GROD: Enhancing Generalization of Transformer with Out-of-Distribution Detection

Yijin Zhou [1] **Yuguang Wang** [1 2 3 4]

Abstract

Transformer networks face challenges in generalizing to Out-of-Distribution (OOD) datasets, that is, data whose distribution differs from that seen during training. Utilizing an OOD detection framework based on Probably Approximately Correct (PAC) theory, the proposed *Generate Rounded OOD Data* (GROD) algorithm, a novel approach to enhancing transformer networks' generalization across various natural language processing and computer vision datasets, improves transformers' ability to in-distribution (ID) data boundary decision-making and detect outliers effectively. By incorporating synthetic outlier generation and penalizing OOD misclassification within the loss function, GROD refines model parameters and ensures robust performance. Empirical evaluations show that GROD achieves state-of-the-art (SOTA) results in natural language processing (NLP) and computer vision (CV) tasks, significantly reducing the SOTA FPR@95 *from* 21.97% *to* 0.12%, and improving AUROC *from* 93.62% *to* 99.98% on image classification tasks, and the SOTA FPR@95 by 12.89% and AUROC by 2.27% in detecting semantic text outliers. The code is available at https://anonymous.4open.science/r/GROD-OOD-Detection-with-transformers-B70F.

1. Introduction

Mainstream machine learning algorithms typically assume data independence, called in-distribution (ID) data (Krizhevsky et al., 2012; He et al., 2015). However, in practical applications, data often follows the "open world" assumption (Drummond & Shearer, 2006), where outliers with different distributions can occur during inference. This real-world challenge frequently degrades the performance of AI models in prediction tasks. One remedy is to incorporate OOD detection techniques. This paper proposes a new algorithm based on OOD detection for transformer networks, which can significantly improve their performance in predicting outlier instances.

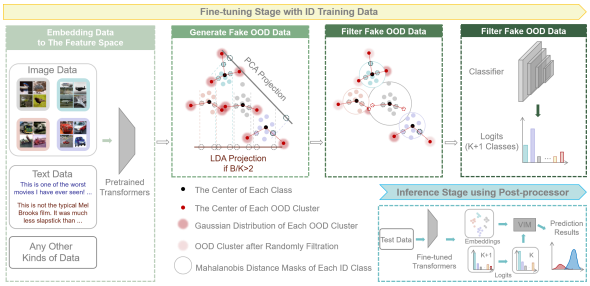

Figure 1. Overview of GROD algorithm: In the fine-tuning stage, GROD generates fake OOD data as part of the training data. GROD then guides the training by incorporating the ID-OOD classifier in the loss. In the inference stage, the features and adjusted LOGITS are input into the post-processor.

The transformer is a deep neural network architecture that leverages an attention mechanism. It is renowned for its powerful capabilities in a variety of deep learning models, such as large language models, computer vision models, and graph neural networks. OOD detection aims to identify and manage semantically distinct outliers, referred to as *OOD data*. It requires the designed algorithm to detect OOD instances and avoid making predictions on them, while maintaining robust performance on ID data. By employing OOD detection, we develop a new algorithm, which we call **G**enerate **R**ounded **OOD** **D**ata (GROD), for fine-tuning a transformer network to enhance its ability to predict the unknown distribution. By taking account of the OOD Detection in network training, we can strengthen the recognition of the in-distribution and out-distribution boundary.

We establish the OOD Detection PAC Learning Theorem (Theorem 2.1). It demonstrates that penalizing the misclassification of OOD data in the training loss of the transformer

[1]School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China [2]Institute of Natural Sciences, Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai, China [3]Shanghai AI Laboratory, Shanghai, China [4]School of Mathematics and Statistics, University of New South Wales, Sydney, Australia. Correspondence to: Yuguang Wang <yuguang.wang@sjtu.edu.cn>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

clarifies the decision boundary between inliers and outliers. This condition ensures that the model possesses *OOD Detection Learnability*. Moreover, we quantify the learnability by proving an error boundary regarding the transformer model's budget (the number of total trainable parameters) (Theorem 2.2). We define GROD following these two theorems. When the network depth is substantial, the GROD-enhanced transformer converges to the target mapping with robust generalization capabilities. Our main contributions are summarized as follows:

- We establish a PAC learning framework for OOD detection applied to transformers, providing necessary and sufficient conditions as well as error boundary estimates for learnability.
- Inspired by the learning theory and empirical validation, we propose a novel OOD detection approach, *Generate Rounded OOD Data* (GROD). This strategy is theoretically grounded and high-quality in generating and representing features regardless of data types.
- We conduct comprehensive experiments to display the state-of-the-art (SOTA) performance of GROD on image and text datasets together with ablation studies and visualizations for interpretability.

2. GROD algorithm

Notation. We introduce some notations regarding OOD detection tasks. Formally, $\mathcal{X}$ and $\mathcal{Y} := \{1, 2, \cdots, K, K+1\}$ denote the whole dataset and its label space. As subsets in $\mathcal{X}$, $\mathcal{X}_{\text{train}}$, $\mathcal{X}_{\text{test}}$ and $\mathcal{X}_I$ represents the training dataset, test dataset and ID dataset, respectively. $\mathcal{Y}_I := \{1, \cdots, K\}$ denote the ID label space. $l(\mathbf{y}_1, \mathbf{y}_2), \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ denotes the paired loss of the prediction and label of one data, and $\mathcal{L}$ denotes the total loss. We depict the basic structure of a transformer network as follows, which includes the following components: input embedding, positional encoding, an encoder, a decoder and an output layer. For OOD detection tasks, which predominantly encompass classification objectives, we directly connect an output layer subsequent to the encoder to streamline the process. For clarity and operational simplicity, we assume that the input data $\mathcal{X}$ is processed by the input embedding and positional encoding mechanisms. The encoder is an assembly of multiple attention blocks, each comprising a self-attention layer and a Feed-forward Fully Connected Network (FFCN). The self-attention layer calculates matrices of key, query, and value, to express the self-attention mechanism, where the hidden dimensions for keys and queries are m_h, and for values are m_V. Each individual data is transformed into τ tokens, with each token having a dimension $\hat{d}$. To quantify the computational overhead of a transformer block, we define the budget $m := (\hat{d}, h, m_h, m_V, r)$, representing the parameter size of one block. More details about notations and preliminaries for theoretical analysis are illustrated in Appendix B.

Framework overview. As illustrated in Figure 1, GROD contains several pivotal steps. Firstly, a binary ID-OOD classification loss function is added for fine-tuning the transformer. This adjustment aligns more closely with the transformer's learnable conditions in the proposed theory. To effectively leverage this binary classification loss, we introduce a novel strategy for synthesizing high-quality OOD data for training. To minimize computational overhead while leveraging high-quality embeddings for enhanced efficiency, GROD generates virtual OOD embeddings, rather than utilizing original data. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) projections are employed to generate global and inter-class outliers respectively, utilizing overall ID information and distinct features for each ID data category. Next, a filtering mechanism is applied to remove synthetic ID-like outliers and maintain a reasonable ratio of ID and OOD. This refined dataset together with the binary loss then serve to fine-tune the transformer under GROD framework. During the testing phase, embeddings and prediction LOGITS are extracted from the GROD-enhanced transformer. These outputs are reformulated for post-processing. The post-processor, VIM (Wang et al., 2022), is applied to get the final prediction.

Recognize boundary ID features by PCA and LDA projections. Let $\mathcal{X}_{\text{train}}$ denote the input to the transformer backbone, which is transformed into a feature representation $\mathcal{F} \in \mathbb{R}^{n \times s}$ in the feature space:

$$\mathcal{F} = \text{Feat} \circ \text{Block}^n(\mathcal{X}_{\text{train}}), \tag{1}$$

where $\text{Feat}(\cdot)$ is the process to obtain features. For instance, in ViT models, $\text{Feat}(\cdot)$ represents extracting CLS tokens. Subsequently, we generate synthetic OOD vectors using PCA for global outliers and LDA for inter-class distinctions. LDA is selected for its ID-separating ability, but is only used when $B/K > 2$ to guarantee the robustness of generated OOD, where B is the batch size. We display formulas for combining PCA and LDA projections as the PCA-only condition is included. Specifically, we first find data with maximum and minimum values of each dimension in projection spaces. $\mathcal{F}$ is projected by

$$\mathcal{F}_{\text{PCA}} = \text{PCA}(\mathcal{F}),\ \mathcal{F}_{\text{LDA},i} = \text{LDA}(\mathcal{F}, \mathcal{Y})|_{\mathbf{y}=i},\ i \in \mathcal{Y}_I. \tag{2}$$

Features are mapped from $\mathbb{R}^d$ to $\mathbb{R}^{num}$, $num \leq d$. Then target vectors are acquired, denoted as $v^M_{\text{PCA},j} = \arg\max_{v \in \mathcal{F}_{\text{PCA}}} v_j$, $v^M_{\text{LDA},i,j} = \arg\max_{v \in \mathcal{F}_{\text{LDA},i}} v_j$ for maximum and $v^m_{\text{PCA},j}$, $v^m_{\text{LDA},i,j}$ for minimum, $i \in \mathcal{Y}_I$, $j \in \mathcal{S} := \{1, \cdots s\}$. The sets $\hat{V}_{\text{PCA}} := \{v^M_{\text{PCA},j}\ and\ v^m_{\text{PCA},j}, j \in \mathcal{S}\}$ and $\hat{V}_{\text{LDA},i} :=$

$\{v^M_{\mathrm{LDA},i,j}$ and $v^m_{\mathrm{LDA},i,j}, j \in \mathcal{S}\}, i \in \mathcal{Y}_I$ are the boundary points in the projection spaces, which are mapped back to the original feature space:

$$\begin{aligned} V_{\mathrm{PCA}} &= \mathrm{PCA}^{-1}(\hat{V}_{\mathrm{PCA}}), \\ V_{\mathrm{LDA},i} &= \mathrm{LDA}^{-1}(\hat{V}_{\mathrm{LDA},i}),\ i \in \mathcal{Y}_I, \end{aligned} \tag{3}$$

where PCA^{-1} and LDA^{-1} are the inverse mappings of PCA and LDA according to set theory.

Generate fake OOD data. Boundary points, while initially within ID, are extended into OOD regions. Firstly, the centers of every training batch and category are calculated by $\mu_{\mathrm{PCA}} = \frac{\sum_{v\in\mathcal{F}} v}{|\mathcal{F}|}$ and $\mu_{\mathrm{LDA},i_k} = \frac{\sum_{v\in\mathcal{F}} v|_{\mathbf{y}=i_k}}{B_{i_k}}$, where $i_k \in \{i = 1, \cdots, K : |\mathcal{F}|_{\mathbf{y}=i}| > 1\} := I$. Then we generate Gaussian mixture fake OOD data with expectations U_{OOD}:

$$\begin{aligned} U_{\mathrm{OOD}} =& \{v + \alpha\frac{v-\mu}{||v-\mu||_2 + \epsilon} : v \in V_{\mathrm{PCA}}, \mu = \mu_{\mathrm{PCA}} \text{ or} \\ & v \in V_{\mathrm{LDA},i_k}, \mu = \mu_{\mathrm{LDA},i_k}, i_k \in I\}, \end{aligned} \tag{4}$$

where $\epsilon = 10^{-7}$, α is a hyperparameter representing extension proportion of L_2 norm. Gaussian mixture fake OOD data are generated with distribution

$$D_{\mathrm{OOD}} = \frac{1}{|U_{\mathrm{OOD}}|} \sum_{\mu_{\mathrm{OOD}} \in U_{\mathrm{OOD}}} \mathcal{N}(\mu_{\mathrm{OOD}}, \alpha/3 \cdot I_{\mathrm{OOD}}), \tag{5}$$

where I_{OOD} is the identity matrix. We denote the set of these fake OOD data as $\hat{\mathcal{F}}_{\mathrm{OOD}} := \hat{\mathcal{F}}^{\mathrm{OOD}}_{\mathrm{PCA}} \cup (\cup_{i_k\in I}\hat{\mathcal{F}}^{\mathrm{OOD}}_{\mathrm{LDA},i_k})$, where $\hat{\mathcal{F}}^{\mathrm{OOD}}_{\mathrm{PCA}}$ and $\hat{\mathcal{F}}^{\mathrm{OOD}}_{\mathrm{LDA},i_k}$ are clusters consist of num data points each, in the Gaussian distribution with expectations μ_{PCA} and μ_{LDA,i_k} respectively.

Filter OOD data. We propose the Mahalanobis distance filtering mechanism and random filtration, improving the generation quality of outliers. The Mahalanobis distance filtering mechanism is a strategy to eliminate ID-like synthetic OOD data. It calculates the Mahalanobis distance from each synthetic outlier to the global and inter-class ID centers, and filters those with the closest Mahalanobis distance less than the average ID distance in the same category. The random filtration keeps a reasonable proportion of ID and OOD, maintaining the stability of fine-tuning. The detailed filtration process can be found in Appendix H. We denote the final generated OOD set as $\mathcal{F}_{\mathrm{OOD}}$.

Train-time and test-time OOD detection. During fine-tuning, training data in the feature space is denoted as $\mathcal{F}_{\text{all}} := \mathcal{F} \cup \mathcal{F}_{\mathrm{OOD}}$, with labels $\mathbf{y} \in \mathcal{Y}$. $\mathcal{F}_{\text{all}}$ is fed into a linear classifier for $K+1$ classes. A loss function $\mathcal{L}$ that integrates a binary ID-OOD classification loss $\mathcal{L}_2$, weighted by the cross-entropy loss $\mathcal{L}_1$, to penalize OOD misclassification and improve ID classification, *i.e.*

$$\mathcal{L} = (1-\gamma)\mathcal{L}_1 + \gamma\mathcal{L}_2, \tag{6}$$

$$\mathcal{L}_1(\mathbf{y}, \mathbf{x}) = -\mathbb{E}_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^{K+1} \mathbf{y}_j \log(\sigma(\mathbf{f} \circ \mathbf{H}(\mathbf{x}))_j), \tag{7}$$

$$\mathcal{L}_2(\mathbf{y}, \mathbf{x}) = -\mathbb{E}_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^{2} \hat{\phi}(\mathbf{y})_j \log(\hat{\phi}(\sigma(\mathbf{f} \circ \mathbf{H}(\mathbf{x})))_j), \tag{8}$$

where $\sigma(\cdot)$ is the softmax function.

During the test time, the feature set $\mathcal{F}_{\text{test}}$ and logit set LOGITS serve as the inputs. The post-processor VIM is utilized due to its capability to leverage both features and LOGITS effectively. To align the data formats, the first K values of LOGITS are preserved and normalized using the softmax function, maintaining the original notation. We then modify LOGITS to yield the logit matrix LOGITS:

$$\mathrm{LOGITS}_i = \begin{cases} \frac{1}{K}\mathbf{1}_K, & \text{if } \arg\max_{i\in\mathcal{Y}} \textsc{Logits}_i = K+1, \\ \textsc{Logits}_i, & \text{else.} \end{cases} \tag{9}$$

Nevertheless, this approach is adaptable to other OOD detection methods, provided that LOGITS is consistently adjusted for the trainer and post-processor. Formally, we also give the pseudocode of GROD displayed in Algorithm 1.

Theoretical guarantee. A crucial aspect of using transformer networks for OOD detection is defining the limits of their OOD detection capabilities. Thus we incorporate OOD detection learning theory into transformer, including conditions for learnability (Theorem C.1) and error approximation of model budgets on transformers (Theorem C.3) in Appendix 2. Theorems are summarized informally below:

Theorem 2.1. *(Informal Theorem C.1, the equivalent conditions for OOD detection learnability on transformer networks) Given the condition $l(\mathbf{y}_2, \mathbf{y}_1) \le l(K+1, \mathbf{y}_1)$ for any in-distribution labels $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, and ID and OOD have no overlap, then there exists one transformer s.t. OOD detection is learnable, if and only if $|\mathcal{X}| = n < +\infty$. Furthermore, if $|\mathcal{X}| < +\infty$, $\exists\delta > 0$ and a transformer with block budget m and l layers, where $m = (\hat{d_0}, 2, 1, 1, 4)$ and $l = \mathcal{O}(\tau(1/\delta)^{(\hat{d_0}\tau)})$, or $m = (K+1)\cdot(2\tau(2\tau\hat{d_0} + 1), 1, 1, \tau(2\tau\hat{d_0}+1), 2\tau(2\tau\hat{d_0}+1))$ and $l = 2$ s.t. OOD detection is learnable.*

Theorem 2.2. *(Informal Theorem C.3, error boundary regarding the transformer's budget) Given the condition $l(\mathbf{y}_2, \mathbf{y}_1) \le l(K+1, \mathbf{y}_1)$, for any in-distribution labels $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, $|\mathcal{X}| = n < +\infty$ and $\tau > K+1$, and set $l = 2$ and $m = (2m_h + 1, 1, m_h, 2\tau\hat{d}_0 + 1, r)$. Using a*

linear classifier c, the probability of OOD detection learnability regarding data distribution $\mathbf{P}$ *defined in Definition C.2 has a lower bound* $\mathbf{P} \geq (1 - \frac{\eta}{|\mathcal{I}|\lambda_0}\tau^2 C_0(\frac{C_1}{m_h^{2\alpha-1}} + \frac{C_2}{r^\beta}(km_h)^\beta))^{(K+1)^{n+1}}$, *where* $C_0, C_1, C_2, \eta, \lambda_0, |\mathcal{I}|, \alpha, \beta$ *can be treated as constants.*

In Theorem 2.1, we establish the necessary and sufficient conditions for OOD detection learnability in transformers *i.e.* finite data and a higher penalty for OOD misclassification. We also discuss the constraints on transformer architecture regarding width and depth. Theorem 2.2 addresses scenarios where transformers' sizes do not meet these specifications, deriving a lower bound for the probability of learnability under similar conditions.

Real-world models are trained with finite data, hence satisfying the condition $l(\mathbf{y}_2, \mathbf{y}_1) \leq l(K+1, \mathbf{y}_1)$ from Theorems 2.1 and 2.2 ensures optimal OOD detection. However, standard cross-entropy loss often fails to meet the condition, according to which we introduce the ID-OOD binary classification loss function $\mathcal{L}_2$ to address this gap. Since training datasets without OOD cannot fully utilize $\mathcal{L}_2$, we propose a novel method to generate high-quality outliers. Therefore, we form the GROD algorithm, which enhances the generalization of transformers through fine-tuning, supported by our theoretical analysis. Notably, a trade-off between ID classification effectiveness and OOD detection capability exists associated with γ, as demonstrated in our ablation study (Appendix K) and experiments on Gaussian mixture datasets (Appendix G). More details on the theoretical analysis and experimental validation are available in Appendix B-G.

3. Experiments

In this section, we provide empirical evidence to validate the effectiveness of GROD across a range of real-world NLP and CV classification tasks and types of outliers. Furthermore, the ablation study and the visualization are displayed in Appendix K and Appendix L, respectively.

3.1. Experimental Setting

Models. In our primary experiments, we use GROD to strengthen the generalization capability of the ViT-B-16 model (Dosovitskiy et al., 2020), pre-trained on **ImageNet-1K** (Russakovsky et al., 2015), as the backbone for image classification tasks. For text classification, we use GROD to update the BERT base model (Devlin et al., 2018), which has been pre-trained using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). Models are fine-tuned without OOD exposure. Details on the training hyper-parameters are provided in Appendix J.1.

Datasets. For image classification tasks, we use four benchmark datasets *i.e* **CIFAR-10** (Krizhevsky et al., 2009), **CIFAR-100** (Krizhevsky et al., 2009), **Tiny ImageNet** (Le & Yang, 2015) and **SVHN** (Netzer et al., 2011). For text classification, we use the dataset **CLINC150** (Larson et al., 2019) for semantic shift, and datasets **IMDB** (Maas et al., 2011) and **Yelp** (Zhang et al., 2015) for background shift. Detailed dataset information can be found in Appendix J.2.

Evaluation metrics. We evaluate models using ID classification accuracy (ID ACC) and three metrics for binary ID-OOD classification: FPR@95 (F), AUROC (A), AUPR for ID test dataset AUPR_IN (I), and AUPR for OOD test dataset AUPR_OUT (O).

3.2. Main Results

Several prevalent methods are used as baselines for comparison, including MSP (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2017), VIM (Wang et al., 2022), GEN (Liu et al., 2023a), and ASH (Djurisic et al., 2022) which require only post-processing, and finetuning models G-ODIN (Hsu et al., 2020), NPOS (Tao et al., 2023), and CIDER (Ming et al., 2022b). All the baselines are offered in the OpenOOD benchmark (Zhang et al., 2023; Yang et al., 2022a;b; 2021; Bitterwolf et al., 2023).

Results for image classification. As discussed in Section 2, GROD employs LDA projection to generate inter-class OOD only when $B/K > 2$ to ensure the stability of the synthesized OOD. To evaluate the performance in both scenarios of $B/K > 2$ and $B/K \leq 2$, we use the training sets of **CIFAR-10** and **CIFAR-100** as ID data, respectively. When $B/K > 2$, the inclusion of both PCA and LDA projections enriches the information in OOD, not only creating virtual OOD around ID but also synthesizing it among ID categories. Correspondingly, the experimental results presented in Table 1 show that GROD surpasses other competitors, achieving SOTA performance across all evaluation metrics. On average, GROD reduces the FPR@95 from 9.41%, achieved by the most competitive method, to 0.12%, while enhancing the AUROC from 97.88% to 99.98%. In scenarios where $B/K \leq 2$, GROD, though not as superior as the LDA-based inter-class OOD generation, still yields competitive outcomes using only PCA, as evidenced in Table 2. Unlike the combined use of PCA and LDA, this approach falls short in capturing features of OOD data between categories. Consequently, GROD excels in detecting far-OOD data rather than near-OOD, as the PCA-reduced features are more similar to those of far-OOD data.

Results for text classification. Table 3 presents the results for text classification. As two ID datasets, **IMDB** and **CLINC150** have two and ten categories respectively, with $B/K > 2$ in both cases. Hence, both PCA and LDA projections are applied to these datasets. Notably, while popular

Table 1. Quantitative comparison with prevalent methods of the ID classification and OOD detection performance, where the backbone ViT-B-16 pre-trained with **ImageNet-1K** is employed. **CIFAR-10** is the ID Dataset and LDA projections are used for generating inter-class fake outliers. The red, blue and bold fonts denote Top 1,2,3 in ranking.

OOD Datasets		-	CIFAR-100				Tiny ImageNet				SVHN				Average			
Evaluate Metrics (%)		ID ACC↑	F↓	A↑	I↑	O↑	F↓	A↑	I↑	O↑	F↓	A↑	I↑	O↑	F↓	A↑	I↑	O↑
Baseline	MSP	96.16	29.31	91.70	92.70	90.28	21.21	94.05	95.54	92.04	15.39	95.11	92.72	97.56	21.97	93.62	93.65	93.29
PostProcess	ODIN	96.16	42.96	91.01	90.69	91.35	14.59	97.10	97.39	96.91	21.49	94.94	90.88	97.89	26.35	94.35	92.99	95.38
	VIM		21.59	95.43	**95.64**	95.38	8.52	98.39	98.68	98.14	3.26	99.39	98.61	99.78	**11.12**	**97.74**	**97.64**	**97.77**
	GEN		27.24	93.51	93.72	93.32	16.99	96.40	97.02	95.86	11.16	97.65	95.50	99.04	18.46	95.85	95.41	96.07
	ASH		26.48	93.64	93.70	93.46	16.87	96.41	96.99	95.87	9.79	98.19	96.55	99.26	17.71	96.08	95.75	96.20
Finetuning+ PostProcess	G-ODIN	95.56	82.60	70.76	68.21	72.86	64.97	83.05	83.88	83.58	62.42	89.48	68.61	95.81	70.00	81.10	73.57	84.08
	NPOS	**96.75**	**21.18**	**95.63**	95.46	**95.68**	15.33	96.85	97.20	96.47	**3.33**	**99.18**	**98.45**	**99.60**	13.28	97.22	97.04	97.25
	CIDER	96.98	14.13	96.99	96.98	96.97	**10.19**	**97.78**	**97.95**	**97.57**	3.91	98.86	98.17	99.41	9.41	97.88	97.70	97.98
	Ours	97.31	0.16	99.97	99.97	99.96	0.11	99.98	99.98	99.97	0.09	99.98	99.97	99.99	0.12	99.98	99.97	99.97

Table 2. Quantitative comparison with prevalent methods using only PCA projection for generating fake OOD data.

OOD Datasets		-	CIFAR-10				Tiny ImageNet				SVHN				Average			
Evaluate Metrics (%)		ID ACC↑	F↓	A↑	I↑	O↑	F↓	A↑	I↑	O↑	F↓	A↑	I↑	O↑	F↓	A↑	I↑	O↑
Baseline	MSP	84.34	71.11	77.17	75.37	77.56	51.34	84.15	86.55	78.08	49.58	82.07	71.41	91.97	57.34	81.13	77.78	82.54
PostProcess	ODIN	84.34	80.29	70.06	67.71	73.54	51.63	88.78	90.12	86.62	57.96	82.07	66.59	91.74	63.29	80.30	74.81	83.97
	VIM		54.97	85.42	84.62	85.71	30.22	92.30	94.69	88.43	23.02	93.93	88.69	**97.15**	**36.07**	90.55	89.33	90.43
	GEN		73.77	80.89	77.28	82.37	45.00	89.06	91.44	84.77	35.83	90.96	81.97	96.17	51.53	86.97	83.56	87.77
	ASH		75.26	80.61	76.87	82.19	44.68	88.98	91.42	84.62	35.87	90.88	81.85	96.12	51.94	86.82	83.38	87.64
Finetuning+ PostProcess	G-ODIN	61.40	89.14	47.52	51.63	47.76	74.07	68.87	77.48	54.99	30.77	93.15	95.55	89.40	64.66	69.85	74.89	64.05
	NPOS	**84.76**	43.53	89.63	89.14	90.42	**33.36**	**91.72**	**94.14**	**88.38**	38.86	90.62	81.67	96.04	38.58	**90.66**	88.32	91.61
	CIDER	84.87	**44.47**	89.41	88.74	90.23	33.08	91.83	94.18	88.60	**30.36**	**93.48**	84.46	97.36	35.97	91.57	**89.13**	92.06
	Ours	86.21	43.38	**88.00**	**88.01**	**87.94**	38.84	91.44	93.46	87.91	23.38	94.59	**87.88**	98.63	35.20	91.34	89.78	**91.49**

Table 3. Quantitative comparison with prevalent methods, where the pre-trained BERT is employed. Experimental results on two typical OOD in the text OOD detection, *i.e.* background shift OOD and semantic shift OOD are reported.

OOD Detection Type		Background Shift					Semantic Shift				
ID Datasets		**IMDB**					**CLINC150** with Intents				
OOD Datasets		**Yelp**					**CLINC150** with Unknown Intents				
Evaluate Metrics (%)		ID ACC↑	F↓	A↑	I↑	O↑	ID ACC↑	F↓	A↑	I↑	O↑
Baseline		91.36	**57.72**	74.28	73.28	74.60	97.78	37.11	92.31	97.70	74.66
PostProcess	VIM	91.36	64.00	**74.61**	70.17	76.05	97.78	29.33	93.58	98.03	80.99
	GEN		57.63	74.28	**73.28**	74.60		**36.27**	92.27	97.47	79.43
	ASH		73.27	71.43	65.11	**76.64**		40.67	**92.56**	**97.60**	**79.70**
Finetuning+ PostProcess	NPOS	90.36	76.31	68.48	61.84	74.56	95.62	49.89	83.57	95.64	48.52
	CIDER	**91.28**	59.71	78.10	75.09	79.07	95.93	45.04	86.39	96.44	55.17
	Ours	91.47	52.89	78.86	77.61	79.63	**97.66**	24.00	94.58	98.52	82.47

OOD detection algorithms are rigorously tested on image datasets, their effectiveness on text datasets does not exhibit marked superiority, as Table 3 illustrates. In addition, methods like ODIN (Liang et al., 2017) and G-ODIN (Hsu et al., 2020), which compute data gradients, necessitate floating-point number inputs. However, the tokenizer-encoded long integers used as input tokens in BERT create data format incompatibilities when attempting to use the BERT model alongside ODIN or G-ODIN. Given their marginal performance on images, they are excluded from text classification. In comparison, GROD consistently enhances model performance across image and text datasets, demonstrating remarkable versatility and wide application potential.

4. Discussion and future work

The universality and flexibility of GROD show more than across various data types. Its independence from and compatibility with various deep learning models and Outlier Exposure (OE) allows models to employ GROD and OE modules simultaneously. Despite its strengths, GROD has limitations. GROD is theoretically guaranteed only to a lower bound but not the infimum, indicating that GROD could be further enhanced with a more advanced theory of transformer expression approximation. Moreover, when handling datasets with numerous categories, GROD generates global outliers without fully exploiting inter-class data, which can potentially limit its effectiveness, calling for a stable method for generating OOD. We have reported representing experimental results across different data modalities and classification discussions, and continuing to test GROD on additional datasets and benchmarks will help validate its effectiveness in diverse settings.

5. Conclusion

In this paper, we introduce GROD, an algorithm designed to improve transformer generalization and enhance OOD detection capabilities. GROD builds on theoretical foundations, incorporating two theorems that establish conditions and error bounds for OOD detection in transformers to form fine-tuning strategies. It has shown superior performance in NLP and CV tasks regardless of data format. This research would suggest promising directions for future research into OOD detection, model generalization and safety.

References

Alberti, S., Dern, N., Thesing, L., and Kutyniok, G. Sumformer: Universal approximation for efficient transformers. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 72–86. PMLR, 2023.

Arora, U., Huang, W., and He, H. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*, 2021.

Bartlett, P. L. and Maass, W. Vapnik-chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*, pp. 1188–1192, 2003.

Bitterwolf, J., Mueller, M., and Hein, M. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023.

Cai, M. and Li, Y. Out-of-distribution detection via frequency-regularized generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5521–5530, 2023.

Chiang, D., Cholak, P., and Pillay, A. Tighter bounds on the expressivity of transformer encoders. In *International Conference on Machine Learning*, pp. 5544–5562. PMLR, 2023.

Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., and Vernekar, S. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DeVore, R., Hanin, B., and Petrova, G. Neural network approximation. *Acta numerica*, 30:327–444, 2021.

Djurisic, A., Bozanic, N., Ashok, A., and Liu, R. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Drummond, N. and Shearer, R. The open world assumption. In *eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web*, volume 15, pp. 1, 2006.

Fang, Z., Lu, J., Liu, A., Liu, F., and Zhang, G. Learning bounds for open-set learning. In *International conference on machine learning*, pp. 3122–3132. PMLR, 2021.

Fang, Z., Li, Y., Lu, J., Dong, J., Han, B., and Liu, F. Is out-of-distribution detection learnable? *Advances in neural information processing systems*, 35:37199–37213, 2022.

Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. *Advances in neural information processing systems*, 34:7068–7081, 2021.

Graham, M. S., Pinaya, W. H., Tudosiu, P.-D., Nachev, P., Ourselin, S., and Cardoso, J. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2947–2956, 2023.

Hahn, M. Theoretical limitations of self-attention in neural sequence models. *Transactions of the association for computational linguistics*, 8:156–171, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

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

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Hsu, Y.-C., Shen, Y., Jin, H., and Kira, Z. 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*, pp. 10951–10960, 2020.

Jiang, H. and Li, Q. Approximation theory of transformer networks for sequence modeling. *arXiv preprint arXiv:2305.18475*, 2023.

Jiang, W., Ge, Y., Cheng, H., Chen, M., Feng, S., and Wang, C. Read: Aggregating reconstruction error into out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14910–14918, 2023.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Koner, R., Sinhamahapatra, P., Roscher, K., Günnemann, S., and Tresp, V. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021.

Kratsios, A., Zamanlooy, B., Liu, T., and Dokmanić, I. Universal approximation under constraints is possible with transformers. *arXiv preprint arXiv:2110.03303*, 2021.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *Learning multiple layers of features from tiny images*, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., et al. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*, 2019.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

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

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Liu, X., Lochman, Y., and Zach, C. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23946–23955, 2023a.

Liu, Y., Ding, K., Liu, H., and Pan, S. Good-d: On unsupervised graph out-of-distribution detection. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 339–347, 2023b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Luo, S., Li, S., Zheng, S., Liu, T.-Y., Wang, L., and He, D. Your transformer may not be as powerful as you expect. *Advances in neural information processing systems*, 35: 4301–4315, 2022.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Merrill, W. and Sabharwal, A. A logic for expressing log-precision transformers. *Advances in neural information processing systems*, 36, 2024.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022a.

Ming, Y., Sun, Y., Dia, O., and Li, Y. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022b.

Morteza, P. and Li, Y. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7831–7840, 2022.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y., et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, Spain, 2011.

Ouyang, Y., Cao, Y., Gao, Y., Wu, Z., Zhang, J., and Dai, X. On prefix-tuning for lightweight out-of-distribution detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1533–1545, 2023.

Podolskiy, A., Lipin, D., Bout, A., Artemova, E., and Piontkovskaya, I. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13675–13682, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International journal of computer vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Strobl, L., Merrill, W., Weiss, G., Chiang, D., and Angluin, D. Transformers as recognizers of formal languages: A survey on expressivity. *arXiv preprint arXiv:2311.00208*, 2023.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Tao, L., Du, X., Zhu, X., and Li, Y. Non-parametric outlier synthesis. *arXiv preprint arXiv:2303.02966*, 2023.

Urysohn, P. Über die mächtigkeit der zusammenhängenden mengen. *Mathematische annalen*, 94(1):262–295, 1925.

Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in neural information processing systems*, 35: 32598–32611, 2022a.

Yang, J., Zhou, K., and Liu, Z. Full-spectrum out-of-distribution detection. *arXiv preprint arXiv:2204.05306*, 2022b.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Yun, C., Chang, Y.-W., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. O (n) connections are expressive enough: Universal approximability of sparse transformers. *Advances in neural information processing systems*, 33:13783–13794, 2020.

Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., Li, Y., Liu, Z., Chen, Y., and Li, H. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

Zhang, L., Goldstein, M., and Ranganath, R. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pp. 12427–12436. PMLR, 2021.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Zhou, W., Liu, F., and Chen, M. Contrastive out-of-distribution detection for pretrained transformers. *arXiv preprint arXiv:2104.08812*, 2021.

A. Related works

Application of OOD detection. The recent advancements in OOD detection models and algorithms have been significant (Sun et al., 2022; Liu et al., 2023b; Cai & Li, 2023). Typically, OOD detection methods leverage both post-processing techniques and training strategies, which can be implemented either separately or in combination (Zhang et al., 2023). Key post-processing techniques include the use of distance functions (Denouden et al., 2018), the development of scoring functions (Ming et al., 2022a), and the integration of disturbance terms (Hsu et al., 2020), among others. Several methods introduce training strategies for OOD detection models. For instance, Tao et al. (2023) suggests loss functions to facilitate the learning of compact representations, while Graham et al. (2023); Jiang et al. (2023) innovatively employ reconstruction models to pinpoint abnormal data. In addition, the transformer architecture has gained popularity in OOD detection, prized for its robust feature representation capabilities (Koner et al., 2021; Fort et al., 2021).

Theory of OOD detection. Theoretical research into OOD detection has recently intensified. Morteza & Li (2022) examines maximum likelihood on mixed Gaussian distributions and introduces a GEM log-likelihood score. Zhang et al. (2021) reveals that even minor errors in density estimation can result in OOD detection failures. Fang et al. (2022) presents the first application of Probably Approximately Correct (PAC) learning theory to OOD detection, deriving the Impossibility Theorem and exploring conditions under which OOD detection can be learned in previously unknown spaces. Moreover, Yang et al. (2021) has pioneered the concept of generalized OOD detection, noting its commonalities with anomaly detection (AD) and open set recognition (OSR) (Fang et al., 2021). To the best of our knowledge, no comprehensive theory of OOD detection for transformers has been established yet.

Transformers and their universal approximation power. We also conduct a literature survey on transformers and their approximation theory related to theoretical analysis. Transformers bring inspiration and progress to OOD detection, with algorithms utilizing their self-attention mechanism achieving noteworthy results (Koner et al., 2021; Hendrycks et al., 2020; Podolskiy et al., 2021; Zhou et al., 2021). Understanding the expressivity of transformers is vital for their application in OOD detection. Current research predominantly explores two main areas: formal language theory and approximation theory (Strobl et al., 2023). The former examines transformers as recognizers of formal languages, clarifying their lower and upper bounds (Hahn, 2020; Chiang et al., 2023; Merrill & Sabharwal, 2024). Our focus, however, lies primarily in approximation theory. The universal approximation property (UAP) of transformers, characterized by fixed width and infinite depth, was initially demonstrated by Yun et al. (2019). Subsequent studies have expanded on this, exploring UAP under various conditions and transformer architectures (Yun et al., 2020; Kratsios et al., 2021; Luo et al., 2022; Alberti et al., 2023). As another important development, Jiang & Li (2023) established the UAP for architectures with a fixed depth and infinite width and provided Jackson-type approximation rates for transformers.

B. Notations and preliminaries

Notations. More notations for theoretical analysis can be found here. $|\cdot|$ indicates the count of elements in a set, and $||\cdot||2$ represents the L_2 norm in Euclidean space. The data priori-unknown distribution spaces include $\mathcal{D}_{XY}^{all}$, which is the total space including all distributions; $\mathcal{D}_{XY}^{s}$, the separate space with distributions that have no ID-OOD overlap; $\mathcal{D}_{XY}^{D_{XY}}$, a single-distribution space for a specific dataset distribution denoted as D_{XY}; $\mathcal{D}_{XY}^{F}$, the Finite-ID-distribution space containing distributions with a finite number of ID examples; and $\mathcal{D}_{XY}^{\mu,b}$, the density-based space characterized by distributions expressed through density functions. A superscript may be added on D_{XY} to denote the number of data points in the distribution. The model hypothesis space is represented by $\mathcal{H}$, and the binary ID-OOD classifier is defined as Φ. These notations, consistent with those used in Fang et al. (2022), facilitate a clear understanding of OOD detection learning theory.

Several notations related to spaces and measures of function approximation also require further clarification to enhance understanding of the theoretical framework. $\mathcal{C}$ and C denote the compact function set and compact data set, respectively. Complexity measures for the self-attention blocks within transformers are denoted as $C_0(\cdot)$ and $C_1^{(\alpha)}(\cdot)$, while $C_2^{(\beta)}(\cdot)$ represents a regularity measure for the feed-forward neural networks within transformers. These measures indicate the approximation capabilities of transformers, with α and β being the convergence orders for Jackson-type estimation. $\widetilde{\mathcal{C}}^{(\alpha,\beta)}$ within $\mathcal{C}$ is the function space where Jackson-type estimation is applicable. Given the complexity of the mathematical definitions and symbols involved, we aim to provide clear conceptions to facilitate a smooth understanding of our theoretical approach. These mathematical definitions regarding function approximation follow those presented by Jiang & Li (2023).

Goal of theory. As an impressive work on OOD detection theory, Fang et al. (2022) defines strong learnability for OOD detection and has applied its PAC learning theory to the FCNN-based and score-based hypothesis spaces:

Definition B.1. (Fang et al., 2022)(Strong learnability) OOD detection is strongly learnable in $\mathcal{D}_{XY}$, if there exists an algorithm $\mathbf{A}$: $\cup_{n=1}^{+\infty}(\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H}$ and a monotonically decreasing sequence $\epsilon(n)$ *s.t.* $\epsilon(n)\to 0$, as $n\to+\infty$, and for any domain $D_{XY}\in\mathcal{D}_{XY}$,

$$\mathbb{E}_{S\sim D^n_{X_I Y_I}}[\mathcal{L}^{\alpha}_D(\mathbf{A}(S)) - inf_{h\in\mathcal{H}}\mathcal{L}^{\alpha}_D(h)] \le \epsilon(n), \forall\alpha\in[0,1].$$

In the data distribution spaces under our study, the equality of strong learnability and PAC learnability has been proved. So we only need to gain strong learnability to verify the proposed theorems.

Theorem B.2. *(Fang et al., 2022) (Informal, learnability in FCNN-based and score-based hypothesis spaces)*

If $l(y_2,y_1)\le l(K+1,y_1)$ for any in-distribution labels y_1 and $y_2\in\mathcal{Y}$, and the hypothesis space $\mathcal{H}$ is FCNN-based or corresponding score-based, then OOD detection is learnable in the separate space $\mathcal{D}^s_{XY}$ for $\mathcal{H}$ if and only if $|\mathcal{X}|<+\infty$.

Inspired by Theorem B.2 and its Proof, the goal of our theory is proposed as follows:

> Goal: Given a transformer hypothesis space $\mathcal{H}_{\mathrm{TOOD}}$, what are necessary or sufficient conditions to ensure the learnability of OOD detection? Furthermore, we try to derive the approximation rates and error bounds of OOD detection.

The transformer hypothesis space. Under the goal of investigating the OOD detection learning theory on transformers, our research defines a fixed transformer hypothesis space for OOD detection $\mathcal{H}_{\mathrm{tood}}$. A transformer block $\mathrm{Block}(\cdot):\mathbb{R}^{\hat{d}\times\tau}\to\mathbb{R}^{\hat{d}\times\tau}$ consists of a self-attention layer $\mathrm{Att}(\cdot)$ and a feed-forward layer $\mathrm{FF}(\cdot)$, *i.e.*

$$\mathrm{Att}(\mathbf{h}_l) = \mathbf{h}_l + \sum_{i=1}^{h} W_O^i W_V^i \mathbf{h}_l \cdot \sigma[(W_K^i\mathbf{h}_l)^T W_Q^i \mathbf{h}_l], \tag{10}$$

$$\mathbf{h}_{l+1} = \mathrm{FF}(\mathbf{h}_l) = \mathrm{Att}(\mathbf{h}_l) + W_2\cdot\mathrm{Relu}(W_1\cdot\mathrm{Att}(\mathbf{h}_l) + \mathbf{b_1}\mathbf{1}_n^T) + \mathbf{b_2}\mathbf{1}_n^T, \tag{11}$$

with $W_O^i\in\mathbb{R}^{\hat{d}\times m_v}$, $W_V^i\in\mathbb{R}^{m_V\times\hat{d}}$, $W_K^i, W_Q^i\in\mathbb{R}^{m_h\times\hat{d}}$, $W_1\in\mathbb{R}^{r\times\hat{d}}$, $W_2\in\mathbb{R}^{\hat{d}\times r}$, $b_1\in\mathbb{R}^r$ and $b_2\in\mathbb{R}^{\hat{d}}$. Besides, $\mathbf{h}_l\in\mathbb{R}^{\hat{d}\times\tau}$ is the hidden state of l-th transformer block with $\mathbf{h}_0\in\mathbb{R}^{\hat{d}\times\tau}$ is the input data $\mathcal{X}\in\mathbb{R}^{(\hat{d_0}\times\tau)\times n}$ (with position encoding) after a one-layer FCNN $F:\mathbb{R}^{\hat{d_0}\times\tau}\to\mathbb{R}^{\hat{d}\times\tau}$, and $\sigma(\cdot)$ is the column-wise softmax function. We denote $d:=\hat{d}\times\tau$ and $d_0:=\hat{d_0}\times\tau$ for convenience.

The computation budget of a transformer block includes the number of heads h, the hidden layer size r of FF, m_h, m_V, and n, denoted by $m=(\hat{d},h,m_h,m_V,r)$. Formally, a classic transformer block with a budget of m and l-th layer can be depicted as $\mathrm{Block}_l^{(m)}(\cdot)=\mathrm{FF}\circ\mathrm{Att}(\cdot)$. Transformer is a composition of transformer blocks, by which we define transformer hypothesis space $\mathcal{H}_{\mathrm{Trans}}$:

Definition B.3. (Transformer hypothesis space) The transformer hypothesis space is $\mathcal{H}_{\mathrm{Trans}}$ is

$$\mathcal{H}_{\mathrm{Trans}} = \cup_l \mathcal{H}^{(l)}_{\mathrm{Trans}} = \cup_l \cup_m \mathcal{H}^{(l,m)}_{\mathrm{Trans}} \tag{12}$$

where $\mathcal{H}^{(l)}_{\mathrm{Trans}}$ is the transformer hypothesis space with l layers, and $\mathcal{H}^{(l,m)}_{\mathrm{Trans}}$ is the transformer hypothesis space with l layers of $\mathrm{Block}_i^{(m)}(\cdot)$, $i\in\{1,2,\ldots,l\}$. More specifically,

$$\mathcal{H}^{(l,m)}_{\mathrm{Trans}} := \{\hat{H}:\hat{H}=\mathrm{Block}_l^{(m)}\circ\mathrm{Block}_{l-1}^{(m)}\circ\cdots\circ\mathrm{Block}_1^{(m)}\circ F\}. \tag{13}$$

By the Definition B.3 that $\forall\hat{H}\in\mathcal{H}_{\mathrm{Trans}}$, $\hat{H}$ is a map from $\mathbb{R}^{d_0\times n}$ to $\mathbb{R}^{d\times n}$. To match the OOD detection task, we insert a classifier $c:\mathbb{R}^d\to\mathcal{Y}$ applied to each data as follows:

Definition B.4. (Classifier) $c:\mathbb{R}^d\to\mathcal{Y}$ is a classical classifier with forms:

$$\text{(maximum value) } c(\mathbf{h}_l) = \arg\max_{k\in\mathcal{Y}} f^k(\mathbf{h}_l), \tag{14}$$

$$\text{(score-based) } c(\mathbf{h}_l) = \begin{cases} K+1, & E(f(\mathbf{h}_l)) < \lambda, \\ \arg\max_{k \in \mathcal{Y}} f^k(\mathbf{h}_l), & E(f(\mathbf{h}_l)) \geq \lambda, \end{cases} \tag{15}$$

where f^k is the k-th coordinate of $f \in \{\hat{f} \in \mathbb{R}^d \to \mathbb{R}^{K+1}\}$, which is defined by

$$f^k(\mathbf{h}_l) = W_{4,k}(W_{3,k}\mathbf{h}_l + b_{3,k})^T + b_{4,k}. \tag{16}$$

$W_{3,k} \in \mathbb{R}^{1\times \hat{d}}$, $W_{4,k}, b_{3,k} \in \mathbb{R}^{1\times\tau}$ and $b_{4,k} \in \mathbb{R}$. And $E(\cdot)$ is the scoring functions like softmax-based function (Hendrycks & Gimpel, 2016) and energy-based function (Liu et al., 2020).

Now, we can naturally derive the Definition of transformer hypothesis space for OOD detection:

Definition B.5. (Transformer hypothesis space for OOD detection)

$$\mathcal{H}_{\text{tood}} := \{H \in \mathbb{R}^{d_0\times n} \to \mathcal{Y}^n : H = c \circ \hat{H}, c \text{ is a classifier in Definition B.4}, \hat{H} \in \mathcal{H}_{\text{Trans}}\} \tag{17}$$

Similarly, we denote $\mathcal{H}^{(l)}_{\text{tood}}$ and $\mathcal{H}^{(l,m)}_{\text{tood}}$ as in Definition B.3.

C. Theoretical results of OOD detection using transformers

In the five priori-unknown spaces defined in Fang et al. (2022), **the total space** $\mathcal{D}^{\text{all}}_{XY}$ has been thoroughly discussed. Following the impossible Theorem, OOD detection is NOT learnable due to the overlapping of datasets, even when the budget $m \to +\infty$. Consequently, we shift our focus to the learning theory of transformers within the other four spaces: $\mathcal{D}^{D_{XY}}_{XY}$, $\mathcal{D}^{s}_{XY}$, $\mathcal{D}^{F}_{XY}$, and $\mathcal{D}^{\mu,b}_{XY}$. For each space, we investigate whether OOD detection is learnable under $\mathcal{H}\text{tood}$, considering the specific constraints, conditions, or assumptions. If OOD detection is found to be learnable, we then explore the approximation of rates and boundaries to further understand the generalization capabilities of transformers.

C.1. OOD detection in the separate space

Since the overlap is a crucial factor preventing models from successfully learning OOD detection, a natural perspective is to explore the separate space $\mathcal{D}^{s}_{XY}$.

Conditions for learning with transformers. Firstly, by Theorem 10 in Fang et al. (2022) and Theorems 5, 8 in Bartlett & Maass (2003), OOD detection is NOT learnable in $\mathcal{D}^{s}_{XY}$. It means OOD detection also has the impossible Theorem in $\mathcal{D}^{s}_{XY}$ for $\mathcal{H}_{\text{tood}}$. So we enquire about the conditions for $\mathcal{H}_{\text{tood}}$ to meet the requirements of learnability, deriving Theorem C.1:

Theorem C.1. *(Necessary and sufficient condition for OOD detection learnability on transformers)*

Given the condition $l(\mathbf{y}_2, \mathbf{y}_1) \leq l(K+1, \mathbf{y}_1)$, for any in-distribution labels $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, then OOD detection is learnable in the separate space $\mathcal{D}^{s}_{XY}$ for $\mathcal{H}_{\text{tood}}$ if and only if $|\mathcal{X}| = n < +\infty$. Furthermore, if $|\mathcal{X}| < +\infty$, $\exists \delta > 0$ and $g \in \mathcal{H}^{(m,l)}_{\text{tood}}$, where $Block(\cdot)$ budget $m = (\hat{d}_0, 2, 1, 1, 4)$ and the number of $Block(\cdot)$ layer $l = \mathcal{O}(\tau(1/\delta)^{(\hat{d}_0\tau)})$, or $m = (K+1)\cdot(2\tau(2\tau\hat{d}_0+1), 1, 1, \tau(2\tau\hat{d}_0+1), 2\tau(2\tau\hat{d}_0+1))$ and $l = 2$ s.t. OOD detection is learnable with g.

Theorem C.1 gives the necessary and sufficient conditions for OOD detection learnability on transformers with a fixed depth or width. Detailed proof and remarks on inspection can be found in the Appendix D.

Approximation of rates and boundaries. To further investigate the approximation rates and boundaries as the budget m grows, we formulate Jackson-type estimates for OOD detection learnability using transformer models. Before presenting the main Theorem C.3, it is essential to define the probability associated with the learnability of OOD detection:

Definition C.2. (Probability of the OOD detection learnability) Given a domain space $\mathcal{D}_{XY}$ and the hypothesis space $\mathcal{H}^{(m,l)}_{\text{tood}}$, $D'^{n}_{XY} \subset D^{n}_{XY} \in \mathcal{D}_{XY}$ is the distribution that for any dataset $\mathcal{X} \sim D'^{n}_{X_I Y_I}$, OOD detection is learnable, where D^{n}_{XY} is any distribution in $\mathcal{D}_{XY}$ with data amount n. The probability of the OOD detection learnability is defined by

$$\mathbf{P} := \varliminf_{D^{n}_{XY}\in\mathcal{D}_{XY}} \varlimsup_{D'^{n}_{XY}\subset D^{n}_{XY}} \frac{\mu(D'^{n}_{XY})}{\mu(D^{n}_{XY})}, \tag{18}$$

where μ is the Lebesgue measure in $\mathbb{R}^d$ and $n \in \mathbb{N}^*$.

Then the main Theorem C.3 of the Jackson-type approximation is formally depicted:

Theorem C.3. *Given the condition $l(\mathbf{y}_2, \mathbf{y}_1) \leq l(K+1, \mathbf{y}_1)$, for any in-distribution labels $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, $|\mathcal{X}| = n < +\infty$ and $\tau > K+1$, and set $l = 2$ and $m = (2m_h + 1, 1, m_h, 2\tau\hat{d}_0 + 1, r)$. Then in $\mathcal{H}_{\text{tood}}^{(m,l)}$ restricted to maximum value classifier c, $\mathbf{P} \geq (1 - \frac{\eta}{|\mathcal{I}|\lambda_0}\tau^2 C_0(r_i)(\frac{C_1^{(\alpha)}(r_i)}{m_h^{2\alpha-1}} + \frac{C_2^{(\beta)}(r_i)}{r^\beta}(km_h)^\beta))^{(K+1)^{n+1}}$, and in $\mathcal{H}_{\text{tood}}^{(m,l)}$ restricted to score-based classifier c, $\mathbf{P} \geq (1 - \frac{\eta}{|\mathcal{I}|\lambda_0}\tau^2 C_0(r_i)(\frac{C_1^{(\alpha)}(r_i)}{m_h^{2\alpha-1}} + \frac{C_2^{(\beta)}(r_i)}{r^\beta}(km_h)^\beta))^{(K+1)^{n+1}+1}$, for any fixed $\lambda_0 > 0$ and r_i defined in Lemma E.4, if $\{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) \geq \lambda\}$ and $\{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) < \lambda\}$ both contain an open ball with the radius R, where $R > ||W_4||_2|\mathcal{I}|(\tau^2 C_0(\phi)(\frac{C_1^{(\alpha)}(\phi)}{m_h^{2\alpha-1}} + \frac{C_2^{(\beta)}(\phi)}{r^\beta}(km_h)^\beta) + \lambda_0)$, ϕ defined in Lemma E.5 and W_4 is determined by ϕ.*

The proof structure leverages the Jackson-type approximation of transformers, as detailed in Jiang & Li (2023), to fulfill one of the sufficient conditions for OOD detection learnability *i.e.* Theorem 7 in Fang et al. (2022). Notably, the Jackson-type approximation of transformers has a global error bound instead of the uniform convergence in UAP theory. Therefore, Markov's inequality is applied to get probabilistic conclusions regarding Definition C.2. This approach establishes a lower bound of error and its convergent rate for OOD detection using transformers. The lower bound is not the infimum because the Jackson-type approximation is sufficient but not necessary. The specific proof and discussion about the convergent rate and insights into transformers are organized in Appendix E.

C.2. OOD detection in other priori-unknown spaces

In **the single-distribution space** $\mathcal{D}_{XY}^{D_{XY}}$, **the Finite-ID-distribution space** $\mathcal{D}_{XY}^{F}$, and **the density-based space** $\mathcal{D}_{XY}^{\mu,b}$, if there exists an overlap between ID and OOD, OOD detection is NOT learnable, which has been discussed in Fang et al. (2022); otherwise, $\mathcal{D}_{XY}^{D_{XY}} \subset \mathcal{D}_{XY}^{s}$, this situation is analyzed in the previous Appendix C.1. Additionally, in **the density-based space** $\mathcal{D}_{XY}^{\mu,b}$, Theorem 9 and Theorem 11 in Fang et al. (2022) are still established in the hypothesis space $\mathcal{H}_{\text{tood}}$, as the proof of these two theorems only need to check the finite Natarajan dimension (Shalev-Shwartz & Ben-David, 2014) of the hypothesis space, which is a weaker condition compared with the finite VC dimension.

Theorem C.1 demonstrates that in $\mathcal{H}_{\text{tood}}$, models are OOD detection learnable given sufficient parameters, thereby providing a theoretical basis for employing transformers in OOD detection algorithms (Koner et al., 2021; Fort et al., 2021). Nevertheless, training models to reach their optimal state poses significant challenges. To overcome these issues, additional strategies such as incorporating extra data (Fort et al., 2021; Tao et al., 2023) and utilizing various distance metrics (Podolskiy et al., 2021) have been developed. Detailed discussions on Gaussian mixture datasets, which explore the discrepancy between theoretical performance and practical outcomes and suggest ways to bridge this gap, can be found in Appendices F and G.

D. Proof and remarks of Theorem C.1

We first propose several Lemmas before proving the Theorem C.1.

Lemma D.1. *The FCNN-based hypothesis space $\mathcal{H}_{\mathbf{q}}^{\text{Relu}} \subseteq \mathcal{H}_{\text{tood}}^{(m,l)}$, where $\mathbf{q} = (l_1, \cdots, l_g), l_1 = d_0, l_g = K+1, l_M = max\{l_1, \cdots, l_g\}, m = (l_M, 1, 1, 1, l_M)$, and $l = g - 3, g > 2$.*

Proof. Given weights $\mathbf{w}_i \in \mathbb{R}^{l_i \times l_{i-1}}$ and bias $\mathbf{b}_i \in \mathbb{R}^{l_i \times 1}$ and considering $\mathbf{x} = \mathbf{h}_0 \in \mathbb{R}^{d_0}$ is a data in the dataset $\mathcal{X}$, the i-layer output of FCNN with architecture $\mathbf{q}$ can be written as

$$\mathbf{f}_i(\mathbf{x}) = \text{Relu}(\mathbf{w}_i \mathbf{f}_{i-1}(\mathbf{x}) + \mathbf{b}_i), \tag{19}$$

and that of the transformer network $\mathbf{H} = c \circ \text{Block}_l^{(m)} \circ \text{Block}_{l-1}^{(m)} \circ \cdots \circ \text{Block}_1^{(m)} \circ F$ in the transformer hypothesis space for OOD detection $\mathcal{H}_{\text{tood}}$ is depicted by Eq. (10) and (11). Particularly, set $W_O^i = \mathbf{0}$, and $m = (l_M, 1, 1, 1, l_M)$, then we get

$$\mathbf{h}_i = \mathbf{h}_i + W_2 \cdot \text{Relu}(W_1 \cdot \mathbf{h}_{i-1} + \mathbf{b}_1) + \mathbf{b}_2, \tag{20}$$

where $\mathbf{h}_i, \mathbf{h}_{i-1}, \mathbf{b}_1 \in \mathcal{R}^{l_M}, W_1, W_2 \in \mathcal{R}^{l_M \times l_M}$. Since $\mathbf{H}$ is composed of l blocks and mappings c at the bottom and F at the top as two layers, a simple case is when $g = 3$, it comes that $l = 0$, $\mathcal{H}_{\text{tood}}^{(m,l)}$ collapse into $\mathcal{H}_{\mathbf{q}'}^{\text{Relu}}$, where $\mathbf{q}' = (l_1, l_M, l_g), l_M = \max\{l_1, l_g\}$. So $\mathcal{H}_{\mathbf{q}}^{\text{Relu}} \subseteq \mathcal{H}_{\mathbf{q}'}^{\text{Relu}}$ according to Lemma 10 in Fang et al. (2022).

When $g > 3$, consider $F(\cdot) : \mathbb{R}^{d_0 \times n} \to \mathbb{R}^{l_M \times n}$, $F(\mathbf{x}) = \mathrm{Relu}(W\mathbf{x}+\mathbf{b})$ column-wise and the first layer of the FCNN-based network $f_1 : \mathbb{R}^{l_1} \to \mathbb{R}^{l_2}$, $f_1(\mathbf{x}) = \mathrm{Relu}(\omega_1\mathbf{x} + \beta_\mathbf{1})$. Since $l_M = max\{l_1, \cdots, l_g\}$, $l_2 \leq l_M$. Let

$$W = [\omega_1, \mathbf{0}]^T,\ \mathbf{b} = [\beta_\mathbf{1}, \mathbf{0}]^T, \tag{21}$$

then $F(\mathbf{x}) = [f_1(\mathbf{x}), \mathbf{0}]^T$. Similarly, we compare $f_i = \mathrm{Relu}(\omega_i f_{i-1}(\mathbf{x}) + \beta_\mathbf{i})$ and Block_{i-2}. Suppose that $h_{i-3} = [f_{i-1}(\mathbf{x}), \mathbf{0}]^T$, let $\mathbf{b}_2 = -h_{i-3}, \mathbf{b}_1 = [\beta_i, \mathbf{0}]^T, W_2 = Id_{l_M \times l_M}$ and

$$W_1 = \begin{bmatrix} \omega_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{22}$$

then it is clear that $h_{i-2} = [f_i(\mathbf{x}), \mathbf{0}]^T$.

By mathematical induction, it follows that $h_{g-3} = [f_{g-1}(\mathbf{x}), \mathbf{0}]^T$ and $f(h_{g-3}) = f_g(\mathbf{x})$, f is defined in Definition B.4. Therefore, for any entry $h_{\mathbf{w},\mathbf{b}} \in \mathcal{H}^{\mathrm{Relu}}_{\mathbf{q}}$, there exists $\mathbf{H} \in \mathcal{H}^{(m,l)}_{\mathrm{tood}}$, m, l defined in the Lemma *s.t.* $\mathbf{H} = h_{\mathbf{w},\mathbf{b}}$. □

Lemma D.2. *For any $\mathbf{h} \in \mathcal{C}(\mathbb{R}^d, \mathbb{R}^{K+1})$, and any compact set $C \in \mathbb{R}^d$, $\epsilon > 0$, there exists a two layer transformer $\hat{\mathbf{H}} \in \mathcal{H}^{(m,2)}_{\mathrm{Trans}}$ and a linear transformation $\mathbf{f}$ s.t. $||\mathbf{f} \circ \hat{\mathbf{H}} - \mathbf{h}||_2 < \epsilon$ in C, where $m = (K+1) \cdot (2\tau(2\tau\hat{d_0}+1), 1, 1, \tau(2\tau\hat{d_0}+1), 2\tau(2\tau\hat{d_0}+1))$.*

Proof. Let $\mathbf{h} = [h_1, \cdots, h_{K+1}]^T$. Based on the UAP of transformers *i.e.* Theorem 4.1 in Jiang & Li (2023), for any $\epsilon > 0$, there exists $\hat{h_i} = \hat{f}_i \circ \bar{H}_i$, where $\hat{f}_i$ is a linear read out and $\bar{H}_i \in \mathcal{H}^{(\hat{m},2)}_{\mathrm{Trans}}, \hat{m} = 2\tau(2\tau\hat{d_0}+1), 1, 1, \tau(2\tau\hat{d_0}+1), 2\tau(2\tau\hat{d_0}+1)$ *s.t.*

$$\max_{\mathbf{x}\in C} ||\hat{h_i}(\mathbf{x}) - h_i(\mathbf{x})||_1 < \epsilon/\sqrt{K+1}, i = 1, 2, \cdots, K+1. \tag{23}$$

We need to construct a transformer network $\hat{\mathbf{H}} \in \mathcal{H}^{(m,2)}_{\mathrm{Trans}}$ and a linear transformation $\mathbf{f}$ *s.t.*

$$(\mathbf{f} \circ \hat{\mathbf{H}})_i = \hat{f}_i \circ \bar{H}_i \tag{24}$$

for all $i \in \{1, \cdots, K+1\}$. The following shows the process of construction:

Denote the one-layer FCNN in $\bar{H}_i$ by $F_i : \mathcal{R}^{d_0 \times n} \to \mathcal{R}^{D \times n}$, where $D = 2n(2nd_0 + 1)$, the set the one-layer FCNN in $\hat{\mathbf{H}}$:

$$\begin{aligned} F : \mathcal{R}^{d_0 \times n} &\to \mathcal{R}^{D(K+1) \times n}, \\ F &= [F_1, \cdots, F_{K+1}]^T, \end{aligned} \tag{25}$$

then $\mathbf{h}_0 = [h_0^1, \cdots, h_0^{K+1}]^T$, where $\mathbf{h}_0$ is the input to transformer blocks in $\hat{\mathbf{H}}$, and h_0^i is that in $\bar{H}_i, i = 1, \cdots, K+1$.

Denote the matrices in $\bar{H}_i$ by $\bar{W}^i_K, \bar{W}^i_Q, \bar{W}^i_V$ and $\bar{W}^i_O$ since each block only has one head. For the i-th head in each block of transformer network $\hat{\mathbf{H}}$, we derive the matrix $W^i_k \in \mathcal{R}^{(K+1)\hat{m}_h \times (K+1)D}$ from $\bar{W}^i_K$ with $\hat{m}_h = 1$:

$$W^i_K = \begin{bmatrix} \mathbf{0}_{(i-1)\hat{m}_h \times (i-1)D} & & \\ & \bar{W}^i_K & \\ & & \mathbf{0}_{(K+1-i)\hat{m}_h \times (K+1-i)D} \end{bmatrix}. \tag{26}$$

Furthermore, we obtain W^i_Q, W^i_V and W^i_O in the same way, then independent operations can be performed on different blocks in the process of computing the matrix $\mathrm{Att}(\mathbf{h}_0) \in \mathcal{R}^{(K+1)D \times n}$. So we can finally get the attention matrix in the following form:

$$\mathrm{Att}(\mathbf{h}_0) = [\mathrm{Att}_1(\mathbf{h}_0), \cdots \mathrm{Att}_{K+1}(\mathbf{h}_0)]^T, \tag{27}$$

where $\mathrm{Att}_i(\mathbf{h}_0) \in \mathcal{R}^{D \times n}, i \in \mathcal{Y}_I + 1$ are attention matrices in $\bar{H}_i$.

Similarly, it is easy to select $W_1, W_2, \mathbf{b}_1, \mathbf{b}_2$ such that $\mathrm{FF}(\mathbf{h}_0) = [\mathrm{FF}_1(\mathbf{h}_0), \cdots \mathrm{FF}_{K+1}(\mathbf{h}_0)]^T$, *i.e.* $\mathbf{h}_1 = [h_1^1, \cdots, h_1^{K+1}]^T$, where the meaning of superscripts resembles to that of h_0^i. Repeat the process, we found that

$$\hat{\mathbf{H}}(\mathcal{X}) = [\bar{H}_1(\mathcal{X}), \cdots \bar{H}_{K+1}(\mathcal{X})]^T. \tag{28}$$

Denote $\hat{f}_i(\bar{H}_i) = w_i \cdot \bar{H}_i + b_i$, then it is natural to construct the linear transformation $\mathbf{f}$ by:

$$\mathbf{f}(\hat{\mathbf{H}}) = [w_1, \cdots, w_{K+1}]^T \cdot \hat{\mathbf{H}} + [b_1, \cdots, b_{K+1}]^T, \tag{29}$$

which satisfies Eq. (24).

By Eq. (23), for any $\epsilon > 0$, there exists $\hat{\mathbf{H}} \in \mathcal{H}_{\text{Trans}}^{(m,2)}$ and the linear transformation $\mathbf{f}$ *s.t.*

$$\begin{aligned} \max_{\mathbf{x}\in C} ||\mathbf{f} \circ \hat{\mathbf{H}} - \mathbf{h}||_2 &\leq \sqrt{\sum_{i=1}^{K+1} (\max_{\mathbf{x}\in C} ||\hat{h_i}(\mathbf{x}) - h_i(\mathbf{x})||_1)^2} \\ &< \sqrt{\sum_{i=1}^{K+1} (\epsilon/\sqrt{K+1})^2} = \epsilon, \end{aligned} \tag{30}$$

where $m = (K+1) \cdot \hat{m}$.

We have completed this Proof. □

Then we prove the proposed Theorem C.1.

Proof. **First**, we prove the sufficiency. By the proposed Lemma D.1 and Theorem 10 in Fang et al. (2022), the sufficiency of Theorem C.1 is obvious.

Furthermore, according to the Proof of Theorem 10 in Fang et al. (2022), to replace the FCNN-based or score-based hypothesis space by the transformer hypothesis space for OOD detection $\mathcal{H}_{\text{tood}}$, the only thing we need to do is to investigate the UAP of transformer networks *s.t.* the UAP of FCNN network *i.e.* Lemma 12 in Fang et al. (2022) can be replaced by that of transformers. Moreover, it is easy to check Lemmas 13$\tilde{1}$6 in Fang et al. (2022) still holds for $\mathcal{H}_{\text{tood}}$. So following the Proof of Theorem 10 in Fang et al. (2022), by Theorem 3 in Yun et al. (2019) and the proposed Lemma D.2, we can obtain the needed layers l and specific budget m which meet the conditions of the learnability for OOD detection tasks.

Second, we prove the necessity. Assume that $|\mathcal{X}| = +\infty$. By Theorems 5, 8 in Bartlett & Maass (2003), $\text{VCdim}(\Phi \circ \mathcal{H}_{\text{tood}}^{(m,l)}) < +\infty$ for any m, l, where Φ maps ID data to 1 and maps OOD data to 2. Additionally, $\sup_{h\in\mathcal{H}_{\text{tood}}^{(m,l)}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| = +\infty$ given $|\mathcal{X}| = +\infty$ for any m, l. By the impossibility Theorem 5 for separate space in Fang et al. (2022), OOD detection is NOT learnable for any finite $m,\ l$. □

Remark D.3. Yun et al. (2019) and Jiang & Li (2023) provide two perspectives of the capacity of transformer networks. The former gives the learning conditions of OOD detection with limited width (or budget of each block) and any depth of networks, and the letter develops the learning conditions with limited depth.

Remark D.4. Define a partial order for the budget m: for $m = (d, h, m_h, m_V, r)$ and $m' = (d', h', m'_h, m'_V, r')$, $m' < m$ if every element in m' is less than the corresponding element in m. $m' \leq m$ if if every element in m' is not greater than the corresponding element in m. So it easily comes to a corollary: $\forall m'$ satisfies $m \leq m'$ and $l \leq l'$, if transformer hypothesis space $\mathcal{H}_{\text{tood}}^{(m,l)}$ is OOD detection learnable, then $\mathcal{H}_{\text{tood}}^{(m',l')}$ is OOD detection learnable.

Remark D.5. It is notable that when $m = +\infty$ or $l = +\infty$, $\text{VCdim}(\Phi \circ \mathcal{H}_{\text{tood}}^{(m,l)})$ may equal to $+\infty$. This suggests the possibility of achieving learnability in OOD detection without the constraint of $|\mathcal{X}| < +\infty$. Although an infinitely capacitated transformer network does not exist in reality, exploring whether the error asymptotically approaches zero as capacity increases remains a valuable theoretical inquiry.

E. Proof and remarks of Theorem C.3

To derive the Theorem C.3, we need to figure out some Lemmas.

Lemma E.1. *For any $\mathbf{h} \in \widetilde{\mathcal{C}}^{(\alpha,\beta)}$, and any compact set $C \in \mathbb{R}^d$, there exists a two layer transformer $\hat{\mathbf{H}} \in \mathcal{H}_{\text{Trans}}^{(m,2)}$ and a linear read out $\mathbf{c} : \mathbb{R}^{\hat{d}\times\tau} \to \mathbb{R}^{1\times\tau}$ s.t. the inequality (34) is established, where $m = (2m_h + 1, 1, m_h, 2\tau\hat{d}_0 + 1, r)$.*

Proof. According to Theorem 4.2 in Jiang & Li (2023), for any $\mathbf{h} \in \widetilde{\mathcal{C}}^{(\alpha,\beta)}$, there exists $\mathbf{H} \in \mathcal{H}^{(m,2)}_{\text{Trans}}$ and a linear read out $\mathbf{c}$ *s.t.*

$$\int_{\mathcal{I}} \sum_{t=1}^{\tau} |\mathbf{c} \circ \mathbf{H}_t(\mathbf{x}) - \mathbf{h}_t(\mathbf{x})| d\mathbf{x} \leq \tau^2 C_0(\mathbf{h}) (\frac{C_1^{(\alpha)}(\mathbf{h})}{m_h^{2\alpha-1}} + \frac{C_2^{(\beta)}(\mathbf{h})}{m_{\text{FF}}^{\beta}} (m_h)^{\beta}). \tag{31}$$

Based on Chebyshev's Inequality,

$$P(\sum_{t=1}^{\tau} |\mathbf{c} \circ \mathbf{H}_t(\mathbf{x})_i - \mathbf{h}_t(\mathbf{x})_i| / |\mathcal{I}| > \text{RHS in Eq. (31)}) + \lambda_0) \leq \frac{\text{RHS in Eq. (31)}}{\lambda_0 |\mathcal{I}|} \tag{32}$$

for any $\lambda_0 > 0$. Additionally,

$$\begin{aligned} ||\mathbf{c} \circ \mathbf{H}(\mathbf{x}) - \mathbf{h}(\mathbf{x})||_2 &= \sqrt{\sum_{t=1}^{\tau} |\mathbf{c} \circ \mathbf{H}_t(\mathbf{x})_i - \mathbf{h}_t(\mathbf{x})_i|^2} \\ &\leq \sum_{t=1}^{\tau} |\mathbf{c} \circ \mathbf{H}_t(\mathbf{x})_i - \mathbf{h}_t(\mathbf{x})_i|. \end{aligned} \tag{33}$$

So we get

$$P(||\mathbf{c} \circ \mathbf{H}(\mathbf{x}) - \mathbf{h}(\mathbf{x})||_2 > |\mathcal{I}| (\text{RHS in Eq. (31)} + \lambda_0) \leq \frac{\text{RHS in Eq. (31)}}{\lambda_0 |\mathcal{I}|} \tag{34}$$

where m_{FF} is usually determined by its number of neurons and layers. As the number of layers in FF is fixed, the budget m_{FF} and r are proportional:

$$r = k \cdot m_{\text{FF}}. \tag{35}$$

So the right side of the equation (31) can be written as

$$\text{RHS} = \tau^2 C_0(\mathbf{h}) (\frac{C_1^{(\alpha)}(\mathbf{h})}{m_h^{2\alpha-1}} + \frac{C_2^{(\beta)}(\mathbf{h})}{r^{\beta}} (k m_h)^{\beta}). \tag{36}$$

We have completed this Proof of the Lemma E.1. □

Given any finite δ hypothesis functions $h_1, \cdots, h_\delta \in \{\mathcal{X} \to \mathcal{Y}\}$, for each h_i, we introduce a correspongding $\mathbf{g}_i$ (defined over $\mathcal{X}$) satisfying that for any $\mathbf{x} \in \mathcal{X}$, $\mathbf{g}_i(\mathbf{x}) = \mathbf{y}_k$ and $W_4 \mathbf{g}_i^T + b_4 = \mathbf{z}_k$ if and only if $h_i(\mathbf{x}) = k$, where $\mathbf{z}_k \in \mathbb{R}^{K+1}$ is the one-hot vector corresponding to the label k with value N. Clearly, $\mathbf{g}_i$ is a continuous mapping in $\mathcal{X}$, because $\mathcal{X}$ is a discrete set. Tietze Extension Theorem (Urysohn, 1925) implies that $\mathbf{g}_i$ can be extended to a continuous function in $\mathbb{R}^d$. If $\tau \geq K + 1$, we can find such $\mathbf{g}_i, W_4, b_4$.

Lemma E.2. *For any introduced $\mathbf{g}_i$ mentioned above, there exists $\hat{\mathbf{g}}_i$ satisfies $\hat{\mathbf{g}}_i \in \widetilde{\mathcal{C}}^{(\alpha,\beta)}$ and $||\hat{\mathbf{g}}_i - \mathbf{g}_i||_2 < \epsilon$.*

Proof. Based on Theorem 7.4 in DeVore et al. (2021), set $G \equiv 0$ and $\rho \equiv 0$, then $\hat{\mathbf{g}}_i \in \widetilde{\mathcal{C}}^{(\alpha,\beta)}$, and there exists a constant C, *s.t.* $||\hat{\mathbf{g}}_i - \mathbf{g}_i||_2 < \frac{C}{(r+1)^{\beta}}$.

Choose r which is great enough, the proof is completed. □

Remark E.3. Note that we can also prove the same result if $\mathbf{g}_i$ is any continuous function from $\mathbb{R}^{\hat{d}}$ to $\mathbb{R}$ with compact support.

Lemma E.4. *Let $|\mathcal{X}| = n < +\infty$, $\tau > K + 1$ and σ be the* Relu *function. Given any finite δ hypothesis functions $h_1, \cdots, h_\delta \in \{\mathcal{X} \to \{1, \cdots, K+1\}\}$, then for any $m_h, r > 0$, $m = (2m_h + 1, 1, m_h, 2\tau\hat{d}_0 + 1, r)$, $P(h_1, \cdots, h_\delta \in \mathcal{H}^{(m,2)}_{\text{tood}}) \geq (1 - \frac{m\text{RHS in Eq. (31)}}{|\mathcal{I}|\lambda_0})^{(K+1)\delta}$ for any $\eta > 1$.*

Proof. Since $\mathcal{X}$ is a compact set, then Lemma E.2 implies that there exists $\hat{\mathbf{g}}_i \in \widetilde{\mathcal{C}}^{(\alpha,\beta)}$ *s.t.*

$$||\mathbf{g}_i - \hat{\mathbf{g}}_i||_2 < \epsilon/||W_4||_2. \tag{37}$$

Denote $r_i = W_4\mathbf{g}_i^T + b_4$ and $\hat{r}_i = W_4\hat{\mathbf{g}}_i^T + b_4$, So we get

$$||r_i - \hat{r}_i||_2 = ||W_4(\mathbf{g}_i - \hat{\mathbf{g}}_i)^T||_2 \leq \epsilon. \tag{38}$$

Then by Lemma E.1, there exists $\hat{\mathbf{H}} \in \mathcal{H}_{\text{Trans}}^{(m,2)}$ and a linear read out $\mathbf{c}$ *s.t.*

$$P(||\mathbf{c} \circ \mathbf{H}(\mathbf{x}) - \mathbf{h}(\mathbf{x})||_2 \leq |\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0) \geq 1 - \frac{\text{RHS in Eq. (31)}}{\lambda_0|\mathcal{I}|}. \tag{39}$$

Thus we get if $h_i(\mathbf{x}) = k$, which is equal to $\mathbf{g}_i(\mathbf{x}) = \mathbf{y}_k$ or $r_i(\mathbf{x}) = \mathbf{z}_k$:

Firstly, denote $\mathbf{f} = W_4\mathbf{c} \circ \mathbf{H}^T + b_4$, and let $\mathbf{h} = \hat{\mathbf{g}}_i$, then

$$P(||\mathbf{f}(\mathbf{x}) - \hat{r}_i(\mathbf{x})||_2 \leq ||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0) \geq 1 - \frac{\text{RHS in Eq. (31)}}{\lambda_0|\mathcal{I}|}. \tag{40}$$

So we obtain that

$$\begin{aligned}
&P(|\mathbf{f}_k - N| \leq ||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0) \\
&\geq P(|\mathbf{f}_k - \hat{r}_{i,k}| + |\hat{r}_{i,k} - r_{i,k}| \leq ||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0)) \\
&\geq P(||\mathbf{f} - \hat{r}_i||_2 + ||\hat{r}_i - r_i||_2 \leq ||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0)) \\
&\geq P(||\mathbf{f} - \hat{r}_i||_2 + \epsilon \leq ||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0)) \\
&= P(||\mathbf{f} - \hat{r}_i||_2 \leq ||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + (\lambda_0 - \frac{\epsilon}{|\mathcal{I}|}))) \\
&\geq 1 - \frac{\text{RHS in Eq. (31)}}{|\mathcal{I}|(\lambda_0 - \frac{\epsilon}{|\mathcal{I}|})} \\
&= 1 - \frac{\text{RHS in Eq. (31)}}{|\mathcal{I}|\lambda_0 - \epsilon}.
\end{aligned} \tag{41}$$

Similarly, for any $j \neq k$, we can also obtain that

$$P(|\mathbf{f}_k| \leq ||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0) \geq 1 - \frac{\text{RHS in Eq. (31)}}{|\mathcal{I}|\lambda_0 - \epsilon}. \tag{42}$$

Therefore, $P(\arg\max_{k\in\mathcal{Y}} \mathbf{f}_k(\mathbf{x}) = h_i(\mathbf{x})) \geq (1 - \frac{\eta\text{RHS in Eq. (31)}}{|\mathcal{I}|\lambda_0})^{K+1}$ for any $\mathbf{x}$, if

$$N > 2||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0) \tag{43}$$

for any $\eta > 1$, *i.e.*

$$P(h_1, \cdots, h_\delta \in \mathcal{H}_{\text{tood}}^{(m,2)}) \geq (1 - \frac{\eta\text{RHS in Eq. (31)}}{|\mathcal{I}|\lambda_0})^{(K+1)\delta}, \tag{44}$$

if

$$N > 2||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0) \tag{45}$$

for any $\eta > 1$. Since N is arbitrary, we can find such N. □

Lemma E.5. *Let the activation function σ be the Relu function. Suppose that $|\mathcal{X}| < +\infty$, and $\tau > K + 1$. If $\{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) \geq \lambda\}$ and $\{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) < \lambda\}$ both contain an open ball with the radius $R > ||W_4||_2|\mathcal{I}|(RHS\ in\ Eq.\ (31)(\phi) + \lambda_0)$, the probability of introduced binary classifier hypothesis space $\mathcal{H}_{\text{tood},E}^{(m,2),\lambda}$ consisting of all binary classifiers $P > (1 - \frac{\eta RHS\ in\ Eq.\ (31)}{|\mathcal{I}|\lambda_0})^{(K+1)\delta+1}$, where $m = (2m_h + 1, 1, m_h, 2\tau\hat{d}_0 + 1, r)$ and $\phi(\mathbf{x})$ is determined by centers of balls, specifically defined in the proof and W_4 is determined by $\phi(\mathbf{x})$.*

Proof. Since $\{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) \geq \lambda\}$ and $\{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) < \lambda\}$ both contain an open ball with the radius $R \geq ||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0)$, we can find $\mathbf{v}_1 \in \{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) \geq \lambda\}$, $\mathbf{v}_2 \in \{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) < \lambda\}$ *s.t.* $B_R(\mathbf{v}_1) \subset \{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) \geq \lambda\}$ and $B_R(\mathbf{v}_2) \subset \{\mathbf{v} \in \mathbb{R}^{K+1} : E(\mathbf{v}) < \lambda\}$, where $B_R(\mathbf{v}_1) := \{\mathbf{v} : ||\mathbf{v} - \mathbf{v}_1||_2 < R\}$ and $B_R(\mathbf{v}_2) := \{\mathbf{v} : ||\mathbf{v} - \mathbf{v}_2||_2 < R\}$.

For any binary classifier h over $\mathcal{X}$, we can induce a vector-valued function as follows. For any $\mathbf{x} \in \mathcal{X}$,

$$\phi(\mathbf{x}) = \begin{cases} \mathbf{v}_1, & \text{if } h(\mathbf{x}) = 1, \\ \mathbf{v}_2, & \text{if } h(\mathbf{x}) = 2. \end{cases} \tag{46}$$

Since $\mathcal{X}$ is a finite set, the Tietze Extension Theorem implies that ϕ can be extended to a continuous function in $\mathbb{R}^d$. Since $\mathcal{X}$ is a compact set, then Lemma E.1 and Lemma E.2 implies that there exists a two layer transformer $\mathbf{H} \in \mathcal{H}_{\text{Trans}}^{(m,2)}$ and f defined in B.4 s.t for any $\eta > 1$,

$$P(||\mathbf{f} \circ \mathbf{H}(\mathbf{x}) - \phi(\mathbf{x})||_2 \leq ||W_4||_2|\mathcal{I}|(\text{RHS in Eq. (31)} + \lambda_0) \geq 1 - \frac{\text{RHS in Eq. (31)}}{|\mathcal{I}|\lambda_0 - \epsilon} \tag{47}$$

Therefore, for any $\mathbf{x} \in \mathcal{X}$, it is easy to check that $E(\mathbf{f} \circ \mathbf{H}(\mathbf{x})) \geq \lambda$ if and only if $h(\mathbf{x}) = 1$, and $E(\mathbf{f} \circ \mathbf{H}(\mathbf{x})) < \lambda$ if and only if $h(\mathbf{x}) = 2$ if the condition in $P(\cdot)$ is established.

Since $|X| < +\infty$, only finite binary classifiers are defined over $\mathcal{X}$. By Lemma E.4, we get

$$P(\mathcal{H}_{\text{all}}^b = \mathcal{H}_{\text{tood},E}^{(m,2),\lambda}) \geq (1 - \frac{\eta\text{RHS in Eq. (31)}}{|\mathcal{I}|\lambda_0})^{(K+1)\delta+1} \tag{48}$$

The proof is completed. □

Now we prove one of the main conclusions *i.e.* Theorem C.3, which provides a sufficient Jackson-type condition for learning of OOD detection in $\mathcal{H}_{\text{tood}}$.

Proof. **First, we consider the case that c is a maximum value classifier.** Since $|\mathcal{X}| < +\infty$, it is clear that $|\mathcal{H}_{\text{all}}| < +\infty$, where $\mathcal{H}_{\text{all}}$ consists of all hypothesis functions from $\mathcal{X}$ to $\mathcal{Y}$. For $|\mathcal{X}| < +\infty$ and $\tau > K + 1$, according to Lemma E.4, $P(\mathcal{H}_{\text{all}} \subset \mathcal{H}_{\text{tood}}^{(m,2)}) \geq (1 - \frac{\eta\text{RHS in Eq. (31)}}{|\mathcal{I}|\lambda_0})^{(K+1)\delta}$ for any $\eta > 1$, where $m = (2m_h+1, 1, m_h, 2nd+1, r)$ and $\delta = (K+1)^n$.

Consistent with the proof of Lemma 13 in Fang et al. (2022), we can prove the correspondence Lemma 13 in the transformer hypothesis space for OOD detection if $\mathcal{H}_{\text{all}} \subset \mathcal{H}_{\text{tood}}^{(m,2)}$, which implies that there exist $\mathcal{H}^{\text{in}}$ and $\mathcal{H}^{\text{b}}$ *s.t.* $\mathcal{H}_{\text{tood}}^{(m,2)} \subset \mathcal{H}^{\text{in}} \circ \mathcal{H}^{\text{b}}$, where $\mathcal{H}^{\text{in}}$ is for ID classification and $\mathcal{H}^{\text{b}}$ for ID-OOD binary classification. So it follows that $\mathcal{H}_{\text{all}} = \mathcal{H}_{\text{tood}}^{(m,2)} = \mathcal{H}^{\text{in}} \circ \mathcal{H}^{\text{b}}$. Therefore, $\mathcal{H}_{\text{b}}$ contains all binary classifiers from $\mathcal{X}$ to $\{1, 2\}$. According to Theorem 7 in (Fang et al., 2022), OOD detection is learnable in $\mathcal{D}_{XY}^s$ for $\mathcal{H}_{\text{tood}}^{(m,2)}$.

Second, we consider the case that c is a score-based classifier. It is easy to figure out the probability of which OOD detection is learnable based on Lemma E.5 and Theorem 7 in Fang et al. (2022).

The proof of Theorem C.3 is completed. □

Remark E.6. Approximation of α: First of all, it is definitely that $\alpha > \frac{1}{2}$ to maintain the conditions in Theorem 4.2 of Jiang & Li (2023). Then, analyze the process of our proof, because of the powerful expressivity of Relu, we only need $G \equiv 0$ to bridge from $\mathcal{C}$ to $\widetilde{\mathcal{C}}^{(\alpha,\beta)}$. So with regard to $\mathcal{H}_{\text{tood}}$, any $\alpha > \frac{1}{2}$ satisfies all conditions. But C_1^α can increase dramatically when α get greater.

Remark E.7. Approximation of β: We denote $\beta \in (0, \beta_{\max}]$. According to Theorem 7.4 in DeVore et al. (2021), $\beta_{\max} \in [1, 2]$.

Remark E.8. By the approximation of α and β, we discuss the trade-off of expressivity and the budget of transformer models. Firstly, the learnability probability $P \to 1$ if and only if $m_h \to +\infty$ and $\frac{r}{m_h} \to +\infty$. For a fixed r, there exists a m_h which achieves the best trade-off. For a fixed m_h, the greater r is, the more powerful the expressivity of transformer models is.

Remark E.9. Different scoring functions E have different ranges. For example, $\max_{k\in\{1,\cdots K\}} \frac{e^{v^k}}{\sum_{c=1}^{K+1} e^{v^c}}$ and $T\log\sum_{c=1}^{K} e^{(\frac{v^c}{T})}$ have ranges contain $(\frac{1}{K+1},1)$ and $(0,+\infty)$, respectively. The Theorem C.3 gives the insight that the domain and range of scoring functions should be considered when dealing with OOD detection tasks using transformers.

Remark E.10. It can be seen from Theorem C.3 that the complexity of the data increases, and the scale of the model must also increase accordingly to ensure the same generalization performance from the perspective of OOD detection. Increasing the category K of data may exponentially reduce the learnable probability of OOD detection, while increasing the amount of data n reduces the learnable probability much more dramatically. Using Taylor expansion for estimation,

$$\begin{aligned}
&(1-\frac{\eta}{|\mathcal{I}|\lambda_0}\tau^2 C_0(r_i)(\frac{C_1^{(\alpha)}(r_i)}{m_h^{2\alpha-1}}+\frac{C_2^{(\beta)}(r_i)}{r^\beta}(km_h)^\beta))^{(K+1)^{n+1}}\\
&=1-(K+1)^{n+1}\frac{\eta}{|\mathcal{I}|\lambda_0}\tau^2 C_0(r_i)(\frac{C_1^{(\alpha)}(r_i)}{m_h^{2\alpha-1}}+\frac{C_2^{(\beta)}(r_i)}{r^\beta}(km_h)^\beta)\\
&+\mathcal{O}((\frac{\eta}{|\mathcal{I}|\lambda_0}\tau^2 C_0(r_i)(\frac{C_1^{(\alpha)}(r_i)}{m_h^{2\alpha-1}}+\frac{C_2^{(\beta)}(r_i)}{r^\beta}(km_h)^\beta))^2)
\end{aligned} \tag{49}$$

for any $\frac{\eta}{|\mathcal{I}|\lambda_0}\tau^2 C_0(r_i)(\frac{C_1^{(\alpha)}(r_i)}{m_h^{2\alpha-1}}+\frac{C_2^{(\beta)}(r_i)}{r^\beta}(km_h)^\beta)<1$. To ensure generalization, increasing the data category K requires a polynomial increase of model parameters; while increasing the amount of data n requires an exponential increase of model parameters. The data with positional coding $\mathcal{X}$ is contained in $\mathcal{I}$. The greater $\mathcal{I}$ is, the more possibility transformers have of OOD detection learnability. Nevertheless, the scoring function needs to meet a stronger condition of R. Theorem C.3 indicates that large models are guaranteed to gain superior generalization performance.

Remark E.11. This theorem has limitations for not determining the exact optimal convergence order and the infimum of the error. More research on function approximation theory would be helpful to develop it in-depth.

F. The gap between theoretical existence and training OOD detection learnable models

We first show the key problems that intrigue the gap by conducting experiments on generated datasets. The specific experiments are described as follows.

F.1. Basic dataset generation

We generated Gaussian mixture datasets consisting of two-dimensional Gaussian distributions. The expectations μ^i and the covariance matrices Σ^i are randomly generated respectively, $i=1,2$ *i.e.* $K=2$:

$$\begin{aligned}
\mu^i &= \frac{i}{10}[|\mathcal{N}(0,1)|,|\mathcal{N}(0,1)|]^T,\\
\Sigma^i &= \begin{bmatrix}\sigma_1^i & 0\\ 0 & \sigma_2^i\end{bmatrix},\ \text{where } \sigma_j^i=\frac{i}{10}|\mathcal{N}(0,1)|+0.1,\ j=1,2,
\end{aligned} \tag{50}$$

and the data whose Euclidean distance from the expectation is greater than 3σ is filtered to construct the separate space. Further, we generated another two-dimensional Gaussian distribution dataset, and also performed outlier filtering operations as OOD data with the expectation μ^O and the covariance matrix Σ^O as

$$\begin{aligned}
\mu^O &= \frac{1}{2}[-|\mathcal{N}(0,1)|,-|\mathcal{N}(0,1)|]^T,\\
\Sigma^O &= \begin{bmatrix}\sigma_1^O & 0\\ 0 & \sigma_2^O\end{bmatrix},\ \text{where } \sigma_j^O=0.2|\mathcal{N}(0,1)|+0.1.
\end{aligned} \tag{51}$$

Formally, the distribution of the generated dataset can be depicted by

$$D_X=\frac{1}{3}(\mathcal{N}(\mu^1,\Sigma^1)+\mathcal{N}(\mu^2,\Sigma^2)+\mathcal{N}(\mu^O,\Sigma^O)) \tag{52}$$

as the quantity of each type of data is almost the same. A visualization of the dataset with a fixed random seed is shown in Figure 2(a).

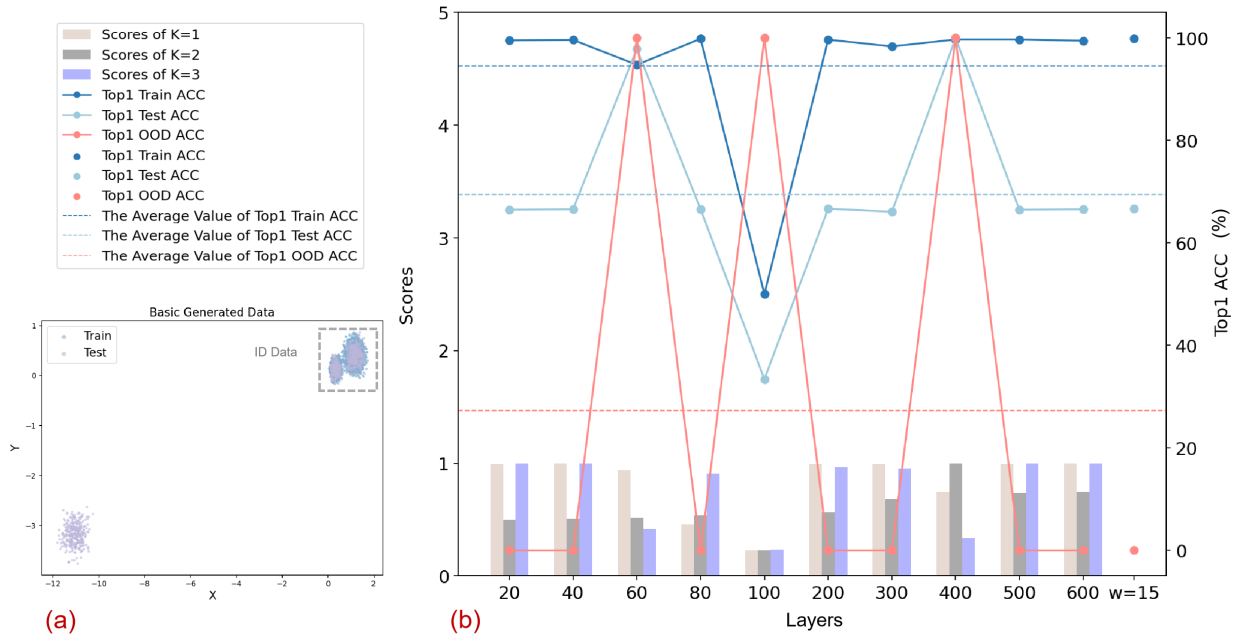

Figure 2. (a) The visualization of the generated two-dimensional Gaussian mixture dataset. (b) Curves show the classification accuracy and OOD detection accuracy of the training stage and test stage with different model budgets. And likelihood score bars demonstrate that the model with the theoretical support is disabled to learn OOD characters, leading to the failure of OOD detection.

F.2. Model construction and gap illustration

We constructed the transformer models strictly following the hypothesis space definition B.5, where $\hat{d}_0 = \hat{d} = 2$ and $\tau = 1$. Our experimental results are shown in Figure 4(b). According to Theorem C.1, in $\mathcal{H}^{(m,l)}_{\text{tood}}$, where $m = (2, 2, 1, 1, 4)$ and l is sufficiently large, or $l = 2$, $m = (2w, 1, 1, w, 2w)$, where $w := \tau(2\tau\hat{d}_0 + 1) = 15$, OOD detection can be learned. Since Theorem C.1 does not give a specific value for l, so we choose a wide range of l for experiments. Figure 4(b) shows that even for a very simple Gaussian mixture distribution dataset, transformer models without additional algorithm design can classify ID data with high accuracy in most cases, but can not correctly classify OOD data, showing severe overfitting and strong bias to classify OOD data into ID categories. By chance, transformers with some l just converge to a learnable state. We have also selected the scoring function $E(f(\mathbf{h}_l)) = max_{k\in\{1,\cdots K\}} \frac{e^{f(\mathbf{h}_l)^k}}{\sum_{c=1}^{K+1} e^{f(\mathbf{h}_l)^c}}$ and visualized the scoring function values for every category by the trained models. It can be seen that in a model that cannot identify OOD data, using the score-based classifier c also can not distinguish the OOD data.

G. Details of optimization and validation on generated datasets

In this section, we analyze the causes of training failures and introduce an algorithm designed to address these challenges. We used five different random seeds for data generation for each dataset type discussed later. The experimental outcomes are illustrated in Figure 3.

G.1. Optimization 1

First of all, considering that the classical cross-entropy loss $\mathcal{L}_1$ does not satisfy the condition $l(\mathbf{y}_2, \mathbf{y}_1) \leq l(K+1, \mathbf{y}_1)$, for any in-distribution labels $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, and there is no instruction for model to learn recognizing OOD data, an additional loss $\mathcal{L}_2$ is added in the loss function:

$$\mathcal{L} = (1-\gamma)\mathcal{L}_1 + \gamma\mathcal{L}_2, \tag{53}$$

$$\mathcal{L}_1(\mathbf{y}, \mathbf{x}) = -\mathbb{E}_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^{K+1} \mathbf{y}_j \log(\text{softmax}(\mathbf{f} \circ \mathbf{H}(\mathbf{x}))_j), \tag{54}$$

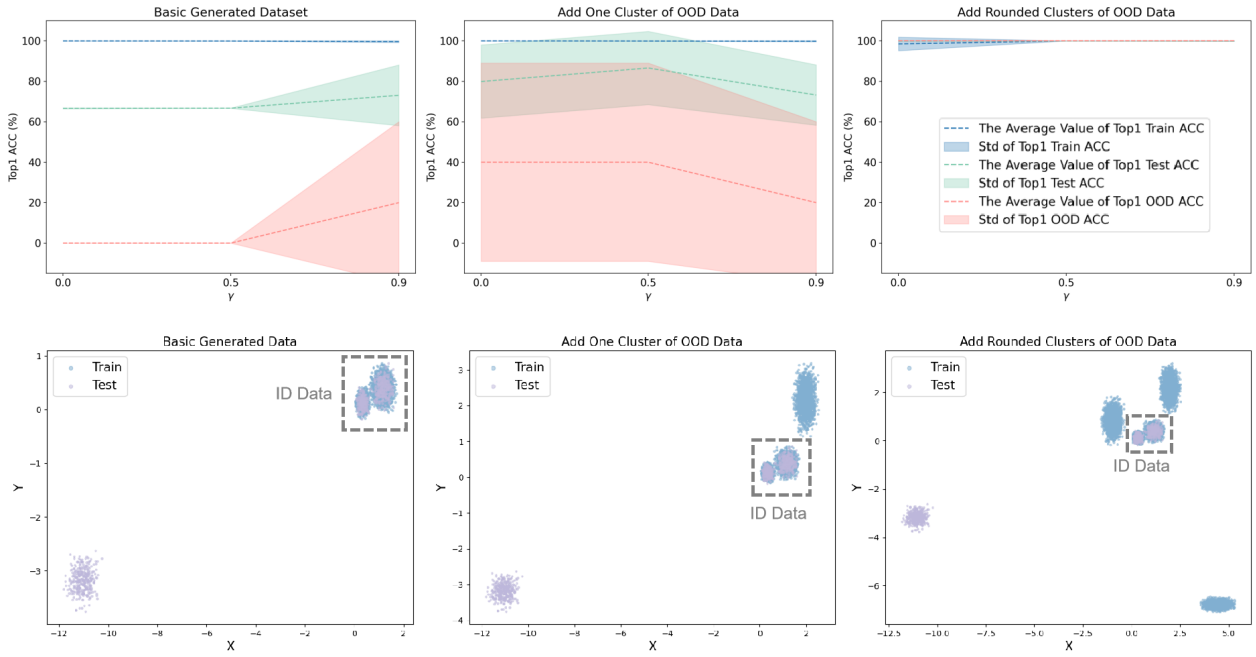

Figure 3. The classification and OOD detection results in the optimization process. The first row of subfigures is the results of experiments under different OOD data distributions. The scAtter plots below show the corresponding training and test set data. The trade-off of loss function $\mathcal{L}$ is shown when picking different γ, and the power of adding rounded OOD data is illustrated with perfect performance in the third column.

$$\mathcal{L}_2(\mathbf{y}, \mathbf{x}) = -\mathbb{E}_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^{2} \hat{\phi}(\mathbf{y})_j \log(\hat{\phi}(\text{softmax}(\mathbf{f} \circ \mathbf{H}(\mathbf{x})))_j), \tag{55}$$

where $\mathbf{H} \in \mathcal{H}_{\text{Trans}}$, $\mathbf{y}$ is the one-hot label vector, $\hat{\phi} : \mathbb{R}^{K+1} \to \mathbb{R}^2$ is depicted as follows:

$$\hat{\phi}(\mathbf{y}) = \begin{bmatrix} \sum_{i=1}^{K} \mathbf{y}_i \\ \mathbf{y}_{K+1} \end{bmatrix}. \tag{56}$$

When the condition is satisfied, the classification loss sensitivity of ID data classification decreases, affecting the classification performance of ID data. Therefore, it is qualitatively evident that the value of γ has a trade-off between the performance of ID data classification and OOD data recognition.

G.2. Optimization 2

Selecting $\gamma = 0.0, 0.5, 0.9$, we observe a nuanced trade-off illustrated in the basic generated dataset (see the first column of Figure 3), as the model will classify ID data randomly, achieving only 50% accuracy in both training and testing phases, if $\gamma = 1$. Modifying the loss function merely increases the probability that the model can learn from OOD data but does not ensure stable training for achieving high-performance OOD detection. This limitation arises because when the model accurately classifies ID data, the value of $\mathbf{f} \circ \mathbf{H}(\mathbf{x})_{K+1}$ remains small, rendering $\mathcal{L}_2$ almost ineffective during training and impeding the model's ability to distinguish between ID and OOD. Without OOD data in the training set, the model tends to classify all test set data as ID. To address these issues, we explore the generation of virtual OOD data. Our experiments, shown in the middle column of Figure 3, indicate that creating a single cluster of virtual OOD data markedly enhances the OOD detection capabilities of transformers, while also illustrating the trade-offs associated with the parameter γ as analysed in Section 2. However, challenges persist in situations where the model correctly classifies ID data but fails to identify OOD data during training. To further enhance performance, we generate three clusters of OOD data surrounding the ID data. As demonstrated in the right column of Figure 3, enriching the content of virtual OOD data enables the model to consistently learn ID classification and extend its generalization to OOD data. Adding rounded clusters of OOD

data significantly diminishes the influence of $\mathcal{L}_2$, emphasizing the importance of generating high-quality fake OOD data. Considering the high dimensionality of most datasets and the challenges of delineating high-dimensional ID data boundaries in Euclidean space due to the curse of dimensionality, we retain the binary loss $\mathcal{L}_2$ in our algorithm.

Our experimental results are also consistent with recent research. For example, Fort et al. (2021) shows that incorporating outlier exposure significantly improves the OOD detection performance of transformers, and Tao et al. (2023) has presented a method for synthesizing OOD data using boundary data from KNN clusters.

H. The filtration process

Specifically, Mahalanobis distance from a sample $\mathbf{x}$ to the distribution of mean μ and covariance Σ is defined as $\text{Dist}(\mathbf{x}, \mu, \Sigma) = (\mathbf{x} - \mu)\Sigma^{-1}(\mathbf{x} - \mu)^T$. To ensure robust computations, the inverse matrix of Σ is calculated with numerical techniques. Firstly, we add a regularization term with small perturbation to Σ, *i.e.* $\Sigma' = \Sigma + \epsilon_0 I_d$, where $\epsilon_0 = 10^{-4}$ and I_d is the identity matrix. Given that Σ' is symmetric and positive definite, the Cholesky decomposition technique is employed whereby $\Sigma' = L \cdot L^T$. L is a lower triangular matrix, facilitating an efficient computation of the inverse $\Sigma^{-1} = (L^{-1})^T \cdot L^{-1}$. Then we filter $\hat{\mathcal{F}}_{\text{OOD}}$ by Mahalanobis distances. The average distances from ID data to their global and inter-class centers *i.e.* $\text{Dist}^{\text{ID}}_{\text{PCA}}$ and $\text{Dist}^{\text{ID}}_{\text{LDA},i_k}$ respectively are obtained by

$$\begin{aligned} \text{Dist}^{\text{ID}}_{\text{PCA}} &= \frac{1}{|\mathcal{F}|} \sum_{v \in \mathcal{F}} \text{Dist}(v, \mu_{\text{PCA}}, \text{cov}(\mathcal{F})), \\ \text{Dist}^{\text{ID}}_{\text{LDA},i_k} &= \frac{1}{|\mathcal{F}|_{\mathbf{y}=i_k}|} \sum_{v \in \mathcal{F}|_{\mathbf{y}=i_k}} \text{Dist}(v, \mu_{\text{LDA},i_k}, \text{cov}(\mathcal{F}|_{\mathbf{y}=i_k})), \end{aligned} \tag{57}$$

where $\text{cov}(\cdot)$ is the operator to calculate the covariance matrix of samples $\mathcal{F}$. In the meanwhile, Mahalanobis distances between OOD and ID are calculated:

$$\text{Dist}^{\text{OOD}}(v) = \begin{cases} \text{Dist}(v, \mu_{\text{PCA}}, \text{cov}(\mathcal{F})), & \text{if } B/K \leq 2, \\ \min\limits_{i_k \in I} \text{Dist}(v, \mu_{\text{LDA},i_k}, \text{cov}(\mathcal{F}|_{\mathbf{y}=i_k})), & \text{if } B/K > 2. \end{cases} \tag{58}$$

And if $B/K > 2$, $i_0 = i_0(v) = \arg\min_{i_k \in I} \text{Dist}(v, \mu_{\text{LDA},i_k}, \text{cov}(\mathcal{F}|_{\mathbf{y}=i_k}))$ is also recorded. The set to be deleted $\mathcal{F}_D$ is

$$\mathcal{F}_D = \begin{cases} \{v \in \hat{\mathcal{F}}_{\text{OOD}} : \text{Dist}^{\text{OOD}}(v) < (1+\Lambda)\text{Dist}^{\text{ID}}_{\text{PCA}}\}, & \text{if } B/K \leq 2, \\ \{v \in \hat{\mathcal{F}}_{\text{OOD}} : \text{Dist}^{\text{OOD}}(v) < (1+\Lambda)\text{Dist}^{\text{ID}}_{\text{LDA},i_0}\}, & \text{if } B/K > 2, \end{cases} \tag{59}$$

where $\Lambda = \lambda \cdot \frac{10}{|\hat{\mathcal{F}}_{\text{OOD}}|} \sum_{v \in \hat{\mathcal{F}}_{\text{OOD}}} (\frac{\text{Dist}^{\text{OOD}}(v)}{\text{Dist}^{\text{ID}}} - 1)$, λ is a learnable parameter with the initial value 0.1. $\text{Dist}^{\text{ID}} = \text{Dist}^{\text{ID}}_{\text{PCA}}$ if $B/K \leq 2$, else $\text{Dist}^{\text{ID}} = \text{Dist}^{\text{ID}}_{\text{LDA},i_0(v)}$. Additionally, we randomly filter the remaining OOD data to no more than $[B/K]+2$, and the filtered set is denoted as $\mathcal{F}_{RD}$. In this way, we obtain the final generated OOD set $\mathcal{F}_{\text{OOD}} := \hat{\mathcal{F}}_{\text{OOD}} - \mathcal{F}_D - \mathcal{F}_{RD}$, with the label $\mathbf{y} = K + 1$.

I. The pseudocode of GROD

The pseudocode of GROD is shown in Alg. 1.

J. Implementation details

J.1. Settings for the fine-tuning stage.

For image classification, we finetune the ViT backbone and GROD model with hyper-parameters as follows: epoch number $= 20$, batch size $= 64$, and the default initial learning rate = 1×10^{-4}. We set parameters $\alpha = 1 \times 10^{-3}$ for PCA and LDA projection and $\alpha = 0.1$ for PCA projection, $num = 1$, and $\gamma = 0.1$. An AdamW (Kingma & Ba, 2014; Loshchilov & Hutter, 2017) optimizer with the weight decay rate 5×10^{-2} is used when training with one Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz and one NVIDIA GeForce RTX 4090 GPU with 24GiB memory. For other OOD detection methods, we adopt the same values of common training hyperparameters for fair comparison, and the parameter selection

Algorithm 1 GROD

Require: The training dataset and labels $\mathcal{X}_{\text{train}}$, $\mathcal{Y}$, the testing dataset and labels $\mathcal{X}_{\text{test}}$, $\mathcal{Y}_{\text{test}}$, the learnable parameter α, fixed parameters γ and the number of each cluster of OOD data num, batch size B, number of ID classes K
Ensure: Trained model M, classification results $\hat{\mathcal{Y}}_{\text{test}}$ for ID data and OOD detection
{**Fine-tuning Stage**}
for ep in training epochs **do**
  for each batch $\mathcal{X}$ in $\mathcal{X}_{\text{train}}$ **do**
    $\mathcal{F} \leftarrow \texttt{NET}(\mathcal{X})$ {Obtain features by Eq. (1)}
    $\mathcal{F}_{\text{PCA}} \leftarrow \text{PCA}(\mathcal{F}, num)$ {PCA projction}
    $V_{\text{PCA}} \leftarrow \texttt{Boundary}(\mathcal{F}_{\text{PCA}})$ {Obtain boundary ID data}
    $\mu_{\text{PCA}} \leftarrow \texttt{MEAN}(\mathcal{F})$ {Obtain centers of ID data}
    $\Sigma_{\text{PCA}} \leftarrow \texttt{COV}(\mathcal{F})$
    $\text{Dist}^{ID}_{\text{PCA}} \leftarrow \texttt{MEAN}(\texttt{DIST}(\mathcal{F}, \mu_{\text{PCA}}, \Sigma_{\text{PCA}}))$ {Obtain average distances of ID by Eq. (57) (the former one)}
    $\hat{\mathcal{F}}^{\text{OOD}}_{\text{PCA}} \leftarrow \texttt{GENERATE}(V_{\text{PCA}}, \mu_{\text{PCA}}, \alpha, num)$ {Generate fake OOD data by Eq. (4)-Eq. (5)}
    if $B/K > 2$ **then**
      $\mathcal{F}_{\text{LDA}} := \cup_{i=1}^{K} \mathcal{F}_{\text{LDA},i} \leftarrow \text{LDA}(\mathcal{F}, \mathcal{Y}, num)$ {Generate inter-class fake OOD data and calculate Mahalanobis distances similar to the above process}
      $V_{\text{LDA},i} \leftarrow \texttt{Boundary}(\mathcal{F}_{\text{LDA},i})$
      for $i_k \in \{1, \cdots, K\}$ **do**
        if $|\mathcal{F}_{\mathbf{y}=i_k}| > 1$ **then**
          $\mu_{\text{LDA},i_k} \leftarrow \texttt{MEAN}(\mathcal{F}|_{\mathbf{y}=i_k})$
          $\Sigma_{\text{LDA},i_k} \leftarrow \texttt{COV}(\mathcal{F}|_{\mathbf{y}=i_k})$
          $\mu^{\text{ID}}_{\text{LDA},i_k} \leftarrow \texttt{MEAN}(\texttt{DIST}(\mathcal{F}|_{\mathbf{y}=i_k}, \mu_{\text{PCA}}, \Sigma_{\text{LDA},i_k}))$
          $\hat{\mathcal{F}}^{\text{OOD}}_{\text{LDA},i_k} \leftarrow \texttt{GENERATE}(V_{\text{LDA},i_k}, \mu_{\text{LDA},i_k}, \alpha, num)$
        end if
      end for
      {Mahalanobis distance filtering mechanism by Eq. (58)-Eq. (59)}
      $\hat{\mathcal{F}}^{\text{OOD}} \leftarrow \hat{\mathcal{F}}^{\text{OOD}}_{\text{PCA}} \cup (\cup_{i_k} \hat{\mathcal{F}}^{\text{OOD}}_{\text{LDA},i_k})$
      $\text{Dist}^{\text{OOD}} \leftarrow \min_{i_k} \texttt{DIST}(\hat{\mathcal{F}}^{\text{OOD}}, \mu_{\text{LDA},i_k}, \Sigma_{\text{LDA},i_k})$
      $I_0 \leftarrow \text{argmin}_{i_k} \texttt{DIST}(\hat{\mathcal{F}}^{\text{OOD}}, \mu_{\text{LDA},i_k}, \Sigma_{\text{LDA},i_k})$
      $\Lambda \leftarrow \texttt{LAMBDA}(\lambda, \hat{\mathcal{F}}^{\text{OOD}}, \text{Dist}^{\text{OOD}}, \text{Dist}^{ID}_{\text{LDA},I_0})$
      $mask = \text{Dist}^{\text{OOD}} \geq (1+\Lambda)\text{Dist}^{ID}_{\text{LDA},I_0}$
    else
      $\hat{\mathcal{F}}^{\text{OOD}} \leftarrow \hat{\mathcal{F}}^{\text{OOD}}_{\text{PCA}}$
      $\text{Dist}^{\text{OOD}} \leftarrow \texttt{DIST}(\hat{\mathcal{F}}^{\text{OOD}}, \mu_{\text{PCA}}, \Sigma_{\text{PCA}})$
      $\Lambda \leftarrow \texttt{LAMBDA}(\lambda, \hat{\mathcal{F}}^{\text{OOD}}, \text{Dist}^{\text{OOD}}, \text{Dist}^{ID}_{\text{PCA}})$
      $mask = \text{Dist}^{\text{OOD}} \geq (1+\Lambda)\text{Dist}^{ID}_{\text{PCA}}$
    end if
    $\mathcal{F}^{\text{OOD}} \leftarrow \hat{\mathcal{F}}^{\text{OOD}}[\text{mask}]$
    if $|\mathcal{F}^{\text{OOD}}| > B/K + 2$ **then**
      $\mathcal{F}^{\text{OOD}} \leftarrow \mathcal{F}^{\text{OOD}}[\text{random mask}]$ {Random filtering mechanism}
    end if
    $\mathcal{F}_{\text{all}} \leftarrow \mathcal{F} \cup \mathcal{F}^{\text{OOD}}$
    $\mathcal{Y}_{\text{all}} \leftarrow \texttt{STACK}(\mathcal{Y}, (K+1)\mathbf{1}_{|\mathcal{F}^{\text{OOD}}|})$
    $\hat{\mathcal{Y}}_{\text{all}}, \text{LOGITS} \leftarrow \texttt{CLASSIFIER}(\mathcal{F}_{\text{all}})$
    Iterate the model parameters according to the loss function $\mathcal{L}$ in Eq. (53)-(55).
  end for
  Save model M with the best performance.
end for
Return M
{**Inference Stage**}
$\mathcal{F}_{\text{test}}, \text{LOGITS}_{\text{test}} \leftarrow M(\mathcal{F}_{\text{test}})$
$\text{LOGITS}_{\text{test}} \leftarrow \texttt{ADJUST}(\text{LOGITS}_{\text{test}})$ {Adjust LOGITS by Eq. (9)}
$\hat{\mathcal{Y}}_{\text{test}} \leftarrow \texttt{PostProcessor}(\mathcal{F}_{\text{test}}, \text{LOGITS}_{\text{test}})$ {Obtain prediction results after post-processing}
Return $\hat{\mathcal{Y}}_{\text{test}}$

and scanning strategy provided by OpenOOD (Zhang et al., 2023; Yang et al., 2022a;b; 2021; Bitterwolf et al., 2023) for some special parameters. For text classification, we employ the pre-trained BERT base model. We modify the default initial learning rate to 2×10^{-5} and the weight decay rate to 1×10^{-3}, and other hyperparameters maintain the same as in image

classification tasks.

We preserve the finetuned model with the highest ID data classification accuracy on the validation dataset and evaluate its performance with test datasets. The training and validation process is conducted without any OOD exposure.

J.2. Dataset details

For image classification tasks, we use four benchmark datasets *i.e* **CIFAR-10** (Krizhevsky et al., 2009), **CIFAR-100** (Krizhevsky et al., 2009), **Tiny ImageNet** (Le & Yang, 2015) and **SVHN** (Netzer et al., 2011). **CIFAR-10** or **CIFAR-100** serve as ID data, respectively, while the other three are OOD data. **SVHN** is uniquely identified as far-OOD data due to its distinct image contents and styles. For text classification, we employ datasets in Ouyang et al. (2023) to experiment with detecting semantic and background shift outliers. The semantic shift task uses the dataset **CLINC150** (Larson et al., 2019), where sentences lacking intents are treated as semantic shift OOD, following Podolskiy et al. (2021). For the background shift task, the movie review dataset **IMDB** (Maas et al., 2011) serves as ID, while the business review dataset **Yelp** (Zhang et al., 2015) is used as background shift OOD, following Arora et al. (2021). We provide details of each dataset as follows:

Image datasets.

- **CIFAR-10** (Krizhevsky et al., 2009): This dataset contains $60,000$ images of 32x32 pixels each, distributed across 10 diverse categories (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck). Each category includes $6,000$ images, split into $50,000$ for training and $10,000$ for testing. It is a standard benchmark for image classification tasks.

- **CIFAR-100** (Krizhevsky et al., 2009): Building on the structure of **CIFAR-10**, **CIFAR-100** offers greater variety with 100 categories, each containing 600 images. This dataset serves as an extension of **CIFAR-10**, providing a deeper pool of images for more complex machine-learning models.

- **Tiny ImageNet** (Le & Yang, 2015): **Tiny ImageNet** comprises $100,000$ images resized to 64×64 pixels, spread across 200 categories, with each category featuring 500 training samples, and 50 samples each for validation and testing. This dataset offers a broad spectrum of challenges in a format similar to the CIFAR datasets but on a larger scale.

- **SVHN** (Netzer et al., 2011): The Street View House Numbers (**SVHN**) dataset, extracted from Google Street View images, focuses on number recognition with 10 classes corresponding to the digits $0, 1, \cdots, 9$. This dataset is particularly suited for developing machine learning techniques as it simplifies preprocessing steps.

Text datasets.

- Semantic shift: Following the approach in Podolskiy et al. (2021), we use the CLINC150 dataset (Larson et al., 2019), which consists of phrases used in voice assistants, representing various intents. The OOD data is set to be phrases with unidentified intents, serving as "out-of-scope" inquiries not aligned with any predefined categories. This dataset is ideal for testing the robustness of intent classification systems against unexpected queries and includes both in-scope and out-of-scope data.

- Background shift: We follow (Arora et al., 2021) to choose the long movie review dataset **IMDB** (Maas et al., 2011) as the ID dataset and a business review dataset **Yelp** (Zhang et al., 2015) as the OOD dataset. The **IMDB** dataset consists of $50,000$ movie reviews, tailored for binary sentiment classification to discern positive and negative critiques. The **Yelp** dataset, which includes a variety of business, review, and user data, represents a shift in the background context and is treated as OOD data, providing a different commercial background from the movie reviews of the **IMDB** dataset.

K. Ablation study

Comprehensive ablation studies are conducted to explore hyper-parameters and optimization strategies, where Figure 4 shows the ablation experiments for GROD concerning key parameters γ, α, and num, and the ablation results of modules in GROD are displayed in Table K.

Ablation on the loss weight γ. Figure 4(a) examines variations in γ within the loss function as detailed in Eq. (6)-(8). As outlined in Section 2, changes in γ show the trade-off within the loss function $\mathcal{L}$. When the value of γ ranges from 0 to 1, the performance under each evaluation metric initially increases and then decreases. When $\gamma = 1$, the model fails to classify ID data. Intriguingly, $\mathcal{L}_2$ and the fake OOD slightly enhance the ID classification performance, surpassing the 10% accuracy threshold of randomness, which explains how GROD simultaneously improves ID data classification and OOD detection performance, as illustrated in Section 3.2. The efficiency of $\mathcal{L}_2$ also indicates that OOD generated by GROD closely mimics OOD from real datasets.

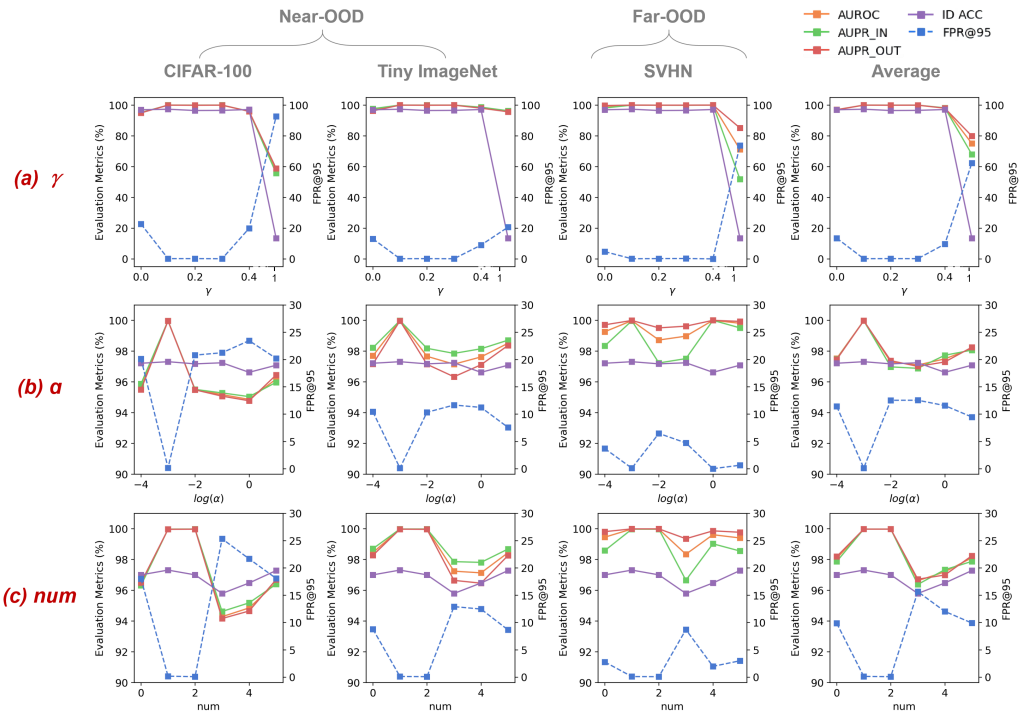

Figure 4. Ablation study on extra hyperparameters in GROD. (a) The weight γ in $\mathcal{L}$. (b) The parameter α adjusts the extending distance of generated OOD data. (c) The number of every OOD cluster num. The ID dataset is **CIFAR-10** and the backbone is the pre-trained ViT-B-16.

Ablation on α in adjusting the ID-OOD distance. In Figure 4(b), the value of α is adjusted, demonstrating that a larger α increases the Mahalanobis distance between ID and synthetic OOD. Empirical results indicate that an α value of 1×10^{-3} achieves optimal performance when using LDA projection. If α is reduced, causing ID and OOD data to be too closely aligned in Mahalanobis distance, the model tends to overfit and fails to discern their differences. Conversely, if α is too high, most inter-class OOD data either become global OOD around ID data or resemble ID from other classes, thus being excluded by the Mahalanobis distance condition in Eq. (58). At this time, inter-class OOD is similar to global OOD typically generated only by PCA, leading to a significant drop in near-OOD detection performance, while far-OOD detection remains consistent. The performance curves of near-OOD detection also indicate that if only PCA projections are used, we can set α in a larger value, as the performance increases after dropping from the top.

Ablation on num in the number of outliers. Figure 4(c) explores how the dimension parameter num influences performance. The model demonstrates superior performance when num is set to 1 or 2, as PCA and LDA effectively retain characteristics of the original data and distinguish clusters of each category. Increasing the dimensions of PCA and LDA projections often results in the selection of less representative features in our filtering mechanism. Besides, maintaining num at 1 or 2 usually ensures a balanced ratio of generated OOD data to ID data. Overall, the model consistently delivers competitive outcomes, affirming the efficiency of GROD in various settings.

Ablation on key modules in GROD. GROD comprises three key modules: adjusting the loss function, generating virtual OOD data, and employing the Mahalanobis distance filtering mechanism, denoted as $\mathcal{L}2$, $\mathcal{F}_{\text{OOD}}$, and $Maha$, respectively.

Table 4. Ablation experiments. The ID dataset is **CIFAR-10** and the backbone is ViT-B-16 pre-trained with **ImageNet-1K**. Respectively, $\mathcal{L}_2$, $\mathcal{F}_{\text{OOD}}$, $Maha$ represent whether to use the binary loss function $\mathcal{L}_2$, fake OOD data generation and Mahalanobis distance filtration.

OOD Datasets			-	CIFAR-100				Tiny ImageNet				SVHN				Average			
Evaluate Metrics (%)			ID ACC↑	F↓	A↑	I↑	O↑	F↓	A↑	I↑	O↑	F↓	A↑	I↑	O↑	F↓	A↑	I↑	O↑
$\mathcal{L}_2$	$\mathcal{F}_{\text{OOD}}$	$Maha$																	
			96.16	21.59	95.43	**95.64**	95.38	8.52	**98.39**	**98.68**	**98.14**	3.26	99.39	**98.61**	99.78	11.12	97.74	97.64	97.77
	✓	✓	**96.96**	22.66	94.98	95.13	94.94	13.04	96.98	97.68	96.27	4.69	99.18	98.11	99.70	13.46	97.05	96.97	96.97
✓			97.00	18.02	96.32	96.32	96.49	**8.78**	98.45	98.70	98.27	**2.76**	**99.45**	98.58	**99.81**	9.85	98.07	**97.87**	98.19
✓	✓		96.68	**21.17**	**95.57**	95.52	**95.78**	9.41	98.27	98.58	98.04	0.49	99.83	99.77	99.88	**10.36**	**97.89**	97.96	**97.90**
✓	✓	✓	97.31	0.16	99.97	99.97	99.96	0.11	99.98	99.98	99.97	0.09	99.98	99.97	99.99	0.12	99.98	99.97	99.97

Table K presents the ablation studies for these modules. $\mathcal{L}2$ alone can enhance model optimization, whereas $\mathcal{F}$OOD and $Maha$ contribute positively when integrated with $\mathcal{L}_2$. Utilizing all three strategies concurrently yields optimal performance, confirming that GROD effectively synergizes these modules to assign penalties associated with OOD and sharpen the precision of the ID-OOD decision boundary. Moreover, features $\mathcal{F}_{\text{all}}$, along with the prediction LOGITS LOGITS of GROD and the baseline, are visualized under t-SNE dimensional embedding (Appendix L), which illustrate the efficiency of GROD directly.

L. Visualization for fake OOD data and prediction likelihood

Feature visualization. As shown in Figure 5, we use the t-SNE dimensionality reduction method to visualize the two-dimensional dataset embeddings in the feature space. All the subfigures are derived from the same fine-tuned ViT-B-16 model.

The ID dataset, the test set of **CIFAR-10**, displays ten distinct clusters after embedding, each clearly separated. Consistent with our inference on GROD, the LDA projection generates fake OOD around each ID data cluster. Despite the high-dimensional feature space where OOD data typically lies outside ID clusters due to GROD's generation and filtering mechanisms, the two-dimensional visualization occasionally shows virtual OOD data within the dense regions of ID. This occurs because the projection from high dimensions to two-dimensional space inevitably results in some loss of feature expression, despite efforts to maintain the integrity of the data distribution.

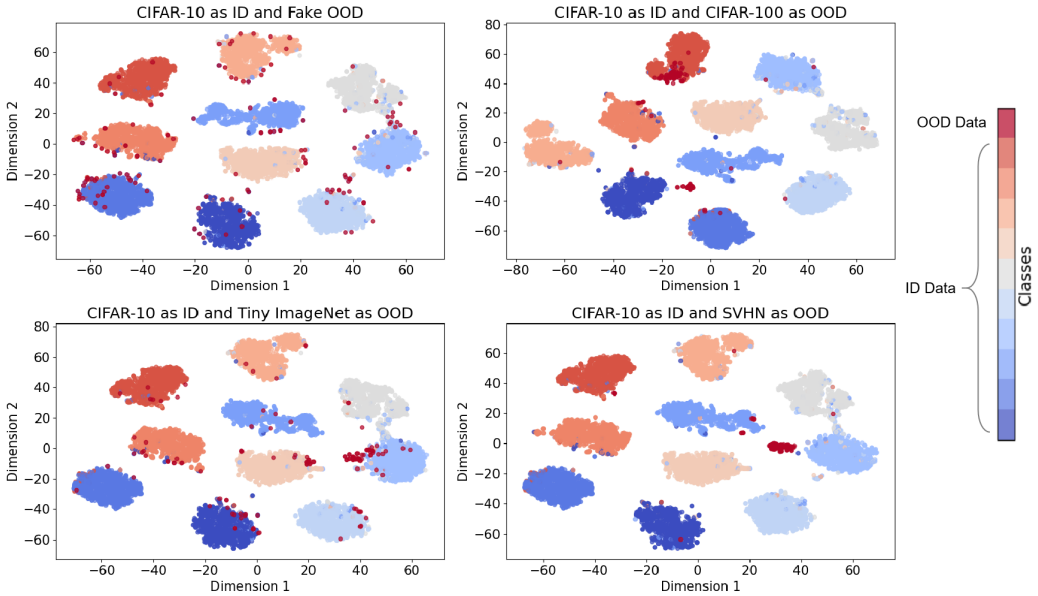

Figure 5. t-SNE visualization of the generated OOD data and test sets in the feature space.

We also visualize real OOD features from near-OOD datasets **CIFAR-100** and **Tiny ImageNet**, and the far-OOD dataset **SVHN**. To distinctively compare the distribution characteristics of fake and real OOD data, we plot an equal number

of real and synthetic OOD samples selected randomly. Near-OOD data resembles our synthetic OOD, both exhibiting inter-class surrounding characteristics, while far-OOD data from SVHN displays a different pattern, mostly clustering far from the ID clusters. Although far-OOD data diverges from synthetic OOD data, the latter contains a richer array of OOD features, facilitating easier detection of far-OOD scenarios. Thus, GROD maintains robust performance in detecting far-OOD instances as well. The visualization results in Figure 5 confirm that GROD can generate high-quality fake OOD data effectively.

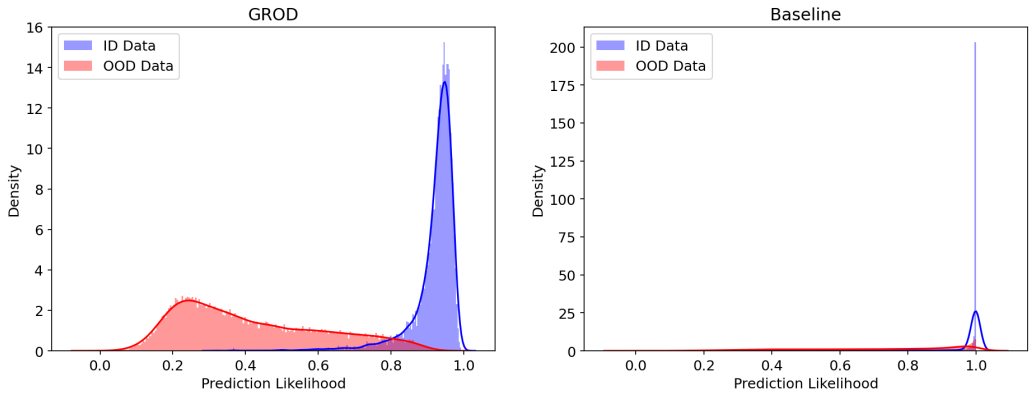

Figure 6. The distribution histograms and probability density curves of prediction likelihoods of ID and OOD test data. Results derived by GROD and the baseline MSP are visualized, with **CIFAR-10** as ID and **SVHN** as OOD.

Likelihood visualization. The process of OOD detection and model performance evaluation follows a standardized protocol, where classification predictions and their likelihood scores are generated and subsequently analyzed. The likelihood scores for OOD data are typically lower than those for ID data, as OOD samples do not fit into any ID category, resulting in a bimodal distribution of likelihood scores of all test data. In this distribution, ID and OOD form distinct high-frequency areas, separated by a zone of lower frequency. A broader likelihood range in this low-frequency zone with minimal overlap between the ID and OOD data signifies the model is more effective for OOD detection.

Comparing the likelihood distributions of the baseline MSP model with GROD as shown in Figure 6, it is evident that GROD significantly enhances the distinction in classification likelihood between ID and OOD, thereby improving OOD detection performance. The enhancements are quantitatively supported by the performance metrics reported in Table 1, where GROD surpasses the baseline by 15.30% in FPR@95 and 4.87% in AUROC on datasets CIFAR-10 and SVHN.