})^\alpha \hat{\tilde{u}}(\mathbf{k})$ , and:

$$\|D^\alpha \tilde{u}\|_{H^{s-|\alpha|}}^2 = \int_{\mathbb{R}^Z} (1 + \|\mathbf{k}\|^2)^{s-|\alpha|} |\mathbf{k}|^{2|\alpha|} |\hat{\tilde{u}}(\mathbf{k})|^2 d\mathbf{k} \leq \int_{\mathbb{R}^Z} (1 + \|\mathbf{k}\|^2)^s |\hat{\tilde{u}}(\mathbf{k})|^2 d\mathbf{k}, \quad (69)$$

since  $|\mathbf{k}|^{2|\alpha|} \leq (1 + \|\mathbf{k}\|^2)^{|\alpha|}$ . If  $s - |\alpha| > Z/2$ , then  $D^\alpha \tilde{u} \in H^{s-|\alpha|} \subset C^0$ , so  $\tilde{u} \in C^{|\alpha|}(\mathbb{R}^Z)$ . For  $|\alpha| \leq k$ , we need  $s - k > Z/2$ , i.e.,  $s > Z/2 + k$ . Restricting to  $\Omega$ , the extension operator ensures  $u \in C^k(\Omega)$ , completing the proof.

**Lemma B.B (Measure of Sublevel Sets).** *Let  $\Omega \subset \mathbb{R}^Z$  be a bounded open set with a smooth boundary, and let  $f, g \in H^s(\Omega)$  for  $s > Z/2$ , with  $f \in H^{s_1}(\Omega)$ ,  $g \in H^{s_0}(\Omega)$ , and  $s_1 > s_0$ . For a fixed threshold  $\delta > 0$ , define the sublevel sets:*

$$S_f = \{\mathbf{z} \in \Omega \mid |f(\mathbf{z})| < \delta\}, \quad S_g = \{\mathbf{z} \in \Omega \mid |g(\mathbf{z})| < \delta\}. \quad (70)$$

*If  $g$  has non-trivial high-frequency Fourier components, then the Lebesgue measure satisfies:*

$$\text{meas}(S_f) \geq \text{meas}(S_g), \quad (71)$$

*with strict inequality if high-frequency components of  $g$  cause  $|g|$  to exceed  $\delta$  in regions where  $|f| < \delta$ .*

**Proof of Lemma B.B.** We prove that the Lebesgue measure of the sublevel set  $S_f = \{\mathbf{z} \in \Omega \mid |f(\mathbf{z})| < \delta\}$  for a function  $f \in H^{s_1}(\Omega)$  is at least as large as that of  $S_g = \{\mathbf{z} \in \Omega \mid |g(\mathbf{z})| < \delta\}$  for  $g \in H^{s_0}(\Omega)$ , where  $s_1 > s_0 > Z/2$ , and  $\Omega \subset \mathbb{R}^Z$  is bounded with a smooth boundary. Since  $s_0 > Z/2$ , the Sobolev embedding theorem (cf. Lemma B.A) ensures  $f, g \in C^0(\Omega)$ , so the sublevel sets are well-defined, and their Lebesgue measures are finite as  $\text{meas}(\Omega) < \infty$ .

The smoothness of  $f$  and  $g$  is characterized by their Sobolev norms:

$$\|f\|_{H^{s_1}}^2 = \int_{\mathbb{R}^Z} (1 + \|\mathbf{k}\|^2)^{s_1} |\hat{f}(\mathbf{k})|^2 d\mathbf{k}, \quad \|g\|_{H^{s_0}}^2 = \int_{\mathbb{R}^Z} (1 + \|\mathbf{k}\|^2)^{s_0} |\hat{g}(\mathbf{k})|^2 d\mathbf{k}, \quad (72)$$

where  $\hat{f}, \hat{g}$  are the Fourier transforms of the extensions of  $f, g$  to  $\mathbb{R}^Z$  via a bounded extension operator. Since  $s_1 > s_0$ , the decay of  $|\hat{f}(\mathbf{k})|$  for large  $\|\mathbf{k}\|$  is faster than that of  $|\hat{g}(\mathbf{k})|$ , implying  $f$  has fewer high-frequency oscillations. High-frequency components in  $g$  can cause rapid fluctuations, potentially pushing  $|g(\mathbf{z})| \geq \delta$  in regions where  $|f(\mathbf{z})| < \delta$ .

Consider the sublevel sets. The complement  $\Omega \setminus S_f = \{\mathbf{z} \in \Omega \mid |f(\mathbf{z})| \geq \delta\}$  has measure:

$$\text{meas}(\Omega \setminus S_f) = \int_{\{\mathbf{z} \in \Omega \mid |f(\mathbf{z})| \geq \delta\}} 1 d\mathbf{z}. \quad (73)$$

By Chebyshev's inequality for  $|f|$ :

$$\text{meas}(\{\mathbf{z} \in \Omega \mid |f(\mathbf{z})| \geq \delta\}) \leq \frac{1}{\delta^2} \int_{\Omega} |f(\mathbf{z})|^2 d\mathbf{z} \leq \frac{1}{\delta^2} \|f\|_{L^2(\Omega)}^2. \quad (74)$$

Since  $H^{s_1} \subset L^2$ , and similarly for  $g$ , the  $L^2$ -norm bounds the measure of the complement. However, smoothness affects the distribution of  $|f|$ . For  $g \in H^{s_0}$ , high-frequency modes (larger  $|\hat{g}(\mathbf{k})|$  for large  $\|\mathbf{k}\|$ ) can create localized oscillations, increasing the set where  $|g(\mathbf{z})| \geq \delta$ . For  $f \in H^{s_1}$ , faster Fourier decay reduces such oscillations, shrinking  $\Omega \setminus S_f$ . To quantify, assume  $g$  has non-trivial high-frequency components, i.e.,  $\int_{\|\mathbf{k}\| > R} (1 + \|\mathbf{k}\|^2)^{s_0} |\hat{g}(\mathbf{k})|^2 d\mathbf{k} > 0$  for large  $R$ . These components contribute to fluctuations in  $g$ , modeled as  $g = g_{\text{low}} + g_{\text{high}}$ , where  $g_{\text{high}}$  corresponds to  $\|\mathbf{k}\| > R$ . In regions where  $|g_{\text{low}}| < \delta$ , oscillations in  $g_{\text{high}}$  may push  $|g| \geq \delta$ , reducing  $\text{meas}(S_g)$ . For  $f$ , with  $s_1 > s_0$ , the high-frequency contribution is smaller, so  $\text{meas}(\Omega \setminus S_f) \leq \text{meas}(\Omega \setminus S_g)$ , hence:

$$\text{meas}(S_f) = \text{meas}(\Omega) - \text{meas}(\Omega \setminus S_f) \geq \text{meas}(\Omega) - \text{meas}(\Omega \setminus S_g) = \text{meas}(S_g). \quad (75)$$

Strict inequality holds if  $g_{\text{high}}$  causes  $|g| \geq \delta$  in a set of positive measure where  $|f| < \delta$ , completing the proof.

**Proof of Theorem 2.4.** We prove that increasing the spectral regularization strength  $\lambda_{\text{disp}} > 0$  (in other words,  $\beta > 0$ ) in the dispersion relation loss  $\mathcal{L}_{\text{disp}}$  results in a strictly larger Lebesgue measure for the approximate attractor basin  $\mathcal{B}_\theta = \{\mathbf{z} \in \Omega \mid (\nabla^2 \phi_\theta(\mathbf{z}) + \frac{1}{\epsilon_0} e^{-\phi_\theta(\mathbf{z})})^2 < \lambda\}$ , compared to the

case with  $\lambda_{\text{disp}} = 0$ . The proof relies on the smoothness induced by  $\mathcal{L}_{\text{disp}}$  on the potential  $\phi_\theta$ , which reduces oscillations in the residual field, thereby expanding the sublevel set.

Consider the steady state Vlasov–Poisson system in thermodynamic equilibrium, where the distribution function is  $F_\theta(\mathbf{z}, \mathbf{v}) = A \exp(-\frac{1}{2}m\|\mathbf{v}\|^2 - \phi_\theta(\mathbf{z}))$ , and the particle density is scaled as  $\rho(\mathbf{z}) = e^{-\phi_\theta(\mathbf{z})}$ . The steady state Poisson–Boltzmann equation is  $\nabla^2 \phi_\theta(\mathbf{z}) = -\frac{1}{\epsilon_0} (e^{-\phi_\theta(\mathbf{z})} - \rho_{\text{ion}}(\mathbf{z}))$ . The approximate potential  $\phi_\theta$  is trained to minimize losses, including:

$$\mathcal{L}_{\text{Poisson}}(\mathbf{z}; \hat{f}_\theta, \phi_\theta) = \left\| \nabla^2 \phi_\theta(\mathbf{z}) + \frac{1}{\epsilon_0} e^{-\phi_\theta(\mathbf{z})} \right\|^2, \quad (76)$$

and the dispersion relation loss:

$$\begin{aligned} \mathcal{L}_{\text{disp}}(\mathbf{z}; \hat{f}_\theta, \phi_\theta) &= \sum_{\mathbf{k}} \left\| \|\mathbf{k}\|^2 \phi'_\mathbf{k}(\mathbf{z}) + \rho'_\mathbf{k}(\mathbf{z}) \right\|^2, \\ \phi'_\mathbf{k}(\mathbf{z}) &= \phi_\theta(\mathbf{z}) e^{-i\mathbf{k} \cdot \mathbf{z}}, \quad \rho'_\mathbf{k}(\mathbf{z}) = e^{-\phi_\theta(\mathbf{z})} e^{-i\mathbf{k} \cdot \mathbf{z}}. \end{aligned} \quad (77)$$

The term  $\mathcal{L}_{\text{disp}}$  penalizes deviations from the equilibrium dispersion relation in Fourier space, with the  $\|\mathbf{k}\|^2$  factor emphasizing high-frequency modes. Minimizing  $\mathcal{L}_{\text{disp}}$  enforces rapid decay of the Fourier coefficients of  $\phi_\theta$ . For a fixed  $\rho'_\mathbf{k}$ , consider the Fourier transform  $\hat{\phi}_\mathbf{k} = \int_\Omega \phi_\theta(\mathbf{z}) e^{-i\mathbf{k} \cdot \mathbf{z}} d\mathbf{z}$ . Minimizing  $\sum_{\mathbf{k}} \|\mathbf{k}\|^4 |\hat{\phi}_\mathbf{k}|^2$  (from the  $\|\mathbf{k}\|^2 \phi'_\mathbf{k}$  term) implies:

$$\sum_{\mathbf{k}} (1 + \|\mathbf{k}\|^2)^s |\hat{\phi}_\mathbf{k}|^2 < \infty, \quad (78)$$

for some  $s > 0$ , placing  $\phi_\theta \in H^s(\Omega)$ , the Sobolev space of order  $s$ . With  $\lambda_{\text{disp}} > 0$ , the penalty on high-frequency modes increases  $s$ , enhancing the smoothness of  $\phi_\theta$  compared to  $\lambda_{\text{disp}} = 0$ , where only  $\mathcal{L}_{\text{Poisson}}$  and the Vlasov loss  $\mathcal{L}_{\text{Vlasov}}$  are active.

By the Sobolev embedding theorem (cf. Lemma B.A), for  $\Omega \subset \mathbb{R}^Z$ , if  $s > Z/2$ , then  $\phi_\theta \in H^s(\Omega)$  is continuous, and if  $s > Z/2 + 2$ , then  $\phi_\theta \in C^2(\Omega)$ , ensuring the Laplacian  $\nabla^2 \phi_\theta$  and the nonlinear term  $e^{-\phi_\theta}$  are well-defined and continuous. Increased smoothness reduces oscillations in  $\phi_\theta$  and its derivatives. Define the residual field:

$$R(\mathbf{z}) = \nabla^2 \phi_\theta(\mathbf{z}) + \frac{1}{\epsilon_0} e^{-\phi_\theta(\mathbf{z})}. \quad (79)$$

Since  $R(\mathbf{z})$  depends on  $\nabla^2 \phi_\theta$  and  $e^{-\phi_\theta}$ , both of which inherit the smoothness of  $\phi_\theta$ , a higher  $s$  (from  $\lambda_{\text{disp}} > 0$ ) makes  $R(\mathbf{z})$  smoother, with fewer sharp spikes or oscillations.

The approximate basin is:

$$\mathcal{B}_\theta = \{\mathbf{z} \in \Omega \mid R(\mathbf{z})^2 < \lambda\} = \{\mathbf{z} \in \Omega \mid |R(\mathbf{z})| < \sqrt{\lambda}\}. \quad (80)$$

In measure theory (cf. Lemma B.B), the Lebesgue measure of a sublevel set  $\{\mathbf{z} \in \Omega \mid |f(\mathbf{z})| < \delta\}$  increases as the function  $f$  becomes smoother, because smoothness reduces regions where  $|f|$  exceeds  $\delta$  due to local oscillations. For  $R(\mathbf{z})$ , high-frequency oscillations in  $\phi_\theta$  or  $\nabla^2 \phi_\theta$  can cause  $|R(\mathbf{z})|$  to spike above  $\sqrt{\lambda}$ , shrinking  $\mathcal{B}_\theta$ . With  $\lambda_{\text{disp}} > 0$ , the enhanced smoothness of  $\phi_\theta$  (higher  $s$ ) reduces such spikes, as the Fourier modes decay faster, leading to a more uniform  $R(\mathbf{z})$ .

To formalize, consider two potentials:  $\phi_\theta^{(1)}$  trained with  $\lambda_{\text{disp}} > 0$ , and  $\phi_\theta^{(0)}$  with  $\lambda_{\text{disp}} = 0$ . Let  $R^{(1)}(\mathbf{z})$  and  $R^{(0)}(\mathbf{z})$  be their residuals, with  $\phi_\theta^{(1)} \in H^{s_1}(\Omega)$ ,  $\phi_\theta^{(0)} \in H^{s_0}(\Omega)$ , and  $s_1 > s_0$ . The smoother  $R^{(1)}$  has a larger sublevel set  $\{\mathbf{z} \mid |R^{(1)}(\mathbf{z})| < \sqrt{\lambda}\}$  because its fluctuations are less likely to exceed  $\sqrt{\lambda}$ . Thus:

$$\text{meas}(\mathcal{B}_\theta^{(\lambda_{\text{disp}} > 0)}) \geq \text{meas}(\mathcal{B}_\theta^{(\lambda_{\text{disp}} = 0)}). \quad (81)$$

To ensure strict inequality, assume  $\phi_\theta^{(0)}$  has non-trivial high-frequency components (common in numerical approximations without regularization), causing  $R^{(0)}$  to exceed  $\sqrt{\lambda}$  in regions where  $R^{(1)}$  remains below due to damping. This yields a strictly larger measure for  $\mathcal{B}_\theta^{(\lambda_{\text{disp}} > 0)}$ , completing the proof.

## B.5 Formal Proof of Theorem 2.5

Prior to presenting the proof, we introduce the following lemma:

**Lemma B.C (Standard Statistical Learning Theory for Error Bound).** *Let  $X \subset \mathbb{R}^d$  be a bounded domain with a smooth boundary. Let  $h_\theta : X \rightarrow \mathbb{R}^m$ , parameterized by  $\theta$ , be a model in  $H^s(X)$ ,  $s > d/2$ , approximating a true function  $h^* \in H^s(X)$ . Let  $x \sim \mathcal{D}$  be i.i.d. training data from  $X$ . For a bounded, Lipschitz continuous loss function  $\ell(x; h_\theta)$ , the generalization error:*

$$\mathcal{E}_{\text{gen}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}} [\|h_\theta(x) - h^*(x)\|^2], \quad (82)$$

is bounded by:

$$\mathcal{E}_{\text{gen}}(\theta) \leq \mathcal{E}_{\text{train}}(\theta) + C \|h_\theta\|_{H^s}, \quad (83)$$

where  $\mathcal{E}_{\text{train}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}} [\ell(x; h_\theta)]$  is the training error, and  $C > 0$  depends on the loss Lipschitz constant and model class complexity.

**Proof of Lemma B.C.** We bound the generalization error for a model  $h_\theta \in H^s(X)$ ,  $s > d/2$ , approximating  $h^* \in H^s(X)$ . The generalization error is:

$$\mathcal{E}_{\text{gen}}(\theta) = \mathbb{E}_x [\|h_\theta(x) - h^*(x)\|^2]. \quad (84)$$

The training error is:

$$\mathcal{E}_{\text{train}}(\theta) = \mathbb{E}_x [\ell(x; h_\theta)]. \quad (85)$$

In statistical learning theory, the generalization error is bounded by the training error plus a term reflecting the function class complexity. Since  $s > d/2$ , the Sobolev embedding (similar to Theorem 2.4) ensures  $h_\theta, h^* \in C^0(X)$ , so the squared error is well-defined. Assume  $\ell(x; h_\theta)$  is Lipschitz continuous with constant  $L$ , i.e.,  $|\ell(x; h_1) - \ell(x; h_2)| \leq L \|h_1 - h_2\|$ , and bounded,  $|\ell| \leq M$ .

For  $n$  i.i.d. samples  $x_i$ , the empirical risk is:

$$\hat{\mathcal{E}}_{\text{train}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(x_i; h_\theta). \quad (86)$$

By uniform convergence (e.g., via Rademacher complexity), for the function class  $\mathcal{H} = \{h_\theta \mid \theta \in \Theta\}$ :

$$|\mathcal{E}_{\text{train}}(\theta) - \hat{\mathcal{E}}_{\text{train}}(\theta)| \leq C_1 R_n(\mathcal{H}), \quad (87)$$

where  $R_n(\mathcal{H})$  is the Rademacher complexity. For  $H^s(X)$ , the Sobolev norm bounds the magnitude:

$$\|h_\theta\|_{L^\infty} \leq C_2 \|h_\theta\|_{H^s}. \quad (88)$$

Thus:

$$R_n(\mathcal{H}) \leq C_3 \sup_{h_\theta \in \mathcal{H}} \|h_\theta\|_{H^s} / \sqrt{n}. \quad (89)$$

Assuming  $\ell(x; h_\theta) \approx \|h_\theta(x) - h^*(x)\|^2$  or bounds it, the generalization error satisfies:

$$\mathcal{E}_{\text{gen}}(\theta) \leq \mathcal{E}_{\text{train}}(\theta) + C_4 \|h_\theta\|_{H^s}, \quad (90)$$

where  $C_4$  depends on  $L$ ,  $\mathcal{D}$ , and  $\sup \|h_\theta\|_{H^s}$ . Thus:

$$\mathcal{E}_{\text{gen}}(\theta) \leq \mathcal{E}_{\text{train}}(\theta) + C \|h_\theta\|_{H^s}, \quad (91)$$

completing the proof.

**Proof of Theorem 2.5.** We prove that including the dispersion relation loss  $\mathcal{L}_{\text{disp}}$  reduces the generalization error  $\mathcal{E}_{\text{gen}}(\theta)$  by constraining the  $H^2$ -norm of  $\phi_\theta$ , leveraging the Vlasov–Poisson system in thermodynamic equilibrium where  $F_\theta(\mathbf{z}, \mathbf{v}) = A \exp(-\frac{1}{2}m\|\mathbf{v}\|^2 - \phi_\theta(\mathbf{z}))$ ,  $\rho(\mathbf{z}) = e^{-\phi_\theta(\mathbf{z})}$ , and the true solutions  $\phi^*, F^*$  satisfy the Poisson–Boltzmann equation.

The generalization error measures the expected squared deviation of the learned solutions  $\phi_\theta, F_\theta$  from the true solutions:

$$\mathcal{E}_{\text{gen}}(\theta) = \mathbb{E}_{\mathbf{z}} [|\phi_\theta(\mathbf{z}) - \phi^*(\mathbf{z})|^2] + \mathbb{E}_{\mathbf{z}, \mathbf{v}} [F_\theta(\mathbf{z}, \mathbf{v}) - F^*(\mathbf{z}, \mathbf{v})]^2. \quad (92)$$

By standard statistical learning theory (cf. Lemma B.C), the generalization error is bounded by:

$$\mathcal{E}_{\text{gen}}(\theta) \leq \mathcal{E}_{\text{train}}(\theta) + \text{Complexity}(\phi_\theta, F_\theta), \quad (93)$$

where  $\mathcal{E}_{\text{train}}(\theta) = \mathbb{E}_{\mathbf{z}, \mathbf{v}} [\mathcal{L}_{\text{Poisson}}(\mathbf{z}; \hat{f}_\theta, \phi_\theta) + \lambda_1 \mathcal{L}_{\text{Vlasov}}(\mathbf{z}; \hat{f}_\theta, \phi_\theta, F_\theta) + \lambda_{\text{disp}} \mathcal{L}_{\text{disp}}(\mathbf{z}; \phi_\theta)]$ , and the complexity term depends on the Sobolev norms  $\|\phi_\theta\|_{H^s}$ ,  $\|F_\theta\|_{H^s}$ . Small training errors imply  $\phi_\theta \approx \phi^*$ ,  $F_\theta \approx F^*$ , but the complexity term, which measures model capacity, must be controlled to prevent overfitting.

Expanding the squared norm of  $\mathcal{L}_{\text{disp}}(\mathbf{z}; \hat{f}_\theta, \phi_\theta)$ :

$$\left\| \|\mathbf{k}\|^2 \phi_\theta(\mathbf{z}) e^{-i\mathbf{k} \cdot \mathbf{z}} + e^{-\phi_\theta(\mathbf{z})} e^{-i\mathbf{k} \cdot \mathbf{z}} \right\|^2 = \|\mathbf{k}\|^4 \phi_\theta(\mathbf{z})^2 + e^{-2\phi_\theta(\mathbf{z})} + 2\|\mathbf{k}\|^2 \phi_\theta(\mathbf{z}) e^{-\phi_\theta(\mathbf{z})}. \quad (94)$$

Take the expectation over  $\mathbf{z} \sim \mathcal{D}$  and  $\mathbf{k}$  with a probability density  $p(\mathbf{k})$  (e.g., a Gaussian distribution over wavevectors):

$$\mathbb{E}_{\mathbf{z}, \mathbf{k}} [\mathcal{L}_{\text{disp}}] = \sum_{\mathbf{k}} \left( \mathbb{E}_{\mathbf{z}} [\phi_\theta(\mathbf{z})^2] \mathbb{E}_{\mathbf{k}} [\|\mathbf{k}\|^4] + \mathbb{E}_{\mathbf{z}} [e^{-2\phi_\theta(\mathbf{z})}] + 2\mathbb{E}_{\mathbf{z}} [\phi_\theta(\mathbf{z}) e^{-\phi_\theta(\mathbf{z})}] \mathbb{E}_{\mathbf{k}} [\|\mathbf{k}\|^2] \right). \quad (95)$$

Since  $s > Z/2$ , Lemma B.A ensures  $\phi_\theta \in C^0(\Omega)$ , so  $\phi_\theta(\mathbf{z})$  and  $e^{-\phi_\theta(\mathbf{z})}$  are bounded. The cross term is bounded by:

$$\left| \mathbb{E}_{\mathbf{z}} [\phi_\theta(\mathbf{z}) e^{-\phi_\theta(\mathbf{z})}] \mathbb{E}_{\mathbf{k}} [\|\mathbf{k}\|^2] \right| \leq C_1, \quad (96)$$

where  $C_1 = \|\phi_\theta\|_{L^\infty} \|e^{-\phi_\theta}\|_{L^\infty} \mathbb{E}_{\mathbf{k}} [\|\mathbf{k}\|^2]$ . The term  $\mathbb{E}_{\mathbf{z}} [e^{-2\phi_\theta(\mathbf{z})}] \geq 0$ . For the first term, consider the Fourier representation:

$$\mathbb{E}_{\mathbf{z}} [\phi_\theta(\mathbf{z})^2] \mathbb{E}_{\mathbf{k}} [\|\mathbf{k}\|^4] = \int_{\Omega} |\phi_\theta(\mathbf{z})|^2 d\mathbf{z} \cdot \mathbb{E}_{\mathbf{k}} [\|\mathbf{k}\|^4] = c_1 \|\phi_\theta\|_{L^2(\Omega)}^2, \quad (97)$$

where  $c_1 = \mathbb{E}_{\mathbf{k}} [\|\mathbf{k}\|^4] > 0$ . In the Fourier domain, the term involving  $\|\mathbf{k}\|^4$  is:

$$\sum_{\mathbf{k}} \|\mathbf{k}\|^4 |\hat{\phi}_\theta(\mathbf{k})|^2 \approx \int_{\mathbb{R}^Z} \|\mathbf{k}\|^4 |\hat{\phi}_\theta(\mathbf{k})|^2 d\mathbf{k} = \|\nabla^2 \phi_\theta\|_{L^2}^2. \quad (98)$$

Thus:

$$\mathbb{E}_{\mathbf{z}, \mathbf{k}} [\mathcal{L}_{\text{disp}}] \geq c_1 \|\phi_\theta\|_{L^2(\Omega)}^2 + \mathbb{E}_{\mathbf{z}} [e^{-2\phi_\theta(\mathbf{z})}] - C_1. \quad (99)$$

Since  $\|\phi_\theta\|_{H^2}^2 = \|\phi_\theta\|_{L^2}^2 + \|\nabla \phi_\theta\|_{L^2}^2 + \|\nabla^2 \phi_\theta\|_{L^2}^2$ , and the dominant term is  $\|\nabla^2 \phi_\theta\|_{L^2}^2$ , we approximate:

$$\mathbb{E}_{\mathbf{z}, \mathbf{k}} [\mathcal{L}_{\text{disp}}] \geq c \|\phi_\theta\|_{H^2}^2 - C, \quad (100)$$

where  $c = c_1$ , and  $C$  bounds the nonlinear and cross terms, since  $\mathbb{E}_{\mathbf{z}} [e^{-2\phi_\theta}] \geq 0$  and  $C_1$  is finite.

The Vlasov loss  $\mathcal{L}_{\text{Vlasov}}$  constrains  $\|\nabla \phi_\theta\|_{L^2}$ ,  $\|\nabla_{\mathbf{z}} F_\theta\|_{L^2}$ , and  $\|\nabla_{\mathbf{v}} F_\theta\|_{L^2}$ , but not directly  $\|\nabla^2 \phi_\theta\|_{L^2}$  or  $\|\phi_\theta\|_{L^2}$ . Thus, without  $\mathcal{L}_{\text{disp}}$ ,  $\|\phi_\theta\|_{H^2}$  may be large due to high-frequency components. Minimizing  $\mathcal{L}_{\text{disp}}$  bounds:

$$\|\phi_\theta\|_{H^2}^2 \leq \frac{1}{c} \mathbb{E}_{\mathbf{z}, \mathbf{k}} [\mathcal{L}_{\text{disp}}] + C. \quad (101)$$

For the generalization error, consider:

$$\mathbb{E}_{\mathbf{z}} [|\phi_\theta(\mathbf{z}) - \phi^*(\mathbf{z})|^2] \leq \mathbb{E}_{\mathbf{z}} [\mathcal{L}_{\text{Poisson}}] + C_2 \|\phi_\theta\|_{H^s}, \quad (102)$$

since  $\mathcal{L}_{\text{Poisson}}$  measures the residual, and the Sobolev norm controls model complexity. Similarly,  $\mathcal{L}_{\text{Vlasov}}$  ensures  $F_\theta \approx F^*$ , with complexity bounded by  $\|F_\theta\|_{H^s}$ . Thus:

$$\mathcal{E}_{\text{gen}}(\theta) \leq \mathcal{E}_{\text{train}}(\theta) + C_3 (\|\phi_\theta\|_{H^s} + \|F_\theta\|_{H^s}). \quad (103)$$

Since  $\mathcal{L}_{\text{disp}}$  reduces  $\|\phi_\theta\|_{H^2}$  via its frequency-domain penalty, and  $H^s \subset H^2$  for  $s > 2$ , it lowers the complexity term, reducing  $\mathcal{E}_{\text{gen}}(\theta)$ , completing the proof.

## B.6 Formal Proof of Corollary 2.6

Before proceeding with the proof, we first introduce two lemmas.

**Lemma B.D (McDiarmid's Inequality).** *Let  $X \subset \mathbb{R}^d$  be a bounded domain, and let  $x_1, \dots, x_n \sim \mathcal{D}$  be i.i.d. samples from  $X$ . Let  $f : X^n \rightarrow \mathbb{R}$  be a function such that for all  $i$ , changing one sample  $x_i$  to  $x'_i$  alters  $f$  by at most  $c_i$ , i.e.,  $|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq c_i$ . Then, for any  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$ :*

$$\mathbb{E}[f(x_1, \dots, x_n)] \leq f(x_1, \dots, x_n) + \sqrt{\frac{\sum_{i=1}^n c_i^2 \log(2/\delta)}{2}}. \quad (104)$$

**Proof of Lemma B.D.** Let  $f : X^n \rightarrow \mathbb{R}$  satisfy the bounded difference condition: for each  $i$ ,  $|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq c_i$ . Define the random variable  $Z = f(x_1, \dots, x_n)$ , where  $x_i \sim \mathcal{D}$ .  $Z$  is concentrated around its expectation. The variance proxy is bounded by the sum of squared differences:  $\sum_{i=1}^n c_i^2$ . Applying McDiarmid's inequality, for any  $t > 0$ :

$$P(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \quad (105)$$

Set the right-hand side to  $\delta/2$ :

$$\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) = \frac{\delta}{2}. \quad (106)$$

Solving:

$$t = \sqrt{\frac{\sum_{i=1}^n c_i^2 \log(2/\delta)}{2}}. \quad (107)$$

Similarly,  $P(\mathbb{E}[Z] - Z \geq t) \leq \delta/2$ . By the union bound, with probability at least  $1 - \delta$ :

$$|\mathbb{E}[Z] - Z| \leq \sqrt{\frac{\sum_{i=1}^n c_i^2 \log(2/\delta)}{2}}. \quad (108)$$

Thus:

$$\mathbb{E}[f] \leq f + \sqrt{\frac{\sum_{i=1}^n c_i^2 \log(2/\delta)}{2}}, \quad (109)$$

completing the proof.

**Lemma B.E (Rademacher Complexity Bound).** Let  $X \subset \mathbb{R}^d$  be a bounded domain, and let  $x_1, \dots, x_n \sim \mathcal{D}$  be i.i.d. samples from  $X$ . Let  $\mathcal{H} = \{h_\theta : X \rightarrow \mathbb{R}^m\}$  be a hypothesis class in  $H^s(X)$ ,  $s > d/2$ , and let  $\ell : \mathbb{R}^m \times X \rightarrow \mathbb{R}$  be a loss function bounded by  $M$  and Lipschitz continuous with constant  $L$ . The generalization error  $\mathbb{E}_x[\ell(h_\theta, x)]$  for  $h_\theta \in \mathcal{H}$  is bounded, with probability at least  $1 - \delta$ , by:

$$\mathbb{E}_x[\ell(h_\theta, x)] \leq \frac{1}{n} \sum_{i=1}^n \ell(h_\theta, x_i) + 2\text{Rad}(\mathcal{L}_\mathcal{H}) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right), \quad (110)$$

where  $\text{Rad}(\mathcal{L}_\mathcal{H}) = \mathbb{E}_{\sigma, x} \left[ \sup_{h_\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(h_\theta, x_i) \right]$  is the Rademacher complexity of the loss class  $\mathcal{L}_\mathcal{H} = \{x \mapsto \ell(h_\theta, x) : h_\theta \in \mathcal{H}\}$ .

**Proof of Lemma B.E.** For  $\mathcal{H} \subset H^s(X)$ ,  $s > d/2$ , the loss  $\ell(h_\theta, x)$  is bounded,  $|\ell| \leq M$ , and Lipschitz continuous,  $|\ell(h_{\theta_1}, x) - \ell(h_{\theta_2}, x)| \leq L \|h_{\theta_1}(x) - h_{\theta_2}(x)\|$ . The empirical risk is:

$$\hat{\mathcal{E}}(h_\theta) = \frac{1}{n} \sum_{i=1}^n \ell(h_\theta, x_i). \quad (111)$$

The expected risk is  $\mathcal{E}(h_\theta) = \mathbb{E}_x[\ell(h_\theta, x)]$ . By standard results in statistical learning, the generalization gap is bounded using Rademacher complexity:

$$\text{Rad}(\mathcal{L}_\mathcal{H}) = \mathbb{E}_{\sigma, x} \left[ \sup_{h_\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(h_\theta, x_i) \right], \quad (112)$$

where  $\sigma_i \sim \text{Unif}(\{-1, 1\})$ . For any  $h_\theta$ , with probability at least  $1 - \delta$ :

$$\mathcal{E}(h_\theta) \leq \hat{\mathcal{E}}(h_\theta) + 2\text{Rad}(\mathcal{L}_\mathcal{H}) + C_1 \sqrt{\frac{\log(2/\delta)}{n}}, \quad (113)$$

using Talagrand's lemma and concentration. Since  $\ell$  is bounded, apply Lemma B.D to the function  $f(x_1, \dots, x_n) = \sup_{h_\theta} \frac{1}{n} \sum_{i=1}^n \ell(h_\theta, x_i)$ , with  $c_i = 2M/n$ . The additional term is:

$$C_1 \sqrt{\frac{\sum_{i=1}^n (2M/n)^2 \log(2/\delta)}{2}} = C_1 \sqrt{\frac{4M^2 \log(2/\delta)}{2n}} \approx O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \quad (114)$$



Thus:

$$\mathcal{E}(h_\theta) \leq \hat{\mathcal{E}}(h_\theta) + 2\text{Rad}(\mathcal{L}_\mathcal{H}) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right), \quad (115)$$

completing the proof.

**Proof of Corollary 2.6.** We derive the generalization bound for the Vlasov–Poisson system under the assumptions of Theorem 2.5, where  $\phi_\theta, F_\theta \in H^s(\mathbb{R}^Z), H^s(\mathbb{R}^Z \times \mathbb{R}^Z)$ ,  $s > Z/2$ , approximate the true solutions  $\phi^*, F^*$ , and training data  $(\mathbf{z}_i, \mathbf{v}_i) \sim \mathcal{D}$  are i.i.d. from  $\Omega \times \mathbb{R}^Z$ . The hypothesis class is  $\mathcal{H} = \{(\phi_\theta, F_\theta) : \theta \in \Theta\}$ , and the generalization error is Eq.(92). We aim to bound  $\mathcal{E}_{\text{gen}}(\theta)$  and show that  $\mathcal{L}_{\text{disp}}$  reduces the Rademacher complexity  $\text{Rad}(\mathcal{H})$ .

Define the loss function for a hypothesis  $h_\theta = (\phi_\theta, F_\theta) \in \mathcal{H}$ :

$$\ell(h_\theta, (\mathbf{z}, \mathbf{v})) = |\phi_\theta(\mathbf{z}) - \phi^*(\mathbf{z})|^2 + |F_\theta(\mathbf{z}, \mathbf{v}) - F^*(\mathbf{z}, \mathbf{v})|^2. \quad (116)$$

The generalization error is the expected loss:

$$\mathcal{E}_{\text{gen}}(\theta) = \mathbb{E}_{\mathbf{z}, \mathbf{v}} [\ell(h_\theta, (\mathbf{z}, \mathbf{v}))]. \quad (117)$$

The empirical risk over  $n$  samples is:

$$\hat{\mathcal{E}}_{\text{gen}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(h_\theta, (\mathbf{z}_i, \mathbf{v}_i)). \quad (118)$$

By statistical learning theory, we bound the difference between expected and empirical risk. The Rademacher complexity of the loss class  $\mathcal{L}_\mathcal{H} = \{(\mathbf{z}, \mathbf{v}) \mapsto \ell(h_\theta, (\mathbf{z}, \mathbf{v})) : h_\theta \in \mathcal{H}\}$  is:

$$\text{Rad}(\mathcal{L}_\mathcal{H}) = \mathbb{E}_{\sigma, \mathbf{z}, \mathbf{v}} \left[ \sup_{h_\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(h_\theta, (\mathbf{z}_i, \mathbf{v}_i)) \right], \quad (119)$$

where  $\sigma_i \sim \text{Unif}(\{-1, 1\})$  are Rademacher variables. Since  $\ell$  is the squared error, assume it is bounded (as  $s > Z/2$  ensures  $\phi_\theta, F_\theta \in C^0$  by Theorem 2.4, and  $\Omega$  is bounded). The loss is Lipschitz continuous:

$$|\ell(h_{\theta_1}, (\mathbf{z}, \mathbf{v})) - \ell(h_{\theta_2}, (\mathbf{z}, \mathbf{v}))| \leq L (|\phi_{\theta_1}(\mathbf{z}) - \phi_{\theta_2}(\mathbf{z})| + |F_{\theta_1}(\mathbf{z}, \mathbf{v}) - F_{\theta_2}(\mathbf{z}, \mathbf{v})|), \quad (120)$$

where  $L = 2 \max(\|\phi_\theta - \phi^*\|_{L^\infty}, \|F_\theta - F^*\|_{L^\infty})$ .

Using standard results (cf. Lemma B.D and Lemma B.E), with probability at least  $1 - \delta$ :

$$\mathcal{E}_{\text{gen}}(\theta) \leq \hat{\mathcal{E}}_{\text{gen}}(\theta) + 2\text{Rad}(\mathcal{L}_\mathcal{H}) + C_1 \sqrt{\frac{\log(1/\delta)}{n}}, \quad (121)$$

where  $C_1$  depends on the bound of  $\ell$ . Since  $\mathcal{L}_{\text{Poisson}}$  and  $\mathcal{L}_{\text{Vlasov}}$  approximate the squared errors (small residuals imply  $\phi_\theta \approx \phi^*, F_\theta \approx F^*$ ), and  $\mathcal{L}_{\text{disp}}$  adds regularization, the training loss  $\mathcal{L}$  bounds  $\ell$ . Thus, the empirical training error  $\hat{\mathcal{E}}_{\text{train}}(\theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\mathbf{z}_i, \mathbf{v}_i; \phi_\theta, F_\theta)$  satisfies:

$$\hat{\mathcal{E}}_{\text{gen}}(\theta) \leq \hat{\mathcal{E}}_{\text{train}}(\theta) + C_2, \quad (122)$$

where  $C_2$  accounts for differences between  $\ell$  and  $\mathcal{L}$ . Taking expectations:

$$\mathcal{E}_{\text{gen}}(\theta) \leq \mathcal{E}_{\text{train}}(\theta) + 2\text{Rad}(\mathcal{L}_\mathcal{H}) + C_1 \sqrt{\frac{\log(1/\delta)}{n}} + C_2. \quad (123)$$

Since  $\text{Rad}(\mathcal{L}_\mathcal{H})$  depends on  $\mathcal{H}$ , we denote  $\text{Rad}(\mathcal{H}) = \text{Rad}(\mathcal{L}_\mathcal{H})$ , and absorb  $C_2$  into the  $O(\sqrt{\log(1/\delta)/n})$  term, yielding:

$$\mathcal{E}_{\text{gen}}(\theta) \leq \mathcal{E}_{\text{train}}(\theta) + 2\text{Rad}(\mathcal{H}) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \quad (124)$$

Now, we show that  $\mathcal{L}_{\text{disp}}$  reduces  $\text{Rad}(\mathcal{H})$ . The Rademacher complexity is bounded by the complexity of  $\mathcal{H}$ :

$$\text{Rad}(\mathcal{H}) \leq C_3 \left( \sup_{\phi_\theta} \|\phi_\theta\|_{H^s} + \sup_{F_\theta} \|F_\theta\|_{H^s} \right) / \sqrt{n}. \quad (125)$$

By Theorem 2.5,  $\mathcal{L}_{\text{disp}}$  satisfies:

$$\mathbb{E}_{\mathbf{z}, \mathbf{k}}[\mathcal{L}_{\text{disp}}] \geq c \|\phi_\theta\|_{H^2}^2 - C. \quad (126)$$

Minimizing  $\mathcal{L}_{\text{disp}}$  constrains:

$$\|\phi_\theta\|_{H^2}^2 \leq \frac{1}{c} \mathbb{E}_{\mathbf{z}, \mathbf{k}}[\mathcal{L}_{\text{disp}}] + C. \quad (127)$$

Define the subset  $\mathcal{H}_\phi = \{\phi_\theta \in H^2 : \mathbb{E}_{\mathbf{z}, \mathbf{k}}[\mathcal{L}_{\text{disp}}] \leq M\}$  for some bound  $M$ . Then:

$$\|\phi_\theta\|_{H^2} \leq \sqrt{\frac{M + C}{c}}, \quad (128)$$

restricting  $\phi_\theta$  to a bounded subset of  $H^2$ . Since  $H^s \subset H^2$  for  $s \geq 2$ , this reduces  $\sup_{\phi_\theta} \|\phi_\theta\|_{H^s}$ . For  $F_\theta = A \exp(-\frac{1}{2}m\|\mathbf{v}\|^2 - \phi_\theta(\mathbf{z}))$ , the norm  $\|F_\theta\|_{H^s}$  depends on  $\|\phi_\theta\|_{H^s}$ , so constraining  $\phi_\theta$  also bounds  $F_\theta$ . Thus,  $\mathcal{L}_{\text{disp}}$  reduces  $\text{Rad}(\mathcal{H})$ , tightening the generalization bound, completing the proof.

## C Algorithm Protocol

Algo. 1 and Algo. 2 give the algorithmic protocol of our framework, which is easy to implement and applicable to common OOD problems.

---

### Algorithm 1 Algorithm pseudocode of *CBD-De*

---

**Require:** Training mini-batches  $\{x_i, y_i\}_{i=1}^B$ , pre-trained feature extractor  $\hat{f}_\theta$ , distribution prediction head  $F_\theta$ , potential prediction head  $\phi_\theta$ , learning rate  $\eta$

**Ensure:** Trained parameters  $\theta$

```

1: Initialize  $\theta \leftarrow \{F_\theta, \phi_\theta\}$ , optimizer  $\mathcal{O}(\theta, \eta)$  /* Fixed  $\hat{f}_\theta$ , post-hoc fashion */
2: for each training step do
3:    $\mathbf{x} \leftarrow \{x_i\}_{i=1}^n$ ,  $\mathbf{y} \leftarrow \{y_i\}_{i=1}^n$ 
4:    $\mathbf{z} \leftarrow \hat{f}_\theta(\mathbf{x})$  /* Extract features */
5:    $\mathbf{v} \leftarrow \text{ones\_like}(\mathbf{z})$  /* Constant velocity input */
6:   Vlasov residual:  $\mathbf{v} \cdot \nabla_{\mathbf{x}} F^* - \nabla \phi^* \cdot \nabla_{\mathbf{v}} F^*$ 
7:    $F^* \leftarrow F_\theta(\mathbf{z}, \mathbf{v})$ 
8:    $\phi^* \leftarrow \phi_\theta(\mathbf{z})$ 
9:    $\nabla_{\mathbf{z}} \phi^* \leftarrow \text{autograd}(\phi^*. \text{sum}(), \mathbf{z})$ 
10:   $\mathbf{E}^* \leftarrow -\nabla_{\mathbf{z}} \phi^*$ 
11:   $\nabla_{\mathbf{x}} F^*, \nabla_{\mathbf{v}} F^* \leftarrow \text{autograd}(F^*. \text{sum}(), [\mathbf{z}, \mathbf{v}])$ 
12:   $\mathcal{L}_{\text{vlasov}} \leftarrow \text{mean}[(\mathbf{v} \cdot \nabla_{\mathbf{x}} F^* + \mathbf{E}^* \cdot \nabla_{\mathbf{v}} F^*)^2]$ 
13:  Poisson residual:  $\nabla^2 \phi^* + \exp(-\phi^*)$ 
14:   $\nabla_{\mathbf{z}} \phi^* \leftarrow \text{autograd}(\phi^*. \text{sum}(), \mathbf{z})$ 
15:  /* Compute second-order derivatives on each dim */
16:  for  $d = 1$  to  $Z$  do
17:     $\nabla \phi_d^* \leftarrow \partial \phi^* / \partial z_d \leftarrow \text{autograd}(\phi^*. \text{sum}(), z_d)$ 
18:     $\partial^2 \phi^* / \partial z_d^2 \leftarrow \text{autograd}(\nabla \phi_d^*. \text{sum}(), z_d)$ 
19:    Accumulate  $\nabla^2 \phi^* \leftarrow \nabla^2 \phi^* + \partial^2 \phi^* / \partial z_d^2$ 
20:  end for
21:   $\mathcal{L}_{\text{poisson}} \leftarrow \text{mean}[(\nabla^2 \phi^* + \exp(-\phi^*))^2]$ 
22:  Total loss and backpropagation
23:   $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{vlasov}} + \mathcal{L}_{\text{poisson}}$ 
24:   $\mathcal{L}_{\text{total}}. \text{backward}()$ 
25:   $\mathcal{O}. \text{step}()$ 
26:   $\mathcal{O}. \text{zero\_grad}()$ 
27: end for

```

---

---

**Algorithm 2** Algorithm pseudocode of *CBD-Gen*


---

**Require:** Training mini-batches  $\{x_i, y_i\}_{i=1}^B$ , feature extractor  $\hat{f}_\theta$ , classifier head  $f_\theta$ , distribution prediction head  $F_\theta$ , potential prediction head  $\phi_\theta$ , learning rate  $\eta$ , loss weight  $\alpha, \beta$ , plane waves number  $N$ , scaling factor  $\sigma$

**Ensure:** Trained parameters  $\theta$

```

1: Initialize  $\theta \leftarrow \{\hat{f}_\theta, f_\theta, F_\theta, \phi_\theta\}$ , optimizer  $\mathcal{O}(\theta, \eta)$ 
2: for each training step do
3:    $\mathbf{x} \leftarrow \{x_i\}_{i=1}^n, \quad \mathbf{y} \leftarrow \{y_i\}_{i=1}^n$ 
4:    $\mathbf{z} \leftarrow \hat{f}_\theta(\mathbf{x})$  /* Extract features */
5:    $\mathbf{v} \leftarrow \text{ones\_like}(\mathbf{z})$  /* Constant velocity input */
6:   Vlasov residual:  $\mathbf{v} \cdot \nabla_{\mathbf{x}} F^* - \nabla \phi^* \cdot \nabla_{\mathbf{v}} F^*$ 
7:    $F^* \leftarrow F_\theta(\mathbf{z}, \mathbf{v})$ 
8:    $\phi^* \leftarrow \phi_\theta(\mathbf{z})$ 
9:    $\nabla_{\mathbf{z}} \phi^* \leftarrow \text{autograd}(\phi^*.sum(), \mathbf{z})$ 
10:   $\mathbf{E}^* \leftarrow -\nabla_{\mathbf{z}} \phi^*$ 
11:   $\nabla_{\mathbf{x}} F^*, \nabla_{\mathbf{v}} F^* \leftarrow \text{autograd}(F^*.sum(), [\mathbf{z}, \mathbf{v}])$ 
12:   $\mathcal{L}_{\text{vlasov}} \leftarrow \text{mean}[(\mathbf{v} \cdot \nabla_{\mathbf{x}} F^* + \mathbf{E}^* \cdot \nabla_{\mathbf{v}} F^*)^2]$ 
13:  Poisson residual:  $\nabla^2 \phi^* + \exp(-\phi^*)$ 
14:   $\nabla_{\mathbf{z}} \phi^* \leftarrow \text{autograd}(\phi^*.sum(), \mathbf{z})$ 
15:  /* Compute second-order derivatives on each dim */
16:  for  $d = 1$  to  $Z$  do
17:     $\nabla \phi_d^* \leftarrow \partial \phi^* / \partial z_d \leftarrow \text{autograd}(\phi^*.sum(), z_d)$ 
18:     $\partial^2 \phi^* / \partial z_d^2 \leftarrow \text{autograd}(\nabla \phi_d^*.sum(), z_d)$ 
19:    Accumulate  $\nabla^2 \phi^* \leftarrow \nabla^2 \phi^* + \partial^2 \phi^* / \partial z_d^2$ 
20:  end for
21:   $\mathcal{L}_{\text{poisson}} \leftarrow \text{mean}[(\nabla^2 \phi^* + \exp(-\phi^*))^2]$ 
22:  Dispersion relation loss:  $\sum_{\mathbf{k}} \|\|\mathbf{k}\|^2 \phi'_k + \rho'_k\|^2$ 
23:  Sample  $\{\mathbf{k}_i\}_{i=1}^N \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ 
24:  for each  $\mathbf{k}_i$  do
25:     $\sigma_i \leftarrow \exp(-j \cdot \mathbf{k}_i^\top \mathbf{z})$ 
26:     $\phi_{k_i}^* \leftarrow \text{mean}(\phi^* \cdot \sigma_i), \quad \rho_{k_i} \leftarrow \text{mean}(\exp(-\phi^*) \cdot \sigma_i)$ 
27:     $r_i \leftarrow \|\mathbf{k}_i\|^2 \cdot \phi_{k_i}^* + \rho_{k_i}$ 
28:  end for
29:   $\mathcal{L}_{\text{disp}} \leftarrow \text{sum}(|r_i|^2)$ 
30:  Classification loss
31:   $\mathbf{y}_{\text{pred}} \leftarrow f_\theta(\mathbf{z})$ 
32:   $\mathcal{L}_{\text{classify}} \leftarrow \text{CrossEntropy}(\mathbf{y}_{\text{pred}}, \mathbf{y})$ 
33:  Total loss and backpropagation
34:   $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{classify}} + \alpha(\mathcal{L}_{\text{vlasov}} + \mathcal{L}_{\text{poisson}}) + \beta \mathcal{L}_{\text{disp}}$ 
35:   $\mathcal{L}_{\text{total}}.\text{backward}()$ 
36:   $\mathcal{O}.\text{step}()$ 
37:   $\mathcal{O}.\text{zero\_grad}()$ 
38: end for

```

---

## D Additional Experiment Settings and Details

We conduct experiments following the latest version of OpenOOD<sup>3</sup> [96, 94] and the DomainBed<sup>4</sup> [23]. In this section, we first provide more details for the utilized baselines (Section D.1), datasets and evaluation protocol (Section D.2), and model architectures (Section D.3). Then, we demonstrate the hyperparameters (Section D.4) and the experiment infrastructures (Section D.5). Finally, we include additional ablation and validation results to better understand the proposed method, including the effect of the initial velocity assumption (Section D.6), sensitivity analysis of the MLP architecture (Section D.7), and the validation of the MLP steady-state approximation (Section D.8).

### D.1 Baselines

**OOD Detection.** To evaluate the effectiveness of *CBD-De* in detecting OOD inputs, we conduct extensive comparisons with 22 representative post-hoc OOD detection methods across three widely adopted benchmarks: CIFAR-10, CIFAR-100, and ImageNet-1k. These baselines span a range of methodological categories, including confidence-based methods such as OpenMax [6] and MSP [28]; calibration and scaling techniques like TempScale [24] and ODIN [47]; distance-based and density-aware approaches such as MDS [43], MDSEns [43], RMDS [62], and KNN [77]; feature-based scoring methods including Gram [66], ReAct [75], RankFeat [73], and VIM [85]; energy- and activation-based models such as EBO [50], MLS [26], KLM [26], and ASH [14]; generative and hybrid models like OpenGAN [37], SHE [97], and GEN [51]; as well as recent advances in neuron-level uncertainty estimation such as GradNorm [32], DICE [76], and NAC-UE [52]. All evaluation results reported in Tables 13, 14, and 15 are derived from the standardized OpenOOD implementations to ensure fair and reproducible comparisons.

**OOD Generalization.** We evaluate the OOD generalization capability of *CBD-Gen* by comparing it with 17 representative baselines spanning diverse algorithmic categories, including distribution matching (MMD [46]), distributionally robust optimization (GroupDRO [65], VREx [39], IRM [2]), data augmentation (MixStyle [103], ARM [98], Mixup [91]), representation learning (SagNet [58], MTL [7], MLDG [44]), alignment-based methods (CORAL [74]), empirical risk minimization (ERM [80]), feature suppression and invariance strategies (RSC [34]), as well as sharpness-aware optimization techniques (SAM [68], GSAM [104], SAGM [86]) and gradient-guided approaches (GGA [3]). Comprehensive experimental results across five widely used OOD generalization benchmarks (PACS, VLCS, OfficeHome, TerraInc, and DomainNet) are summarized in Tables 16 to 20.

**Comparison to Traditional Statistical Methods.** Beyond empirical performance, it is essential to contextualize our proposed method within the broader landscape of traditional statistical approaches for OOD learning. While most conventional methods (*e.g.*, energy-based modeling, density estimation, or confidence scoring) focus on point-wise statistics or local sample relationships, they typically overlook the structured interactions among features or samples. In contrast, our Vlasov–Poisson formulation provides a principled, physically interpretable mechanism that inherently captures *global* inter-feature dynamics through self-consistent field evolution. This collective interaction allows our model to reflect structural dependencies that classical scoring-based or probabilistic models cannot explicitly encode. Table 5 summarizes this conceptual distinction. We compare our method with several representative approaches across multiple dimensions, including their modeling assumptions, underlying OOD mechanisms, sample interaction properties, and interpretability.

### D.2 Benchmarks

**OOD Detection.** We primarily evaluate *CBD-De* using the Far-OOD track from the OpenOOD benchmark suite, which offers a standardized and well-established evaluation protocol adopted by numerous prior works.

CIFAR-10 and CIFAR-100, consisting of 10 and 100 classes respectively, serve as the in-distribution (InD) datasets in our experiments. Following the official OpenOOD setup, we adopt a consistent data split for both benchmarks. Specifically, we use the full training set (50,000 images) for model training. From the official test set, 1,000 samples are reserved as the InD validation set, while the remaining 9,000 images are used as the InD test set. For OOD validation, we follow the OpenOOD protocol by

<sup>3</sup><https://github.com/Jingkang50/OpenOOD>.

<sup>4</sup><https://github.com/facebookresearch/DomainBed>.

Table 5: Comparison between our *CBD*-base method and traditional statistical OOD Approaches.

| Approaches                   | Modeling Assumption                | OOD Mechanism                                        | Sample Interactions                         | Key Formula                                                                                            | Interpretability                              |
|------------------------------|------------------------------------|------------------------------------------------------|---------------------------------------------|--------------------------------------------------------------------------------------------------------|-----------------------------------------------|
| Energy-based Models          | Energy functions in feature space  | Low energy $\rightarrow$ InD, high $\rightarrow$ OOD | [X] No interactions                         | $\mathcal{E}(x) = -\log p_\theta(x)$                                                                   | Intuitive locally, hard globally              |
| Density Estimation           | Probability density modeling       | Low density $\rightarrow$ OOD                        | $[\Delta]$ Local interactions via density   | $p(x) = \prod_i p(x_i)$                                                                                | Captures local trends, lacks high-dim clarity |
| Confidence Scoring           | Model output confidence            | Low confidence $\rightarrow$ OOD                     | [X] No interactions                         | $C(x) = \text{softmax}(f(x))$                                                                          | Heuristic, limited by generalization          |
| <b>Ours (Vlasov-Poisson)</b> | Particles in self-consistent field | Cannot reach steady state                            | $[\checkmark]$ Global explicit interactions | $\frac{\partial F}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{z}} F + \nabla \cdot \mathbf{E} = 0$ | Physics-based via particle dynamics           |

selecting 1,000 images from Tiny ImageNet [41], evenly sampled across 20 unseen categories. This ensures a controlled setting for hyperparameter tuning and model calibration. To comprehensively assess the generalization ability of OOD detection methods, we evaluate performance on four widely used OOD test datasets—each disjoint from both the InD and OOD validation sets. These datasets include:

- MNIST [13] is a benchmark dataset for handwritten digit classification, consisting of 70,000 grayscale images of digits (60,000 for training and 10,000 for test), each sized 28×28 pixels. We utilize the entire test set for OOD detection.
- SVHN [59] (Street View House Numbers) contains over 600,000 color images of digits captured from real-world house numbers in Google Street View. Each image is 32×32 pixels and includes significant background clutter, lighting variation, and perspective distortion. We utilize the entire test set (26,032 images) for OOD detection.
- Textures [10] is a collection of texture-centric images designed for texture recognition and segmentation tasks. The dataset comprises 5,640 images from 47 texture categories, such as "bark", "brick", or "bubble", with each image exhibiting rich local patterns and high-frequency variations. We employ the entire dataset.
- Places365 [101] is a large-scale scene recognition dataset comprising over 1.8 million images across 365 diverse scene categories. The standard test set includes 900 images per category. For OOD detection, we adopt the entire test set after removing 1,305 images that exhibit semantic overlap with in-distribution classes, following the filtering protocol established in prior works.

To evaluate our method under large-scale and high-resolution conditions, we adopt ImageNet-1K [12] as the InD dataset. This dataset comprises approximately 1.2 million training images across 1,000 diverse object categories. Following the evaluation protocol established by OpenOOD, we use 45,000 images from the official ImageNet validation set as the InD test set, while the remaining 5,000 images are held out as the InD validation set. To facilitate hyperparameter tuning and avoid information leakage from the test sets, we construct an OOD validation set using 1,763 images from OpenImage-O [84], a curated subset of OpenImages whose categories are explicitly disjoint from ImageNet classes. For final evaluation, we benchmark our method against three widely used OOD test datasets that cover diverse semantic domains and visual characteristics:

- iNaturalist [79] is a large-scale, fine-grained dataset containing 859,000 images of over 5,000 natural species. Captured in unconstrained environments, the images exhibit complex backgrounds, lighting variation, and high intra-class diversity, making the dataset visually and semantically distinct from object-centric benchmarks like ImageNet. Following prior works[31, 94], we use a subset of 10,000 images from 110 non-overlapping classes for OOD evaluation.
- Textures [10] comprises 5,640 real-world texture images spanning 47 categories. For evaluation, we employ the full dataset without further filtering.
- OpenImage-O [84] is a curated subset of the OpenImages-v3 dataset. It comprises 17,632 images that are carefully selected to be semantically disjoint from ImageNet-1K classes. To ensure a rigorous evaluation setting, the subset underwent manual filtering to eliminate any

category overlap. Following the OpenOOD protocol, we utilize the entire dataset for OOD testing, excluding the samples reserved for OOD validation.

**OOD Generalization.** We adopt the training and evaluation protocol established in DomainBed [23], ensuring consistency in dataset splits, training schedules, and model selection criteria. Specifically, we follow the leave-one-domain-out evaluation strategy, where the model is trained on all but one domain and evaluated on the held-out domain to assess its generalization ability to unseen distributions. This process is repeated for each domain, treating it as the target in turn. Model selection is performed using the average accuracy across all validation sets corresponding to the training domains, ensuring that the chosen model is not tuned on the target domain. We conduct experiments on five widely adopted domain generalization benchmarks included in DomainBed:

- PACS [45] consists of four visually distinct domains: Photo (P), Art Painting (A), Cartoon (C), and Sketch (S). It contains a total of 9,991 images spanning seven object categories common across all domains. The dataset presents significant domain shift, especially between natural (Photo) and abstract (Sketch) styles, challenging models to capture domain-invariant features.
- VLCS [16] combines images from four datasets: PASCAL VOC2007 (V), LabelMe (L), Caltech-101 (C), and SUN09 (S). It contains five shared object classes and a total of 10,729 images. The main source of domain shift lies in differences in collection environments, image resolution, and labeling conventions across the constituent datasets.
- OfficeHome [82] is a more challenging benchmark with 15,588 images across four domains: Art (A), Clipart (C), Product (P), and Real World (R). It spans 65 object categories related to office and home environments. The dataset is characterized by high intra-class variation and large appearance gaps between stylized (Clipart, Art) and natural (Real World) domains.
- TerraInc [5] is a wildlife image dataset collected from camera traps in different geographical locations. It includes 24,788 images across four domains (locations): L100, L38, L43, and L46, and contains ten animal species. The domain shift stems from environmental variation (*e.g.*, lighting, terrain, vegetation) and class imbalance.
- DomainNet [60] is the largest and most diverse benchmark in the domain generalization literature, comprising approximately 600,000 images across six distinct domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. It covers 345 object categories shared among all domains. The dataset exhibits a wide range of domain shifts, including differences in abstraction level, drawing style, color richness, and semantic representation. Among the domains, Quickdraw is particularly challenging due to its highly simplified and noisy stroke-based illustrations.

### D.3 Model Architecture

For OOD detection, we adopt a post-hoc setting by freezing the classification network and training only the lightweight distribution prediction head  $F_\theta$  and potential prediction head  $\phi_\theta$ . In contrast, for OOD generalization, all model parameters are optimized jointly.

#### D.3.1 OOD Detection

For the CIFAR-10 and CIFAR-100 benchmarks, we adopt the ResNet-18 architecture [25]. In accordance with the standardized protocol in OpenOOD, we train each model for 100 epochs using the official training splits. To account for potential variability across training stages, we evaluate OOD detection performance at three checkpoints saved throughout the training process. This setting ensures fair and consistent comparisons with prior post-hoc methods. Additional training configurations (*e.g.*, learning rate, optimizer, data augmentation) follow the default settings provided in the OpenOOD.

For the large-scale ImageNet-1k benchmark, we evaluate OOD detection using two representative model architectures that reflect both convolutional and transformer-based paradigms: *i)* ResNet-50 [25], a 50-layer residual network pretrained on ImageNet-1K, is employed as a high-capacity convolutional backbone. During inference, all test images are resized to 224×224 pixels to match the model’s input resolution. We use the official pretrained weights released by PyTorch to ensure reproducibility. *ii)* ViT-b16 [15] serves as a transformer-based alternative, also pretrained



on ImageNet-1K and evaluated with the same input resolution of 224×224 pixels. The ViT model introduces a fundamentally different inductive bias compared to CNNs, allowing us to examine the generality of our method across architectural paradigms. We employ the official PyTorch checkpoints for all ViT experiments.

**Distribution prediction head  $F_\theta$  and potential prediction head  $\phi_\theta$ :** For the CIFAR-10 and CIFAR-100 benchmarks, both the distribution prediction head  $F_\theta$  and the potential prediction head  $\phi_\theta$  are implemented as two-layer MLPs with an input dimension of 512, a hidden dimension of 512, and an output dimension of 1. For the large-scale ImageNet-1k benchmark, the hidden dimension is set to 512 when using ResNet-50 as the backbone, and 1,536 when using ViT-b16. All heads use the SiLU activation function. Detailed configurations are provided in Table 6.

Table 6: Details of distribution prediction head and potential prediction head.

|              | Backbone  | Input Dim. | Hiddle Dim. | Output Dim. | #Head Param. | Activation |
|--------------|-----------|------------|-------------|-------------|--------------|------------|
| CIFAR-10/100 | ResNet-18 | 512        | 512         | 1           | 0.53M        | SiLU       |
| ImageNet-1k  | ResNet-50 | 2048       | 512         | 1           | 2.10M        | SiLU       |
|              | ViT-b16   | 768        | 1536        | 1           | 2.36M        | SiLU       |

### D.3.2 OOD Generalization

For all benchmarks, we adopt ResNet-50 pretrained on ImageNet [12] as the default backbone and initialization. Both the distribution prediction head  $F_\theta$  and the potential prediction head  $\phi_\theta$  are implemented as two-layer MLPs with a hidden dimension of 512 (same as Table 6). Although these auxiliary branches introduce a small number of additional parameters during training (approximately 8% of the total), they serve solely to guide the backbone toward learning more robust representations via explicit supervision. Importantly, these branches can be safely removed at inference time, preserving the backbone’s structure and incurring no extra cost in prediction. Our approach maintains a clear separation between the auxiliary supervision and the primary prediction pathway, ensuring that the backbone remains unaltered throughout both training and inference. By decoupling training-time supervision from inference-time computation, it achieves an efficient balance between representational robustness and deployment practicality, enabling scalable application across diverse generalization scenarios.

## D.4 Hyperparameters

**Shared Hyperparameters.** For the constant  $\epsilon_0$  in the loss function, we uniformly assign the value 1. For each  $\mathbf{z}$ , the initial velocity  $\mathbf{v}_0$  of the corresponding  $\mathbf{v}$  is set to a vector of ones in  $\mathbb{R}^Z$ .

**OOD Detection.** We train the distribution prediction head  $F_\theta$  and the potential prediction head  $\phi_\theta$  using the Adam optimizer with a learning rate of  $5 \times 10^{-4}$  and a batch size of 128, adhering to standard post-hoc configurations. Training proceeds for 50 epochs.

**OOD Generalization.** In our experiments, the wavevector  $\mathbf{k}$  in the dispersion relation loss  $\mathcal{L}_{\text{disp}}$  is randomly sampled from a Gaussian distribution per batch, with a scaling factor of 0.5 and a wavenumber  $N$  of 250. The loss weights  $\alpha$  and  $\beta$  are both set to 0.01. Additional hyperparameters, including learning rate, weight decay, and dropout rate, are tuned following [9] and detailed in Table 3. We adopt early stopping and utilize the Adam optimizer. Consistent with DomainBed, the batch size is set to 32 for all datasets, except for DomainNet, which uses a batch size of 24.

Table 7: Hyperparameters for OOD generalization experiments.

| Hyperparameter | PACS | VLCS | OfficeHome | TerraInc | DomainNet |
|----------------|------|------|------------|----------|-----------|
| Learning rate  | 3e-5 | 1e-5 | 1e-5       | 1e-5     | 3e-5      |
| Dropout        | 0.5  | 0.5  | 0.5        | 0.5      | 0.5       |
| Weight decay   | 1e-4 | 1e-4 | 1e-4       | 1e-4     | 1e-4      |
| Training Steps | 5000 | 5000 | 5000       | 5000     | 15000     |

## D.5 Experiment Infrastructure

Our experiments are conducted using a combination of NVIDIA GeForce RTX 3090 Ti and NVIDIA A100 GPUs. Specifically, we utilized four 3090 Ti GPUs and one A100 GPU with 40GB of memory. The OOD detection experiments are performed exclusively on the 3090 Ti GPUs, while the OOD generalization experiments were carried out on both the 3090 Ti and A100 GPUs.

## D.6 Ablation on Initial Velocity Assumption

To evaluate the robustness of our uniform initial velocity assumption, we conducted additional experiments using different initialization strategies: *i)* uniform vectors with all elements set to 1, *ii)* uniform vectors with values of 0.5 matching the feature dimension, and *iii)* random vectors of the same dimension. As shown in Tables 8 and 9, random initialization slightly underperforms the uniform settings, but the performance gaps are minimal, demonstrating the robustness of our approach. The uniform initialization assumption thus maintains effectiveness while simplifying computation.

Table 8: OOD Detection results on CIFAR-100 with different initial velocity settings.

| Initial Velocity | FPR95 ↓ | AUROC ↑ |
|------------------|---------|---------|
| 1                | 34.81   | 90.90   |
| 0.5              | 35.20   | 90.47   |
| Random           | 38.29   | 89.52   |

Table 9: OOD Generalization results on PACS with different initial velocity settings.

| Initial Velocity | Accuracy ↑ |
|------------------|------------|
| 1                | 87.7       |
| 0.5              | 87.7       |
| Random           | 87.4       |

## D.7 Sensitivity Analysis of MLP Architecture

We further investigate the sensitivity of our method to key architectural hyperparameters of the MLP, which serves as the steady-state predictor. Specifically, we examine the impact of layer depth and hidden dimension on OOD detection performance using CIFAR-100 as the InD dataset. Table 10 presents results for varying layer depths with a fixed hidden dimension of 512, while Table 11 shows the effect of different hidden dimensions using a fixed depth of 2 layers. The results demonstrate that performance scales favorably with the number of MLP parameters. Notably, increasing the layer depth from 1 to 2 yields substantial improvements, with AUROC increasing from 87.14% to 90.90% and FPR95 decreasing from 39.30 to 34.81. Similarly, expanding the hidden dimension from 128 to 512 progressively enhances performance. However, returns diminish beyond a certain threshold (*e.g.*, hidden dimension of 512 vs. 1024, with AUROC of 90.90% vs. 90.83%), highlighting the importance of balancing model capacity with computational efficiency. These insights inform practical hyperparameter selection for deploying *CBD-De* in real-world applications.

Table 10: OOD detection on CIFAR-100 with MLP hidden dimension of 512.

| Layer Depth | FPR95 ↓      | AUROC ↑      |
|-------------|--------------|--------------|
| 1           | 39.30        | 87.14        |
| 2           | <b>34.81</b> | <b>90.90</b> |

Table 11: OOD detection on CIFAR-100 with MLP depth of 2 layers.

| Hidden Dimension | FPR95 ↓      | AUROC ↑      |
|------------------|--------------|--------------|
| 128              | 40.35        | 86.93        |
| 256              | 38.52        | 88.20        |
| 512              | <b>34.81</b> | <b>90.90</b> |
| 1024             | 35.25        | 90.83        |

### D.8 Validation of the MLP Steady-State Approximation

Our formulation of steady-state prediction is originally defined through PDE. To make the solution computationally tractable, we replace the numerical PDE integration with an MLP-based steady-state predictor that directly approximates the stationary solution in a single forward pass. To assess the fidelity of this approximation, we conduct experiments using a numerical PDE solver implemented with DeepXDE [53], which follows the same architecture and parameter settings (2 layers, 512 hidden dimensions). The PDE solver is executed with 10 and 30 integration steps to estimate the steady-state distribution. Table 12 presents the OOD detection results on CIFAR-10 and CIFAR-100 datasets. For simpler datasets (CIFAR-10), performance differences are negligible (0.02-0.06 AUROC), while for more complex datasets (CIFAR-100), numerical integration shows moderate improvements (0.30-0.42 AUROC). However, these gains require 7-23 $\times$  longer inference time and substantial memory overhead for intermediate computations. Given the minimal accuracy differences on simpler distributions and the significant computational advantages, our direct prediction approach offers a practical solution for efficient OOD detection in real-world applications.

Table 12: Comparison of the MLP steady-state approximation against a numerical PDE solver implemented with DeepXDE [53]. AUROC and relative inference time are reported.

| Dataset                | Sample Size | Classes | MLP AUROC ↑ | PDE Solver AUROC ↑ | Forward Pass Time ↓ |
|------------------------|-------------|---------|-------------|--------------------|---------------------|
| <i>Time Steps = 10</i> |             |         |             |                    |                     |
| CIFAR-10               | 50,000      | 10      | 96.39       | 96.41 (+0.02)      | 7 $\times$ slower   |
| CIFAR-100              | 50,000      | 100     | 90.90       | 91.20 (+0.30)      | 7 $\times$ slower   |
| <i>Time Steps = 30</i> |             |         |             |                    |                     |
| CIFAR-10               | 50,000      | 10      | 96.39       | 96.45 (+0.06)      | 23 $\times$ slower  |
| CIFAR-100              | 50,000      | 100     | 90.90       | 91.32 (+0.42)      | 23 $\times$ slower  |

## E Full OOD Detetion Results

Table 13: OOD detection performance on the CIFAR-10 benchmark. Results are formatted as **first**, second, and third best. Following the OpenOOD protocol, we report the average performance over three ResNet-18 checkpoints, each trained exclusively on the InD dataset.

| Method                    | MINIST             |                    |  | SVHN               |                    |  | Textures           |                    |  | Places365          |                    |  | Average            |                    |
|---------------------------|--------------------|--------------------|--|--------------------|--------------------|--|--------------------|--------------------|--|--------------------|--------------------|--|--------------------|--------------------|
|                           | FPR95↓             | AUROC↑             |  | FPR95↓             | AUROC↑             |  | FPR95↓             | AUROC↑             |  | FPR95↓             | AUROC↑             |  | FPR95↓             | AUROC↑             |
| <i>CIFAR-10 Benchmark</i> |                    |                    |  |                    |                    |  |                    |                    |  |                    |                    |  |                    |                    |
| OpenMax                   | 23.33±4.67         | 90.50±0.44         |  | 25.40±1.47         | 89.77±0.45         |  | 31.50±4.05         | 89.58±0.60         |  | 38.52±2.27         | 88.63±0.28         |  | 29.69±1.21         | 89.62±0.19         |
| MSP                       | 23.64±5.81         | 92.63±1.57         |  | 25.82±1.64         | 91.46±0.40         |  | 34.96±4.64         | 89.89±0.71         |  | 42.47±3.81         | 88.92±0.47         |  | 31.72±1.84         | 90.73±0.43         |
| TempScale                 | 23.53±7.05         | 93.11±1.77         |  | 26.97±2.65         | 91.66±0.52         |  | 38.16±5.89         | 90.01±0.74         |  | 45.27±4.50         | 89.11±0.52         |  | 33.48±2.39         | 90.97±0.52         |
| ODIN                      | 23.83±12.34        | 95.24±1.96         |  | 68.61±0.52         | 84.58±0.77         |  | 67.70±11.06        | 86.94±2.26         |  | 70.36±6.96         | 85.07±1.24         |  | 57.62±4.24         | 87.96±0.61         |
| MDS                       | 27.30±3.55         | 90.10±2.41         |  | 25.96±2.52         | 91.18±0.47         |  | 27.94±4.20         | 92.69±1.06         |  | 47.67±4.54         | 84.90±2.54         |  | 32.22±3.40         | 89.72±1.36         |
| MDSEns                    | <b>1.30</b> ±0.51  | <b>99.17</b> ±0.41 |  | 74.34±1.04         | 66.56±0.58         |  | 76.07±0.17         | 77.40±0.28         |  | 94.16±0.33         | 52.47±0.15         |  | 61.47±0.48         | 73.90±0.27         |
| RMDS                      | 21.49±2.32         | 93.22±0.80         |  | 23.46±1.48         | 91.84±0.26         |  | 25.25±0.53         | 92.23±0.23         |  | 31.20±0.28         | 91.51±0.11         |  | 25.35±0.73         | 92.20±0.21         |
| Gram                      | 70.30±8.96         | 72.64±2.34         |  | 33.91±17.35        | 91.52±4.45         |  | 94.64±2.71         | 62.34±8.27         |  | 90.49±1.93         | 60.44±3.41         |  | 72.34±6.73         | 71.73±3.20         |
| EBO                       | 24.99±12.93        | 94.32±2.53         |  | 35.12±6.11         | 91.79±0.98         |  | 51.82±6.11         | 89.47±0.70         |  | 54.85±6.52         | 89.25±0.78         |  | 41.69±5.32         | 91.21±0.92         |
| OpenGAN                   | 79.54±19.71        | 56.14±24.08        |  | 75.27±26.93        | 52.81±27.60        |  | 83.95±14.89        | 56.14±18.26        |  | 95.32±4.45         | 53.34±5.79         |  | 83.52±11.63        | 54.61±15.51        |
| GradNorm                  | 85.41±4.85         | 63.72±7.37         |  | 91.65±2.42         | 53.91±6.36         |  | 98.09±0.49         | 52.07±4.09         |  | 92.46±2.28         | 60.50±5.33         |  | 91.90±2.23         | 57.55±3.22         |
| ReAct                     | 33.77±18.00        | 92.81±3.03         |  | 50.23±15.98        | 89.12±3.19         |  | 51.42±11.42        | 89.38±1.49         |  | 44.20±3.35         | 90.35±0.78         |  | 44.90±8.37         | 90.42±1.41         |
| MLS                       | 25.06±12.87        | 94.15±2.48         |  | 35.09±6.09         | 91.69±0.94         |  | 51.73±6.13         | 89.41±0.71         |  | 54.84±6.51         | 89.14±0.76         |  | 41.68±5.27         | 91.10±0.89         |
| KLM                       | 76.22±12.09        | 85.00±2.04         |  | 59.47±7.06         | 84.99±1.18         |  | 81.95±9.95         | 82.35±0.33         |  | 95.58±2.12         | 78.37±0.33         |  | 78.31±4.84         | 82.68±0.21         |
| VIM                       | 18.36±1.42         | 94.76±0.38         |  | 19.29±0.41         | 94.50±0.48         |  | 21.14±1.83         | 95.15±0.34         |  | 41.43±2.17         | 89.49±0.39         |  | 25.05±0.52         | 93.48±0.24         |
| KNN                       | 20.05±1.36         | 94.26±0.38         |  | 22.60±1.26         | 92.67±0.30         |  | 24.06±0.55         | 93.16±0.24         |  | 30.38±0.63         | 91.77±0.23         |  | 24.27±0.40         | 92.96±0.14         |
| DICE                      | 30.83±10.54        | 90.37±5.97         |  | 36.61±4.74         | 90.02±1.77         |  | 62.42±4.79         | 81.86±2.35         |  | 77.19±12.60        | 74.67±4.98         |  | 51.76±4.42         | 84.23±1.89         |
| RankFeat                  | 61.86±12.78        | 75.87±5.22         |  | 64.49±7.38         | 68.15±7.44         |  | 59.71±9.79         | 73.46±6.49         |  | 43.70±7.39         | 85.99±3.04         |  | 57.44±7.99         | 75.87±5.06         |
| ASH                       | 70.00±10.56        | 83.16±4.66         |  | 83.64±6.48         | 73.46±6.41         |  | 84.59±1.74         | 77.45±2.39         |  | 77.89±7.28         | 79.89±3.69         |  | 79.03±4.22         | 78.49±2.58         |
| SHE                       | 42.22±20.59        | 90.43±4.76         |  | 62.74±4.01         | 86.38±1.32         |  | 84.60±5.30         | 81.57±1.21         |  | 76.36±5.32         | 82.89±1.22         |  | 66.48±5.98         | 85.32±1.43         |
| GEN                       | 23.00±7.75         | 93.83±2.14         |  | 28.14±2.59         | 91.97±0.66         |  | 40.74±6.61         | 90.14±0.76         |  | 47.03±3.22         | 89.46±0.65         |  | 34.73±1.58         | 91.35±0.69         |
| NAC-UE                    | 15.14±2.60         | 94.86±1.36         |  | 14.33±1.24         | 96.05±0.47         |  | <b>17.03</b> ±0.59 | 95.64±0.44         |  | 26.73±0.80         | 91.85±0.28         |  | 18.31±0.92         | 94.60±0.50         |
| <b>CBD-De</b>             | <b>15.20</b> ±0.59 | <b>97.48</b> ±1.29 |  | <b>14.22</b> ±1.50 | <b>97.85</b> ±0.68 |  | 18.44±1.82         | <b>97.03</b> ±0.95 |  | <b>22.75</b> ±2.21 | <b>93.18</b> ±0.73 |  | <b>17.65</b> ±1.03 | <b>96.39</b> ±0.81 |

Table 14: OOD detection performance on the CIFAR-100 benchmark. Results are formatted as **first**, second, and third best. Following the OpenOOD protocol, we report the average performance over three ResNet-18 checkpoints, each trained exclusively on the InD dataset.

| Method                     | MINIST             |                    |  | SVHN               |                    |  | Textures           |                    |  | Places365          |                    |  | Average            |                    |
|----------------------------|--------------------|--------------------|--|--------------------|--------------------|--|--------------------|--------------------|--|--------------------|--------------------|--|--------------------|--------------------|
|                            | FPR95↓             | AUROC↑             |  | FPR95↓             | AUROC↑             |  | FPR95↓             | AUROC↑             |  | FPR95↓             | AUROC↑             |  | FPR95↓             | AUROC↑             |
| <i>CIFAR-100 Benchmark</i> |                    |                    |  |                    |                    |  |                    |                    |  |                    |                    |  |                    |                    |
| OpenMax                    | 53.82±4.74         | 76.01±1.39         |  | 53.20±1.78         | 82.07±1.53         |  | 56.12±1.91         | 80.56±0.09         |  | 54.85±1.42         | 79.29±0.40         |  | 54.50±0.68         | 79.48±0.41         |
| MSP                        | 57.23±4.68         | 76.08±1.86         |  | 59.07±2.53         | 78.42±0.89         |  | 61.88±1.28         | 77.32±0.71         |  | 56.62±0.87         | 79.22±0.29         |  | 58.70±1.06         | 77.76±0.44         |
| TempScale                  | 56.05±4.61         | 77.27±1.85         |  | 57.71±2.68         | 79.79±1.05         |  | 61.56±1.43         | 78.11±0.72         |  | 56.46±0.94         | 79.80±0.25         |  | 57.94±1.14         | 78.74±0.51         |
| ODIN                       | 45.94±3.29         | 83.79±1.31         |  | 67.41±3.88         | 74.54±0.76         |  | 62.37±2.96         | 79.33±1.08         |  | 59.71±0.92         | 79.45±0.26         |  | 58.86±0.79         | 79.28±0.21         |
| MDS                        | 71.72±2.94         | 67.47±0.81         |  | 67.21±6.09         | 70.68±6.40         |  | 70.49±2.48         | 76.26±0.69         |  | 79.61±0.34         | 63.15±0.49         |  | 72.26±1.56         | 69.39±1.39         |
| MDSEns                     | <b>2.83</b> ±0.86  | <b>98.21</b> ±0.78 |  | 82.57±2.58         | 53.76±1.63         |  | 84.94±0.83         | 69.75±1.14         |  | 96.61±0.17         | 42.27±0.73         |  | 66.74±1.04         | 66.00±0.69         |
| RMDS                       | 52.05±6.28         | 79.74±2.49         |  | 51.65±3.68         | 84.89±1.10         |  | 53.99±1.06         | 83.65±0.51         |  | <b>53.57</b> ±0.43 | <b>83.40</b> ±0.46 |  | 52.81±0.63         | 82.92±0.42         |
| Gram                       | 53.53±7.45         | 80.71±4.15         |  | 20.06±1.96         | 95.55±0.60         |  | 89.51±2.54         | 70.79±1.32         |  | 94.67±0.60         | 46.38±1.21         |  | 64.44±2.37         | 73.36±1.08         |
| EBO                        | 52.62±3.83         | 79.18±1.37         |  | 53.62±3.14         | 82.03±1.74         |  | 62.35±2.06         | 78.35±0.83         |  | 57.75±0.86         | 79.52±0.23         |  | 56.59±1.38         | 79.77±0.61         |
| OpenGAN                    | 63.09±23.25        | 68.14±18.78        |  | 70.35±2.06         | 68.40±2.15         |  | 74.77±1.78         | 65.84±3.43         |  | 73.75±8.32         | 69.13±7.08         |  | 70.49±7.38         | 67.88±7.16         |
| GradNorm                   | 86.97±1.44         | 65.35±1.12         |  | 69.90±7.94         | 76.95±4.73         |  | 92.51±0.61         | 64.58±0.13         |  | 85.32±0.44         | 69.69±0.17         |  | 83.68±1.92         | 69.14±1.05         |
| ReAct                      | 56.04±3.66         | 78.37±1.59         |  | 50.41±2.02         | 83.01±0.97         |  | 55.04±0.82         | 80.15±0.46         |  | 55.30±0.41         | 80.03±0.11         |  | 54.20±1.56         | 80.39±0.49         |
| MLS                        | 52.95±3.82         | 78.91±1.47         |  | 53.90±3.04         | 81.65±1.49         |  | 62.39±2.13         | 78.39±0.84         |  | 57.68±0.91         | 79.75±0.24         |  | 56.73±1.33         | 79.67±0.57         |
| KLM                        | 73.09±6.67         | 74.15±2.59         |  | 50.30±7.04         | 79.34±0.44         |  | 81.80±5.80         | 75.77±0.45         |  | 81.40±1.58         | 75.70±0.24         |  | 71.65±2.01         | 76.24±0.52         |
| VIM                        | 48.32±1.07         | 81.89±1.02         |  | 46.22±5.46         | 83.14±3.71         |  | 46.86±2.29         | 85.91±0.78         |  | 61.57±0.77         | 75.85±0.37         |  | 50.74±1.00         | 81.70±0.62         |
| KNIN                       | 48.58±4.67         | 82.36±1.52         |  | 51.75±3.12         | 84.15±1.09         |  | 53.56±2.32         | 83.66±0.83         |  | 60.70±1.03         | 79.43±0.47         |  | 53.65±0.28         | 82.40±0.17         |
| DICE                       | 51.79±3.67         | 79.86±1.89         |  | 49.58±3.32         | 84.22±2.00         |  | 64.23±1.65         | 77.63±0.34         |  | 59.39±1.25         | 78.33±0.66         |  | 56.25±0.60         | 80.01±0.18         |
| RankFeat                   | 75.01±5.83         | 63.03±3.86         |  | 58.49±2.30         | 72.14±1.39         |  | 66.87±3.80         | 69.40±3.08         |  | 77.42±1.96         | 63.82±1.83         |  | 69.45±1.01         | 67.10±1.42         |
| ASH                        | 66.58±3.88         | 77.23±0.46         |  | 46.00±2.67         | 85.60±1.40         |  | 61.27±2.74         | 80.72±0.70         |  | 62.95±0.99         | 78.76±0.16         |  | 59.20±2.46         | 80.58±0.66         |
| SHE                        | 58.78±2.70         | 76.76±1.07         |  | 59.15±7.61         | 80.97±3.98         |  | 73.29±3.22         | 73.64±1.28         |  | 65.24±0.98         | 76.30±0.51         |  | 64.12±2.70         | 76.92±1.16         |
| GEN                        | 53.92±5.71         | 78.29±2.05         |  | 55.45±2.76         | 81.41±1.50         |  | 61.23±1.40         | 78.74±0.81         |  | 56.25±1.01         | 80.28±0.27         |  | 56.71±1.59         | 79.68±0.75         |
| NAC-UE                     | 21.97±6.62         | 93.15±1.63         |  | 24.39±4.66         | 92.40±1.26         |  | 40.65±1.94         | 89.32±0.55         |  | 73.57±1.16         | 73.05±0.68         |  | 40.14±1.86         | 86.98±0.37         |
| <b>CBD-De</b>              | <u>27.93</u> ±7.73 | <u>96.29</u> ±1.81 |  | <b>18.72</b> ±6.12 | <b>95.80</b> ±2.05 |  | <b>37.34</b> ±2.30 | <b>90.02</b> ±1.52 |  | <u>55.23</u> ±0.88 | <u>81.50</u> ±1.16 |  | <b>34.81</b> ±1.60 | <b>90.90</b> ±0.68 |

Table 15: OOD detection performance on the ImageNet-1k benchmark. Results are formatted as first, second, and third best. Following the OpenOOD protocol, we report the average performance over ResNet-50 and Vit-b16 backbones, each trained exclusively on the InD dataset.

| Method        | iNaturalist  |              |              | OpenImage-O  |              |              | Textures     |              |              |
|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|               | ResNet-50    | Vit-b16      | Average      | ResNet-50    | Vit-b16      | Average      | ResNet-50    | Vit-b16      | Average      |
| OpenMax       | 92.05        | 94.93        | <u>93.49</u> | 87.62        | 87.36        | 87.49        | 88.10        | 85.52        | 86.81        |
| MSP           | 88.41        | 88.19        | 88.30        | 84.86        | 84.86        | 84.86        | 82.43        | 85.06        | 83.75        |
| TempScale     | 90.50        | 88.54        | 89.52        | 87.22        | 85.04        | 86.13        | 84.95        | 85.39        | 85.17        |
| ODIN          | 91.17        | /            | 91.17        | 88.23        | /            | 88.23        | 89.00        | /            | 89.00        |
| MDS           | 63.67        | <u>96.01</u> | 79.84        | 69.27        | <b>92.38</b> | 80.83        | 89.80        | 89.41        | 89.61        |
| MDSEns        | 61.82        | /            | 61.82        | 60.80        | /            | 60.80        | 79.94        | /            | 79.94        |
| RMDS          | 87.24        | <u>96.10</u> | 91.67        | 85.84        | <u>92.32</u> | 89.08        | 86.08        | 89.38        | 87.73        |
| Gram          | 76.67        | /            | 76.67        | 74.43        | /            | 74.43        | 88.02        | /            | 88.02        |
| EBO           | 90.63        | 79.30        | 84.97        | 89.06        | 76.48        | 82.77        | 88.70        | 81.17        | 84.94        |
| OpenGAN       | /            | /            | /            | /            | /            | /            | /            | /            | /            |
| GradNorm      | 93.89        | 42.42        | 68.16        | 84.82        | 37.82        | 61.32        | 92.05        | 44.99        | 68.52        |
| ReAct         | 96.34        | 86.11        | 91.23        | <u>91.87</u> | 84.29        | 88.08        | 92.79        | 86.66        | 89.73        |
| MLS           | 91.17        | 85.29        | 88.23        | 89.17        | 81.60        | 85.39        | 88.39        | 83.74        | 86.07        |
| KLM           | 90.78        | 89.59        | 90.19        | 87.30        | 87.03        | 87.17        | 84.72        | 86.49        | 85.61        |
| VIM           | 89.56        | 95.72        | 92.64        | 90.50        | 92.18        | <u>91.34</u> | 97.97        | 90.61        | <u>94.29</u> |
| KNN           | 86.41        | 91.46        | 88.94        | 87.04        | <u>89.86</u> | 88.45        | 97.09        | 91.12        | 94.11        |
| DICE          | 92.54        | 82.50        | 87.52        | 88.26        | 82.22        | 85.24        | 92.04        | 82.21        | 87.13        |
| RankFeat      | 40.06        | /            | 40.06        | 50.83        | /            | 50.83        | 70.90        | /            | 70.90        |
| ASH           | <b>97.07</b> | 50.62        | 73.85        | <b>93.26</b> | 55.51        | 74.39        | 96.90        | 48.53        | 72.72        |
| SHE           | 92.65        | 93.57        | 93.11        | 86.52        | 91.04        | 88.78        | 93.60        | 92.65        | 93.13        |
| GEN           | 92.44        | 93.54        | 92.99        | 89.26        | 90.27        | 89.77        | 87.59        | <u>90.23</u> | 88.91        |
| NAC-UE        | 96.52        | 93.72        | 95.12        | 91.45        | 91.58        | 91.52        | 97.90        | 94.17        | 96.04        |
| <b>CBD-De</b> | <u>97.06</u> | <b>97.26</b> | <b>97.16</b> | <u>92.65</u> | 91.89        | <b>92.27</b> | <b>98.31</b> | <b>94.30</b> | <b>96.31</b> |



## F Full OOD Generalization Results

Table 16: Out-of-domain accuracies (%) on PACS. Results are formatted as **first**, second, and third best.

| Method         | A                              | C                              | P                              | S                              | Avg         |
|----------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|-------------|
| IRM            | 84.8 $\pm$ 1.3                 | 76.4 $\pm$ 1.1                 | 96.7 $\pm$ 0.6                 | 76.1 $\pm$ 1.0                 | 83.5        |
| ERM            | 85.7 $\pm$ 0.6                 | 77.1 $\pm$ 0.8                 | 97.4 $\pm$ 0.4                 | 76.6 $\pm$ 0.7                 | 84.2        |
| GroupDRO       | 83.5 $\pm$ 0.9                 | 79.1 $\pm$ 0.6                 | 96.7 $\pm$ 0.7                 | 78.3 $\pm$ 2.0                 | 84.4        |
| MTL            | 87.5 $\pm$ 0.8                 | 77.1 $\pm$ 0.5                 | 96.4 $\pm$ 0.8                 | 77.3 $\pm$ 1.8                 | 84.6        |
| Mixup          | 86.1 $\pm$ 0.5                 | 78.9 $\pm$ 0.8                 | 97.6 $\pm$ 0.1                 | 75.8 $\pm$ 1.8                 | 84.6        |
| MMD            | 86.1 $\pm$ 1.4                 | 79.4 $\pm$ 0.9                 | 96.6 $\pm$ 0.2                 | 76.5 $\pm$ 0.5                 | 84.7        |
| VREx           | 86.0 $\pm$ 1.6                 | 79.1 $\pm$ 0.6                 | 96.9 $\pm$ 0.5                 | 77.7 $\pm$ 1.7                 | 84.9        |
| MLDG           | 85.5 $\pm$ 1.4                 | 80.1 $\pm$ 1.7                 | 97.4 $\pm$ 0.3                 | 76.6 $\pm$ 1.1                 | 84.9        |
| ARM            | 86.8 $\pm$ 0.6                 | 76.8 $\pm$ 0.5                 | 97.4 $\pm$ 0.3                 | 79.3 $\pm$ 1.2                 | 85.1        |
| Mixstyle       | 86.8 $\pm$ 0.5                 | 79.0 $\pm$ 1.4                 | 96.6 $\pm$ 0.1                 | 78.5 $\pm$ 2.3                 | 85.2        |
| CORAL          | 88.3 $\pm$ 0.2                 | 80.0 $\pm$ 0.5                 | 97.5 $\pm$ 0.3                 | 78.8 $\pm$ 1.3                 | 86.2        |
| SagNet         | 87.4 $\pm$ 0.2                 | 80.7 $\pm$ 0.5                 | 97.1 $\pm$ 0.1                 | 80.0 $\pm$ 1.0                 | 86.3        |
| RSC            | 85.4 $\pm$ 0.9                 | 79.7 $\pm$ 0.5                 | 97.6 $\pm$ 0.9                 | 78.2 $\pm$ 1.0                 | 85.2        |
| SAM            | 85.6 $\pm$ 2.1                 | 80.9 $\pm$ 1.2                 | 97.0 $\pm$ 0.4                 | 79.6 $\pm$ 1.6                 | 85.8        |
| GSAM           | 86.9 $\pm$ 0.1                 | 80.4 $\pm$ 0.2                 | 97.5 $\pm$ 0.0                 | 78.7 $\pm$ 0.8                 | 85.9        |
| SAGM           | 87.4 $\pm$ 0.2                 | 80.2 $\pm$ 0.3                 | 98.0 $\pm$ 0.2                 | 80.8 $\pm$ 0.6                 | 86.6        |
| GGA            | 88.8 $\pm$ 0.2                 | 80.1 $\pm$ 0.3                 | 97.3 $\pm$ 0.2                 | 81.2 $\pm$ 0.5                 | 87.3        |
| <b>CBD-Gen</b> | <b>89.0<math>\pm</math>0.6</b> | <b>81.6<math>\pm</math>0.4</b> | <b>98.1<math>\pm</math>0.3</b> | <b>82.1<math>\pm</math>0.8</b> | <b>87.7</b> |

Table 17: Out-of-domain accuracies (%) on VLCS. Results are formatted as **first**, second, and third best.

| Method         | C                              | L                              | S                              | V                              | Avg         |
|----------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|-------------|
| GroupDRO       | 97.3 $\pm$ 0.3                 | 63.4 $\pm$ 0.9                 | 69.5 $\pm$ 0.8                 | 76.7 $\pm$ 0.7                 | 76.7        |
| MLDG           | 97.4 $\pm$ 0.2                 | 65.2 $\pm$ 0.7                 | 71.0 $\pm$ 1.4                 | 75.3 $\pm$ 1.0                 | 77.2        |
| MTL            | 97.8 $\pm$ 0.4                 | 64.3 $\pm$ 0.3                 | 71.5 $\pm$ 0.7                 | 75.3 $\pm$ 1.7                 | 77.2        |
| ERM            | 98.0 $\pm$ 0.3                 | 64.7 $\pm$ 1.2                 | 71.4 $\pm$ 1.2                 | 75.2 $\pm$ 1.6                 | 77.3        |
| Mixup          | 98.3 $\pm$ 0.6                 | 64.8 $\pm$ 1.0                 | 72.1 $\pm$ 0.5                 | 74.3 $\pm$ 0.8                 | 77.4        |
| MMD            | 97.7 $\pm$ 0.1                 | 64.0 $\pm$ 1.1                 | 72.8 $\pm$ 0.2                 | 75.3 $\pm$ 3.3                 | 77.5        |
| ARM            | 98.7 $\pm$ 0.2                 | 63.6 $\pm$ 0.7                 | 71.3 $\pm$ 1.2                 | 76.7 $\pm$ 0.6                 | 77.6        |
| SagNet         | 97.9 $\pm$ 0.4                 | 64.5 $\pm$ 0.5                 | 71.4 $\pm$ 1.3                 | 77.5 $\pm$ 0.5                 | 77.8        |
| Mixstyle       | 98.6 $\pm$ 0.3                 | 64.5 $\pm$ 1.1                 | 72.6 $\pm$ 0.5                 | 75.7 $\pm$ 1.7                 | 77.9        |
| VREx           | 98.4 $\pm$ 0.3                 | 64.4 $\pm$ 1.4                 | 74.1 $\pm$ 0.4                 | 76.2 $\pm$ 1.3                 | 78.3        |
| IRM            | 98.6 $\pm$ 0.1                 | 64.9 $\pm$ 0.9                 | 73.4 $\pm$ 0.6                 | 77.3 $\pm$ 0.9                 | 78.6        |
| CORAL          | 98.3 $\pm$ 0.3                 | 66.1 $\pm$ 0.6                 | 73.4 $\pm$ 0.3                 | 77.5 $\pm$ 1.0                 | 78.8        |
| RSC            | 97.9 $\pm$ 0.1                 | 62.5 $\pm$ 0.7                 | 72.3 $\pm$ 1.2                 | 75.6 $\pm$ 0.8                 | 77.1        |
| GSAM           | 98.7 $\pm$ 0.3                 | 64.9 $\pm$ 0.2                 | 74.3 $\pm$ 0.0                 | 78.5 $\pm$ 0.8                 | 79.1        |
| SAM            | 99.1 $\pm$ 0.2                 | 65.0 $\pm$ 1.0                 | 73.7 $\pm$ 1.0                 | 79.8 $\pm$ 0.1                 | 79.4        |
| SAGM           | 99.0 $\pm$ 0.2                 | 65.2 $\pm$ 0.4                 | 75.1 $\pm$ 0.3                 | 80.7 $\pm$ 0.8                 | 80.0        |
| GGA            | 99.1 $\pm$ 0.2                 | 67.5 $\pm$ 0.6                 | 75.1 $\pm$ 0.3                 | 78.0 $\pm$ 0.1                 | 79.9        |
| <b>CBD-Gen</b> | <b>99.1<math>\pm</math>0.2</b> | <b>67.3<math>\pm</math>0.7</b> | <b>75.1<math>\pm</math>0.8</b> | <b>80.5<math>\pm</math>1.0</b> | <b>80.5</b> |

Table 18: Out-of-domain accuracies (%) on OfficeHome. Results are formatted as **first**, second, and third best.

| Algorithm      | A                     | C                     | P                     | R                     | Avg         |
|----------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------|
| Mixstyle       | 51.1 $\pm$ 0.3        | 53.2 $\pm$ 0.4        | 68.2 $\pm$ 0.7        | 69.2 $\pm$ 0.6        | 60.4        |
| IRM            | 58.9 $\pm$ 2.3        | 52.2 $\pm$ 1.6        | 72.1 $\pm$ 2.9        | 74.0 $\pm$ 2.5        | 64.3        |
| ARM            | 58.9 $\pm$ 0.8        | 51.0 $\pm$ 0.5        | 74.1 $\pm$ 0.1        | 75.2 $\pm$ 0.3        | 64.8        |
| GroupDRO       | 60.4 $\pm$ 0.7        | 52.7 $\pm$ 1.0        | 75.0 $\pm$ 0.7        | 76.0 $\pm$ 0.7        | 66.0        |
| MMD            | 60.4 $\pm$ 0.2        | 53.3 $\pm$ 0.3        | 74.3 $\pm$ 0.1        | 77.4 $\pm$ 0.6        | 66.4        |
| MTL            | 61.5 $\pm$ 0.7        | 52.4 $\pm$ 0.6        | 74.9 $\pm$ 0.4        | 76.8 $\pm$ 0.4        | 66.4        |
| VREx           | 60.7 $\pm$ 0.9        | 53.0 $\pm$ 0.9        | 75.3 $\pm$ 0.1        | 76.6 $\pm$ 0.5        | 66.4        |
| ERM            | 63.1 $\pm$ 0.3        | 51.9 $\pm$ 0.4        | 77.2 $\pm$ 0.5        | 78.1 $\pm$ 0.2        | 67.6        |
| MLDG           | 61.5 $\pm$ 0.9        | 53.2 $\pm$ 0.6        | 75.0 $\pm$ 1.2        | 77.5 $\pm$ 0.4        | 66.8        |
| Mixup          | 62.4 $\pm$ 0.8        | 54.8 $\pm$ 0.6        | 76.9 $\pm$ 0.3        | 78.3 $\pm$ 0.2        | 68.1        |
| SagNet         | 63.4 $\pm$ 0.2        | 54.8 $\pm$ 0.4        | 75.8 $\pm$ 0.4        | 78.3 $\pm$ 0.3        | 68.1        |
| CORAL          | <u>65.3</u> $\pm$ 0.3 | 54.4 $\pm$ 0.6        | 76.5 $\pm$ 0.3        | 78.4 $\pm$ 1.0        | 68.7        |
| RSC            | 60.7 $\pm$ 1.4        | 51.4 $\pm$ 0.3        | 74.8 $\pm$ 1.1        | 75.1 $\pm$ 1.3        | 65.5        |
| GSAM           | 64.9 $\pm$ 0.1        | 55.2 $\pm$ 0.2        | <u>77.8</u> $\pm$ 0.0 | 79.2 $\pm$ 0.2        | 69.3        |
| SAM            | 64.5 $\pm$ 0.3        | <u>56.5</u> $\pm$ 0.2 | 77.4 $\pm$ 0.1        | <u>79.8</u> $\pm$ 0.4 | <u>69.6</u> |
| SAGM           | <u>65.4</u> $\pm$ 0.4 | <u>57.0</u> $\pm$ 0.3 | <u>78.0</u> $\pm$ 0.3 | <u>80.0</u> $\pm$ 0.2 | <u>70.1</u> |
| GGA            | 64.3 $\pm$ 0.1        | 54.4 $\pm$ 0.2        | 76.5 $\pm$ 0.3        | 78.9 $\pm$ 0.2        | 68.5        |
| <b>CBD-Gen</b> | <b>66.5</b> $\pm$ 0.9 | <b>58.7</b> $\pm$ 0.7 | <b>78.8</b> $\pm$ 1.2 | <b>81.6</b> $\pm$ 0.8 | <b>71.4</b> |

Table 19: Out-of-domain accuracies (%) on TerraInc. Results are formatted as **first**, second, and third best.

| Method         | L100                  | L38                   | L43                   | L46                   | Avg         |
|----------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------|
| MMD            | 41.9 $\pm$ 3.0        | 34.8 $\pm$ 1.0        | 57.0 $\pm$ 1.9        | 35.2 $\pm$ 1.8        | 42.2        |
| GroupDRO       | 41.2 $\pm$ 0.7        | 38.6 $\pm$ 2.1        | 56.7 $\pm$ 0.9        | 36.4 $\pm$ 2.1        | 43.2        |
| Mixstyle       | 54.3 $\pm$ 1.1        | 34.1 $\pm$ 1.1        | 55.9 $\pm$ 1.1        | 31.7 $\pm$ 2.1        | 44.0        |
| ARM            | 49.3 $\pm$ 0.7        | 38.3 $\pm$ 0.7        | 55.8 $\pm$ 0.8        | 38.7 $\pm$ 1.3        | 45.5        |
| MTL            | 49.3 $\pm$ 1.2        | 39.6 $\pm$ 6.3        | 55.6 $\pm$ 1.1        | 37.8 $\pm$ 0.8        | 45.6        |
| ERM            | 49.8 $\pm$ 4.4        | 42.1 $\pm$ 1.4        | 56.9 $\pm$ 1.8        | 35.7 $\pm$ 3.9        | 46.1        |
| VREx           | 48.2 $\pm$ 4.3        | 41.7 $\pm$ 1.3        | 56.8 $\pm$ 0.8        | 38.7 $\pm$ 3.1        | 46.4        |
| IRM            | 54.6 $\pm$ 1.3        | 39.8 $\pm$ 1.9        | 56.2 $\pm$ 1.8        | 39.6 $\pm$ 0.8        | 47.6        |
| CORAL          | 51.6 $\pm$ 2.4        | 42.2 $\pm$ 1.0        | 57.0 $\pm$ 1.0        | 39.8 $\pm$ 2.9        | 47.6        |
| MLDG           | 54.2 $\pm$ 3.0        | <u>44.3</u> $\pm$ 1.1 | 55.6 $\pm$ 0.3        | 36.9 $\pm$ 2.2        | 47.8        |
| Mixup          | <b>59.6</b> $\pm$ 2.0 | 42.2 $\pm$ 1.4        | 55.9 $\pm$ 0.8        | 33.9 $\pm$ 1.4        | 47.9        |
| SagNet         | 53.0 $\pm$ 2.0        | 43.0 $\pm$ 1.4        | 57.9 $\pm$ 0.8        | 40.4 $\pm$ 1.4        | 48.6        |
| SAM            | 46.3 $\pm$ 1.0        | 38.4 $\pm$ 2.4        | 54.0 $\pm$ 1.0        | 34.5 $\pm$ 0.8        | 43.3        |
| RSC            | 50.2 $\pm$ 2.2        | 39.2 $\pm$ 1.4        | 56.3 $\pm$ 1.4        | 40.8 $\pm$ 0.6        | 46.6        |
| GSAM           | 50.8 $\pm$ 0.1        | 39.3 $\pm$ 0.2        | <u>59.6</u> $\pm$ 0.0 | 38.2 $\pm$ 0.8        | 47.0        |
| SAGM           | 54.8 $\pm$ 1.3        | 41.4 $\pm$ 0.8        | 57.7 $\pm$ 0.6        | <u>41.3</u> $\pm$ 0.4 | <u>48.8</u> |
| GGA            | <u>55.9</u> $\pm$ 0.1 | <b>45.5</b> $\pm$ 0.1 | <b>59.7</b> $\pm$ 0.1 | <b>41.5</b> $\pm$ 0.1 | <b>50.6</b> |
| <b>CBD-Gen</b> | <u>56.4</u> $\pm$ 2.1 | <u>45.0</u> $\pm$ 1.2 | <u>59.6</u> $\pm$ 0.6 | <u>41.0</u> $\pm$ 1.5 | <u>50.5</u> |

Table 20: Out-of-domain accuracies (%) on DomainNet. Results are formatted as **first**, second, and third best.

| Algorithm      | clip                  | info                  | paint                 | quick                 | real                  | sketch                | Avg         |
|----------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------|
| MMD            | 32.1 $\pm$ 13.3       | 11.0 $\pm$ 4.6        | 26.8 $\pm$ 11.3       | 8.7 $\pm$ 2.1         | 32.7 $\pm$ 13.8       | 28.9 $\pm$ 11.9       | 23.4        |
| GroupDRO       | 47.2 $\pm$ 0.5        | 17.5 $\pm$ 0.4        | 33.8 $\pm$ 0.5        | 9.3 $\pm$ 0.3         | 51.6 $\pm$ 0.4        | 40.1 $\pm$ 0.6        | 33.3        |
| VREx           | 47.3 $\pm$ 3.5        | 16.0 $\pm$ 1.5        | 35.8 $\pm$ 4.6        | 10.9 $\pm$ 0.3        | 49.6 $\pm$ 4.9        | 42.0 $\pm$ 3.0        | 33.6        |
| IRM            | 48.5 $\pm$ 2.8        | 15.0 $\pm$ 1.5        | 38.3 $\pm$ 4.3        | 10.9 $\pm$ 0.5        | 48.2 $\pm$ 5.2        | 42.3 $\pm$ 3.1        | 33.9        |
| Mixstyle       | 51.9 $\pm$ 0.4        | 13.3 $\pm$ 0.2        | 37.0 $\pm$ 0.5        | 12.3 $\pm$ 0.1        | 46.1 $\pm$ 0.3        | 43.4 $\pm$ 0.4        | 34.0        |
| ARM            | 49.7 $\pm$ 0.3        | 16.3 $\pm$ 0.5        | 40.9 $\pm$ 1.1        | 9.4 $\pm$ 0.1         | 53.4 $\pm$ 0.4        | 43.5 $\pm$ 0.4        | 35.5        |
| Mixup          | 55.7 $\pm$ 0.3        | 18.5 $\pm$ 0.5        | 44.3 $\pm$ 0.5        | 12.5 $\pm$ 0.4        | 55.8 $\pm$ 0.3        | 48.2 $\pm$ 0.5        | 39.2        |
| SagNet         | 57.7 $\pm$ 0.3        | 19.0 $\pm$ 0.2        | 45.3 $\pm$ 0.3        | 12.7 $\pm$ 0.5        | 58.1 $\pm$ 0.5        | 48.8 $\pm$ 0.2        | 40.3        |
| MTL            | 57.9 $\pm$ 0.5        | 18.5 $\pm$ 0.4        | 46.0 $\pm$ 0.1        | 12.5 $\pm$ 0.1        | 59.5 $\pm$ 0.3        | 49.2 $\pm$ 0.1        | 40.6        |
| MLDG           | 59.1 $\pm$ 0.2        | 19.1 $\pm$ 0.3        | 45.8 $\pm$ 0.7        | 13.4 $\pm$ 0.3        | 59.6 $\pm$ 0.2        | 50.2 $\pm$ 0.4        | 41.2        |
| CORAL          | 59.2 $\pm$ 0.1        | 19.7 $\pm$ 0.2        | 46.6 $\pm$ 0.3        | 13.4 $\pm$ 0.4        | 59.8 $\pm$ 0.2        | 50.1 $\pm$ 0.6        | 41.5        |
| ERM            | 63.0 $\pm$ 0.2        | <u>21.2</u> $\pm$ 0.2 | 50.1 $\pm$ 0.4        | 13.9 $\pm$ 0.5        | 63.7 $\pm$ 0.2        | 52.0 $\pm$ 0.5        | 43.8        |
| RSC            | 55.0 $\pm$ 1.2        | 18.3 $\pm$ 0.5        | 44.4 $\pm$ 0.6        | 12.2 $\pm$ 0.2        | 55.7 $\pm$ 0.7        | 47.8 $\pm$ 0.9        | 38.9        |
| SAM            | <u>64.5</u> $\pm$ 0.3 | 20.7 $\pm$ 0.2        | 50.2 $\pm$ 0.1        | <b>15.1</b> $\pm$ 0.3 | 62.6 $\pm$ 0.2        | 52.7 $\pm$ 0.3        | 44.3        |
| GSAM           | <u>64.2</u> $\pm$ 0.3 | 20.8 $\pm$ 0.2        | 50.9 $\pm$ 0.0        | <u>14.4</u> $\pm$ 0.8 | 63.5 $\pm$ 0.2        | <u>53.9</u> $\pm$ 0.2 | 44.6        |
| SAGM           | <b>64.9</b> $\pm$ 0.2 | 21.1 $\pm$ 0.3        | <u>51.5</u> $\pm$ 0.2 | <u>14.8</u> $\pm$ 0.2 | 64.1 $\pm$ 0.2        | 53.6 $\pm$ 0.2        | <u>45.0</u> |
| GGA            | 64.0 $\pm$ 0.2        | <u>22.2</u> $\pm$ 0.3 | <u>51.7</u> $\pm$ 0.1 | 14.3 $\pm$ 0.2        | <u>64.1</u> $\pm$ 0.4 | <u>54.3</u> $\pm$ 0.3 | <u>45.2</u> |
| <b>CBD-Gen</b> | <b>64.9</b> $\pm$ 0.4 | <b>22.8</b> $\pm$ 0.3 | <b>52.5</b> $\pm$ 0.3 | <b>15.1</b> $\pm$ 0.6 | <b>64.7</b> $\pm$ 0.4 | <b>55.4</b> $\pm$ 0.7 | <b>45.9</b> |

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The experimental results mentioned in the abstract and the contributions mentioned in the introduction accurately reflect the contributions and scope of the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: A large wave vector  $k$  will reduce the training efficiency. In addition, the algorithm involves retaining a large number of gradient maps, which poses a challenge to efficient and GPU memory, and is mentioned in the experimental section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: For the six theories in the main text, we provide complete proofs in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide implementation details in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Time is tight, the code has a lot of redundancy and is not organized. We will release the code after published.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Detailed datasets and hyperparameters are provided in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The upper and lower limits of error are given in the experimental table.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).



- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Implementation details of the Appendix mentions graphics card resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: Our work does not involve ethical issues.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: No need to discuss these.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release any new datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The code, data, and models used are all public.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This question is not suitable for our work.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human subjects involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This question is not suitable for our work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This question is not suitable for our work.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.