Textual Training for the Hassle-Free Removal of Unwanted Visual Data : Case Studies on OOD and Hateful Image Detection

Saehyung Lee[1]* **Jisoo Mok**[1] **Sangha Park**[1] **Yongho Shin**[2] **Dahuin Jung**[3]† **Sungroh Yoon**[1,4]†
[1]Department of Electrical and Computer Engineering, Seoul National University
[2]Qualcomm Korea YH, Seoul, South Korea
[3]School of Computer Science and Engineering, Soongsil University
[4]Interdisciplinary Program in Artificial Intelligence, Seoul National University
{halo8218, magicshop1118, wiarae}@snu.ac.kr,
yshin@qti.qualcomm.com, dahuin.jung@ssu.ac.kr, sryoon@snu.ac.kr

Abstract

In our study, we explore methods for detecting unwanted content lurking in visual datasets. We provide a theoretical analysis demonstrating that a model capable of successfully partitioning visual data can be obtained using only textual data. Based on the analysis, we propose Hassle-Free Textual Training (HFTT), a streamlined method capable of acquiring detectors for unwanted visual content, using only synthetic textual data in conjunction with pre-trained vision-language models. HFTT features an innovative objective function that significantly reduces the necessity for human involvement in data annotation. Furthermore, HFTT employs a clever textual data synthesis method, effectively emulating the integration of unknown visual data distribution into the training process at no extra cost. The unique characteristics of HFTT extend its utility beyond traditional out-of-distribution detection, making it applicable to tasks that address more abstract concepts. We complement our analyses with experiments in out-of-distribution detection and hateful image detection. Our codes are available at https://github.com/Saehyung-Lee/HFTT

1 Introduction

We are currently in the midst of what is known as the large-scale AI era. The growth in both the size of deep neural networks and training datasets led to unparalleled achievements in a wide array of tasks [4, 43]. However, this transition to large-scale AI presents new, unforeseen challenges. Particularly, the recent reports on the biased behavior of large AI models raise significant concerns surrounding the continuously expanding size of training datasets without proper quality control and regulation [2, 41]. The massive scale of visual training datasets necessary to train large-scale models presents a challenge in curating data to ensure unbiased and safe datasets. This is primarily due to the impracticality of manually selecting and removing unwanted content from such an extensive collection of images. This issue of data curation has traditionally been addressed by: (i) creating a supervised dataset for a specific objective; (ii) training a model on this dataset; and then (iii) utilizing the model to develop a larger dataset [28]. However, this approach exploits considerable human labor and needs to be re-initiated from the beginning whenever there are changes in the training objective.

*This publication was created by Saehyung Lee while an intern at QTI and currently attends Seoul National University.

†Corresponding Authors

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

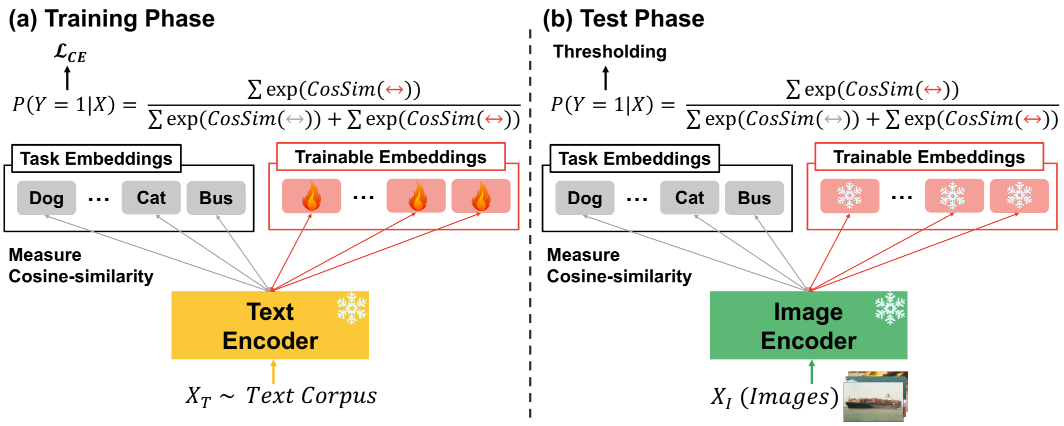
(a) Training Phase
$\mathcal{L}_{CE}$
$P(Y=1|X) = \frac{\sum \exp(CosSim(\leftrightarrow))}{\sum \exp(CosSim(\leftrightarrow)) + \sum \exp(CosSim(\leftrightarrow))}$
Task Embeddings
Dog
Cat
Bus
Trainable Embeddings
Measure Cosine-similarity
Text Encoder
$X_T \sim Text\ Corpus$
(b) Test Phase
Thresholding
$P(Y=1|X) = \frac{\sum \exp(CosSim(\leftrightarrow))}{\sum \exp(CosSim(\leftrightarrow)) + \sum \exp(CosSim(\leftrightarrow))}$
Task Embeddings
Dog
Cat
Bus
Trainable Embeddings
Measure Cosine-similarity
Image Encoder
$X_I\ (Images)$

Figure 1: Overview of our proposed method. **Task embeddings** define the task to be performed. For example, in the case of hateful image detection, hate speeches would serve as task embeddings, while in OOD detection, the names of classes from the training distribution would be the task embeddings. **Trainable embeddings** are the only parameters that are trained in our method, defined in the joint embedding space. During the training phase, only textual data are used, and in the testing phase, these trained parameters are employed to classify images. Detailed explanations are provided in Sections 3.

The field of out-of-distribution (OOD) detection, which aims identify OOD data that lie outside the training data distribution, can be considered a sub-branch of data curation research. Recent works in OOD detection utilize vision-language models (VLMs) [40, 29, 30] to take advantage of the rich and human-aligned representations learned by these models. For instance, Esmaeilpour et al. [11] augmented the pre-trained CLIP model [40] with an additional decoder module, trained on a vision-language dataset, for visual OOD detection. In a similar vein, Wang et al. [49] incorporated an extra "no" text encoder, trained on a vision-language dataset, into CLIP. These previous approaches, however, suffer from a significant limitation: They require a vast amount of additional vision-language data. Using data samples targeted for detection can improve sample efficiency, but this leads to the dilemma of needing to collect unwanted data for the purpose of removing them.

In this work, we propose a novel method that no longer relies on additional visual data or a computationally expensive training process. We first outline our theoretical rationale, particularly demonstrating that with a successfully trained model on a bimodal dataset, like CLIP, one can obtain a classifier to partition data from one mode using solely the data from the other mode. Building upon our motivation, we propose a method called Hassle-Free Textual Training (HFTT). HFTT consists of a newly proposed loss and a clever textual data synthesis method, updating trainable parameters defined in the joint embedding space to improve the detection of undesirable visual content. Specifically, we decompose the weighted cross-entropy loss into a formula that includes a regularization term, which is tailored to our use. Additionally, to achieve higher detection accuracy, we introduce the concept of focal loss [33]. Moreover, our textual data synthesis method, which combines prompt templates and words, can effectively imitate the involvement of the entire visual data distribution in the training process. We illustrate an overview of our proposed method in Figure 1.

The proposed loss function brings considerable convenience in achieving our objective. To train an unwanted data detection model, it is necessary to define out-distribution for a given data distribution (in-distribution), which is not always straightforward due to vague boundaries between the two. For instance, in hateful content detection tasks [13], the divide between what is hateful and what is not is influenced by various contexts, *e.g.*, historical and social backgrounds. Our proposed loss eliminates the need for human labor to annotate out-distribution data because it does not involve a clearly defined set of out-distribution data. Furthermore, our textual data synthesis method requires no cost, employing a rule-based approach using only prompt templates and a set of words.

Based on the principle that HFTT can detect out-distribution samples by merely defining the in-distribution in natural language, we propose that this method can be extended to tasks beyond traditional OOD detection, including hateful image detection. Current OOD detection methods often fail in such extended tasks for two main reasons: firstly, they are based on the assumption of a distinct boundary between in- and out-distributions, which is unsuitable for tasks with abstract concepts.

Secondly, methods requiring training images may lead to ethical concerns. Our proposed method, however, is not subject to these limitations.

Through empirical validation, we verify that HFTT can enhance the performance of VLMs in identifying unwanted visual data, whether it be OOD instances or hateful images. Additionally, we demonstrate through feature visualization results that HFTT, despite not observing any visual data in the training phase, appears to have been trained as if it had. Furthermore, we provide various analyses of HFTT, including comparative results of using different textual data synthesis methods.

In summary, our **contributions** are as follows: (i) We theoretically demonstrate how textual data can serve as a substitute for visual data in our scenario; (ii) We introduce a new loss function. Its use eliminates the need for labor in annotating out-distribution data; (iii) We propose a textual data synthesis method that can efficiently imitate the visual data distribution in our training; (iv) We empirically analyze HFTT, a method composed of the above proposals. Our experiments show that HFTT is effective in a range of scenarios, from traditional OOD detection to situations involving abstract concepts, like the identification of hateful images.

2 Related Work

Vision-language models. With the advancements in deep learning, tackling sophisticated tasks that demand an understanding of both vision and language modalities has become viable. The methodologies employed to encode image and text data exhibit notable distinctions owing to their inherent differences. Prominent within this domain are dual-stream models exemplified by CLIP [40], ALIGN [23], and FILIP [51]. These models employ separate encoders for text and image data, optimizing them through contrastive objectives to align semantically similar features across heterogeneous modalities. Primarily, VLMs integrate transformer-based encoders for text data, while a variety of architectures, encompassing convolutional neural networks [25, 15] and vision transformers [9], are deployed for image encoding. The success of CLIP-like models has spurred numerous subsequent inquiries, with a focus on enhancing data efficiency and adaptability for diverse downstream tasks.

Out-of-distribution detection. Traditionally, OOD detection has evolved by defining post-hoc OOD scores [18, 32, 27, 34] or formulating learning algorithms based on outlier exposure methods [26, 20, 10]. With the advancement of VLMs, methods for OOD detection that leverage both image and text embeddings have also progressed. Post-hoc OOD score methods based on VLM typically involve utilizing the OOD class name [12, 11] or defining OOD scores using the top similarity values between images and class names [37]. In the case of outlier exposure, which requires a training algorithm, some approaches employ prompt learning or fine-tune the image encoder [45] of models like CLIP. Transitioning from conventional methods to VLM-based approaches, none of these methods have attempted text-only training and subsequently applied their techniques to tasks such as hateful image detection.

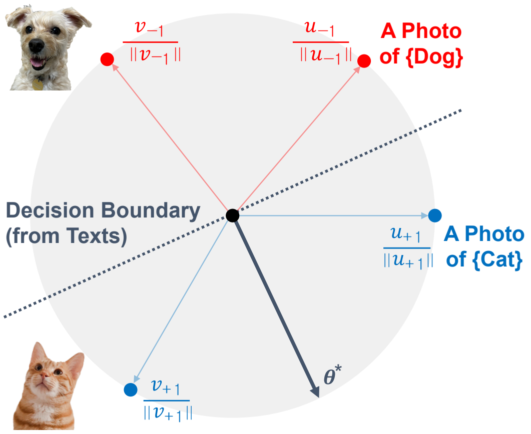
$\frac{v_{-1}}{\|v_{-1}\|}$
$\frac{u_{-1}}{\|u_{-1}\|}$
A Photo of {Dog}
Decision Boundary (from Texts)
$\frac{u_{+1}}{\|u_{+1}\|}$
A Photo of {Cat}
$\frac{v_{+1}}{\|v_{+1}\|}$
θ^*

Figure 2: Overview of Section 3.1. The red and blue colors symbolize the two classes -1 and $+1$, respectively. In our theoretical model, u and v can be interpreted as text and image, respectively.

Text-only training for vision tasks. Given the progress in VLMs, there have been numerous studies aimed at replacing images with textual representations in vision and multimodal tasks. Textual data presents the advantage of being easily collectible compared to visual data. Previous studies have demonstrated the effectiveness of using only textual information for various vision tasks, including image classification [38], image captioning [31], and multi-label image recognition [14]. Our work represents a pioneering effort in applying text-only supervision to unwanted visual data detection.

3 Method

In this section, we propose a new textual training method for the convenient and successful removal of unwanted visual data. In Section 3.1, we theoretically demonstrate, through a motivating example, that when there is a well-trained model on a bimodal dataset, such as CLIP, it is possible to train a binary classifier successfully partitioning data from one mode using only data from the other. This theoretical revelation leads us to our novel loss function that allows hassle-free training of an unwanted visual data detector in Section 3.2. Lastly, in Section 3.3, we present our proposed method that includes a simple yet effective synthesis of textual data. The proposed method is executable even without access to the parameters of the backbone model, making it lightweight and applicable to black-box foundational models.

3.1 A Motivating Example

We theoretically demonstrate that when there exists a well-trained bimodal model $F : \mathcal{G} \times \mathcal{H} \to \mathcal{Z}$ for a given bimodal data distribution, it is possible to train a classifier successfully partitioning data from one mode ($\mathcal{H}$) using only the dataset from the other mode ($\mathcal{G}$). To align with the operations of VLMs like CLIP in our scenario, we assume that the output vectors of F are normalized. We define a bimodal dataset D as follows:

$$D = \{(g_i, h_i, y_i)\}_{i=1}^{N}, \text{ where } y \overset{u.a.r.}{\sim} \{-1, +1\} \text{ and } (g, h) \overset{i.i.d.}{\sim} \mathcal{G}_y \times \mathcal{H}_y.$$

(g_i, h_i) represents the input vectors from the two modes for a given data sample, where y_i denotes the binary class of the i-th sample. We can partition the dataset D as follows:

$$D = D_{-1} \cup D_{+1}, \text{ where } D_y = \{(g_i, h_i) \,|y = y_i\}.$$

We assume that samples belonging to the same class in the dataset D exhibit similar semantic patterns. Given F that successfully builds the joint embedding space for the bimodal data distribution, we can posit the following:

$$u_{+1}^\top v_{+1} > u_{+1}^\top v_{-1} \text{ and } u_{-1}^\top v_{+1} < u_{-1}^\top v_{-1}, \text{ where } u_y = \mathop{\mathbb{E}}_{u \in U_y}[u], v_y = \mathop{\mathbb{E}}_{v \in V_y}[v],$$
$$U_y = \{F(g_i) \,|g_i \in D_y\}, \text{ and } V_y = \{F(h_i) \,|h_i \in D_y\}.$$

U_y and V_y are class-conditional embedding sets for each of the two modes, respectively. For simplicity, we assume that the variances of the angular distributions relative to their mean vectors are equal for sets U_{-1} and U_{+1}, as well as for sets V_{-1} and V_{+1}.

We investigate whether the cosine-similarity classifier $\theta^\star$ trained solely on the unimodal dataset $(g_i, y_i)_{i=1}^{N}$ using F can successfully be applied to $(h_i, y_i)_{i=1}^{N}$. We establish the following theorem:

Theorem 1. *For the quadratic loss function $L(u, y; \theta) = \left(1 - y\theta^\top u\right)^2$, the optimal cosine-similarity classifier $\theta^\star$ that classifies sets U_{-1} and U_{+1} is*

$$\arg\min_\theta \mathop{\mathbb{E}}_{u \in U_{-1}}[L(u, -1; \theta)] + \mathop{\mathbb{E}}_{u \in U_{+1}}[L(u, +1; \theta)] = \frac{u_{+1} - u_{-1}}{\|u_{+1} - u_{-1}\|}.$$

Proofs are in Appendix A. Theorem 1 demonstrates the optimal classifier $\theta^\star$ is orthogonal to $u_{-1} + u_{+1}$. We present an illustration in Figure 2 to enhance understanding of both the problem under investigation and the results of our analysis.

Applying the classifier $\theta^\star$ trained to classify U_{-1} and U_{+1} to distinguish between V_{-1} and V_{+1} leads to the following:

Corollary 1. *The classifier $\theta^\star$, with respect to V_{-1} and V_{+1}, satisfies the double inequalities of*

$$\mathop{\mathbb{E}}_{v \in V_{-1}}\left[\theta^{\star\top} v\right] < 0 < \mathop{\mathbb{E}}_{v \in V_{+1}}\left[\theta^{\star\top} v\right].$$

This implies that we can successfully classify V_{-1} and V_{+1} by observing the cosine similarities with $\theta^\star$. Motivated by these theoretical examples, we hypothesize that classifiers obtained solely using textual data can operate on visual data as well. Section 4 empirically demonstrates that the arguments developed based on our theoretical model can be applied to modern machine-learning settings.

3.2 Our Proposed Loss Function

Our objective is to distinguish in-distribution data samples (D_{in}), conforming to given data distribution, from out-distribution data samples (D_{out}). The development of our new loss function begins with defining the binary cross-entropy loss L as follows:

$$L(u,y) = -\frac{1+y}{2}\log p\left(u\right) - \frac{1-y}{2}\log\left(1-p(u)\right).$$

We employ the notations introduced in Section 3.1. $p(u)$ denotes the probability that the label of an embedding u is $+1$, where $+1$ signifies an out-distribution. With respect to datasets U_{-1} and U_{+1}, we minimize

$$\sum_{u\in U_{-1}} \lambda L\left(u,-1\right) + \sum_{u\in U_{+1}} \left(1-\lambda\right) L\left(u,+1\right). \tag{1}$$

We introduce a hyper-parameter, $\lambda \in [0,1]$, to adjust the balance between in-distribution learning (the first term) and out-distribution learning (the second term). Equation (1) can be reformulated as

$$= \sum_{u\in U_{-1}} L\left(u,-1\right) - \sum_{u\in U_{-1}} \left(1-\lambda\right) L\left(u,-1\right) + \sum_{u\in U_{+1}} \left(1-\lambda\right) L\left(u,+1\right). \tag{2}$$

The second term can be understood as regularization for in-distribution learning. As λ approaches 0, in-distribution learning is more heavily impeded. Rather than employing the original regularization term, $-\sum_{u\in U_{-1}}\left(1-\lambda\right)L\left(u,-1\right)$, we propose changing it to

$$\sum_{u\in U_{-1}} \left(1-\lambda\right) L\left(u,+1\right).$$

Before analyzing the significance of this modification to the objective function, we first examine the effects resulting from this change. Along with the modification, our objective function can be formulated as follows:

$$\begin{aligned}\sum_{u\in U_{-1}} L\left(u,-1\right) + \sum_{u\in U_{-1}} \left(1-\lambda\right) L\left(u,+1\right) + \sum_{u\in U_{+1}} \left(1-\lambda\right) L\left(u,+1\right) \\ = \sum_{u\in U_{-1}} L\left(u,-1\right) + \sum_{u\in U_{-1}\cup U_{+1}} \left(1-\lambda\right) L\left(u,+1\right).\end{aligned} \tag{3}$$

To minimize Eq. 2, it is imperative to distinguish between the in-distribution dataset and the out-distribution dataset. In-distribution data aligns with the objective of the given task, and any data not included in it becomes out-distribution data. However, in real-world scenarios, distinguishing between these distributions is not straightforward. For instance, if we consider U as the text embedding space, collecting out-distribution texts for a given set of in-distribution texts involves considerations such as homonyms, synonyms, and various forms of linguistic variations. Particularly, in tasks where the boundaries between in-distribution and out-distribution are ambiguous, as seen in challenges such as hate content detection [13], constructing a dataset for Eq. 2 becomes difficult and requires considerable human labor. However, the utilization of Eq. 3 alleviates us from such challenges. In other words, the union of the two sets $U_{-1} \cup U_{+1}$ in Eq. 3 allows us to treat all data samples as out-distribution without the need to ponder their relationship with the in-distribution, providing a solution to the intricacies involved in dataset construction. The distinction between Eq. 2 and 3 becomes evident when comparing the gradient signals produced by the two different regularization terms. Gradients of the regularization terms in Eq. 2 and 3 can be computed as follows:

$$\begin{aligned}&\text{(original)} \quad -\sum_{u\in U_{-1}} \frac{\partial L\left(u,-1\right)}{\partial p(u)} = \sum_{u\in U_{-1}} \frac{-1}{1-p(u)},\\ &\text{(proposed)} \quad \sum_{u\in U_{-1}} \frac{\partial L\left(u,+1\right)}{\partial p(u)} = \sum_{u\in U_{-1}} \frac{-1}{p(u)}.\end{aligned}$$

The original regularization term weakly regularizes in-distribution samples that were sufficiently learned by the model (*i.e.*, samples with low $p(u)$). However, the proposed regularization term does the exact opposite; it imposes stronger regularization on in-distribution samples with low

$p(u)$. In essence, our proposed regularization prevents the model from exhibiting high confidence in in-distribution samples and enforces the decision boundary to be created near the in-distribution.

Recent studies show that the closer the decision boundary of the out-distribution data detector is to the in-distribution, the more effective the detector is at identifying various out-distribution data [26, 19, 10, 39]. Subsequent research efforts have been directed at obtaining out-distribution samples that reside close to the in-distribution while training a detector to bring its decision boundary closer to the in-distribution. For instance, Lee et al. [26] utilizes a generative adversarial network to acquire samples placed on the in-distribution boundary. Du et al. [10] models the in-distribution using a Gaussian distribution and samples embeddings from the low-likelihood regions of the defined Gaussian distribution. Likewise, we focus on training samples situated in the region close to the in-distribution by incorporating an additional focal loss.

The focal loss was initially proposed to forcefully suppress the gradients for background pixels, which dominate the image, and intensify the learning signals from foreground pixels. Under our scenario, the in-distribution, like foreground pixels, tends to inhabit a small portion of the entire embedding space. In light of the similarity between the in-distribution and foreground pixels, we utilize the focal loss to restrain the loss from far out-distribution samples and amplify learning signals from samples near the in-distribution. The proposed loss can thus be defined as:

Definition 1. *Let $B_{-1} = \{x_i\}_{i=1}^{N}$ and $B = \{\tilde{x}_i\}_{i=1}^{N}$ denote mini-batches that are respectively drawn from the specified in-distribution and the overall data distribution. Let L be the cross-entropy loss. Then, our proposed loss function is*

$$\sum_{x_i \in B_{-1}} L\left(x_i, -1\right) + \left(1 - \lambda\right) \sum_{x_j \in B} \beta_j L\left(x_j, +1\right); \;\; \beta_j = \frac{N \alpha_j}{\sum_{x_k \in B} \alpha_k} \text{ and } \alpha_j = \left(1 - p\left(x_j\right)\right)^{\gamma}. \quad (4)$$

$p\left(x\right)$ is the predictive probability of x belonging in out-distribution. $\gamma \geq 0$ is treated as a hyperparameter of the focal loss. In Section C, we compare the results of using loss terms in Eqs. 2 and 3 and demonstrate the particular effectiveness of the focal loss.

3.3 Hassle-Free Textual Training (HFTT)

So far, we have assumed access to data sampled from the out-distribution. However, we may not always be able to anticipate the out-distribution in advance, and even if we can, sampling a subset of data that is representative of the entire distribution is not a straightforward problem in nature. In our scenario, we solely utilize textual data to learn an unwanted visual data detector. Therefore, the proposed scenario requires texts to define the in-distribution and a comprehensive corpus of textual data that can replace the entirety of visual data. VLMs, such as CLIP, obtained impressive zero-shot classification accuracy on diverse visual data benchmarks through the usage of prompts, *e.g.*, "a photo of a { }." Inspired by the success of prompting in VLMs, we conjecture that all visual data can be expressed textually through prompts. This assumption allows the textual dataset to replace the unknown visual data distribution by integrating words associated with the visual data into prompts, drastically simplifying the process of textual data sampling in our method. One example of a prompt design utilized in our approach is: "This is a photo of a { }." To emulate the effect of using the entire visual data distribution, we adopt a word set[3] that includes approximately 370k English words. We report the results of using other prompt designs or textual data acquisition processes in Appendix C. While the optimization procedure in our method additionally requires in-distribution textual data according to Eq. 4, these can be obtained with minimal effort by creating arbitrary sentences or prompts related to the given task. Section 4 details how in-distribution textual data are attained for each experimental setting. Even though the task of unwanted visual data detection is a type of binary classification problem, learning a plain linear classifier in the embedding space of pre-trained VLMs through approaches like linear probing is not necessarily compatible with our task because the in- and out-distributions are not expected to be linearly separable. To accurately estimate the probability that input x belongs in the out-distribution $p\left(x\right)$, we must take advantage of the informative signals in the text encoder of pre-trained VLMs. In our method, $p\left(x\right)$ is computed as follows:

1. Obtaining embeddings $\{w_i^{\text{in}}\}_{i=1}^{K}$ for K texts that effectively represent in-distribution visual data is equivalent to defining the task. This process is akin to obtaining zero-shot classifiers using VLMs. We will refer to these text embeddings as **task embeddings**.

[3] https://github.com/dwyl/english-words?tab=readme-ov-file

Algorithm 1 Hassle-Free Textual Training (HFTT)

```
Require: word set $\mathcal{W}$, prompt templates $\mathcal{P}$, in-distribution textual data $\mathcal{G}_{-1}$, task embeddings
    $\{w_i^{\text{in}}\}_{i=1}^{K}$, trainable embeddings $\{w_j^{\text{out}}\}_{j=1}^{N}$, pre-trained model $F$, hyper-parameter $\lambda$
Initialize the trainable embeddings $\{w_j^{\text{out}}\}_{j=1}^{N}$
for mini-batches $(B_{-1}, \texttt{words}) \sim (\mathcal{G}_{-1}, \mathcal{W})$ do
    $B \leftarrow \texttt{word2data}(\texttt{words}, \mathcal{P})$    # textual data synthesis (the entire data distribution)
    Compute the proposed loss by Eq. 4
    Update $\{w_j^{\text{out}}\}_{j=1}^{N}$    # the incurred cost is negligible
end for
Output: out-distribution data detector $\left(F, \{w_i^{\text{in}}\}_{i=1}^{K}, \{w_j^{\text{out}}\}_{j=1}^{N}\right)$
```

2. With a pre-trained vision-language model F and the set of N **trainable embeddings** $\{w_j^{\text{out}}\}_{j=1}^{N}$ defined in the joint embedding space, $p\left(x\right)$ is obtained as

$$\frac{\sum_{j=1}^{N} \exp\left(F\left(x\right)^{\top} w_j^{\text{out}}\right)}{\sum_{i=1}^{K} \exp\left(F\left(x\right)^{\top} w_i^{\text{in}}\right) + \sum_{j=1}^{N} \exp\left(F\left(x\right)^{\top} w_j^{\text{out}}\right)}.$$

Our method minimizes the custom loss defined in Eq. 4 by learning $\{w_j^{\text{out}}\}_{j=1}^{N}$ only with textual data with the task embeddings $\{w_i^{\text{in}}\}_{i=1}^{K}$ and the model F kept frozen. Because the trainable embeddings are tuned in the output space of the backbone network, the proposed method results in little memory and computational cost. Furthermore, no need to access the parameters of the backbone network makes the proposed method extensible to black-box foundation models. The overall procedure of the proposed method is summarized in Algorithm 1.

4 Experimental Results and Discussion

4.1 Experimental Setup

We complement our analysis with case studies conducted on OOD and hateful image detection. For the OOD detection task, ImageNet-1k [7] is treated as in-distribution, and the following datasets are used as out-distribution data: iNaturalist [47], SUN [50], Places [52], and Textures [5]. We specifically utilize OOD datasets that are carefully curated to be disjoint from ImageNet, as described in [22]. For hateful image detection, we utilize a dataset containing 892 Antisemitic/Islamophobic images and 420 phrases (Hate) [13]. The Hate dataset is a human-annotated dataset whose usage is limited to individuals with academic purposes to prevent its unethical and unregulated use.

We adopt CLIP, the most extensively studied VLM, specifically using ViT-B/16 as the vision backbone. Unless specified otherwise, we set the batch size=256, learning rate=1.0, epoch=1, γ=1.0 (the focal loss hyper-parameter), λ=0, and N=10 (the number of trainable embeddings) for all experiments. Note that in the majority of scenarios, there is a substantial predominance of out-distribution textual data compared to in-distribution textual data, and our approach involves mini-batch sampling. Consequently, given the rarity of in-distribution data sampling, the training on in-distribution data remains largely unaffected even though $\lambda = 0$. All values presented in the tables of this paper are the average results over five runs. We conduct a comparative analysis of our approach against existing methods requiring in-distribution images, namely Mahalanobis [27], MSP [18], KNN [44], and NPOS [46]. Additionally, we include methods that do not necessitate in-distribution data, Energy [34], ZOC [11], MaxLogit [21], and MCM [36], in our comparison. The evaluation is performed using OOD scores proposed by the aforementioned works, as well as the scores introduced in this paper (refer to $p\left(x\right)$ in Sec 3.3). The calculation of Area Under the Receiver Operating Characteristic (AUROC) and False Positive Rate at 95% True Positive Rate (FPR95) is based on these scores.

HFTT training and inference costs. The backbone model for HFTT remains untrained, while only trainable parameters (trainable embeddings) defined in the model output space are updated. Thus, the cost of this update process is almost equivalent to the forward propagation cost of the synthetic textual data. For the corpus (∼370k samples) and CLIP utilized in the experiments, the update process takes less than 2 minutes with a single V100 GPU. During inference time, the computation of HFTT

Table 1: Comparison of HFTT and competitive baselines with and without in-distribution image requirements on the ImageNet-1K dataset. The best and second-best results are indicated in bold and underlined, respectively. Our method surpasses even strong baselines that utilize in-distribution images. This complements our analysis in Section 3.1, demonstrating that textual data can substitute for visual data in such tasks.

	OOD Dataset									
Method	iNaturalist		SUN		Places		Texture		Average	
	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
In-distribution images required										
Mahalanobis	99.33	55.89	99.41	59.94	98.54	65.96	98.46	64.23	98.94	61.51
MSP	40.17	89.76	63.99	79.40	63.50	80.19	67.01	79.33	58.67	82.17
KNN	29.17	94.52	35.62	92.67	**39.61**	**91.02**	64.35	85.67	42.19	90.97
NPOS	**16.58**	**96.19**	43.77	90.44	45.27	89.44	46.12	**88.80**	37.94	91.22
In-distribution images not required										
Energy	34.70	90.55	32.33	90.58	40.29	89.32	51.24	72.36	39.64	85.70
MaxLogit	35.03	89.46	32.86	90.33	41.15	89.60	68.17	75.63	44.30	86.26
ZOC	87.30	86.09	81.51	81.20	73.06	83.39	98.90	76.46	85.19	81.79
MCM	34.33	91.36	32.27	91.86	47.48	88.68	50.90	87.52	41.25	89.86
HFTT (ours)	27.44	93.27	**19.24**	**95.28**	43.54	90.26	**43.08**	88.23	**33.33**	**91.76**

involves obtaining cosine similarities between trainable embeddings and input embeddings. This cost amounts to 2×(batch size)×(embedding dimension)×(the number of trainable embeddings) FLOPS, which is negligible compared to the inference cost of the entire model.

4.2 Out-of-Distribution Detection

In OOD detection experiments, we utilize weights of zero-shot classifiers of pre-trained VLMs as task embeddings for HFTT. In-distribution textual data are obtained via combinations of various prompt templates and class names of in-distribution data, which, in our experimental setting, are equivalent to 1,000 ImageNet classes. We adopt the prompt set released by OpenAI for prompt ensembling[4] as prompt templates. The comparison results are reported in Table 1. Despite the simple and lightweight nature of the proposed method, it outperforms even strong baselines that utilize images, on average.

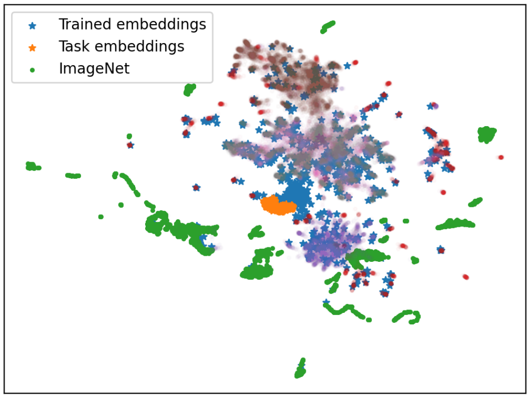

Figure 3: UMAP [35] visualization of the joint embedding space of CLIP. The dispersed, transparent markers represent the OOD data samples used in our experiment (iNaturalist: brown; SUN: grey; Places: pink; Texture: purple; NINCO [3]: red). The trained embeddings (blue stars) are located in a sub-region of the embedding space occupied by OOD data. We trained 2000 embeddings for this plot. It is important to note that these trainable embeddings did not incorporate any information about the OOD data during their training time.

To understand how the trained embeddings provide the task embeddings with informative signals for identifying OOD data points, we visually analyze the joint embedding space of CLIP in Figure 3. Even though visual OOD data were not involved in the training process, the trained embeddings are positioned on a sub-region of the embedding space inhabited by OOD data and thus can function as additional pointers for where OOD data lie in the joint embedding space. Therefore, this deliberate positioning of trained embeddings precludes the task embeddings from accidentally confusing out-distribution data as in-distribution by refining the decision boundary to intricately separate in-distribution and out-distribution regions.

[4] https://github.com/openai/CLIP

Table 2: Comparison of HFTT with state-of-the-art methods for OOD detection that do not require in-distribution images, conducted on the Hate dataset. The best result in each column is in bold. HFTT outperforms baseline approaches, showing that it can effectively be used for the general purpose of unwanted data detection.

Method	Innocuous Dataset											
	iNaturalist		SUN		Places		Texture		NINCO		Average	
	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
Energy	12.30	97.53	3.88	98.51	13.87	97.40	39.70	94.98	26.89	96.12	17.43	97.10
MaxLogit	23.65	96.89	18.49	97.49	27.84	96.48	33.33	96.00	33.03	95.99	25.82	96.71
ZOC	87.76	71.05	66.51	85.23	69.96	82.57	65.48	83.22	81.06	78.36	74.15	84.09
MCM	80.53	76.70	87.54	69.38	81.37	74.12	60.39	84.97	81.95	78.00	77.45	76.29
CLIPN	47.71	92.78	36.36	95.16	40.62	94.52	53.36	92.36	68.40	89.58	49.29	92.88
NegLabel	**0.03**	**99.84**	1.09	99.10	5.16	98.50	3.56	98.82	12.62	97.86	4.49	98.82
HFTT (ours)	0.17	99.44	**1.05**	**99.13**	**4.38**	**98.60**	**1.73**	**99.08**	**4.18**	**98.52**	**1.83**	**99.06**

This visualization result can be attributed to two methodological characteristics that are unique to our method. First, our method directly optimizes the trainable embeddings and is not bounded by the modality gap between texts and images. Second, our method successfully places the trained embeddings on top of OOD data only through textual data; this consolidates that textual data can replace visual data, providing strong empirical support for the theory presented in Section 3.1. Together, these two aspects of HFTT yield the joint embedding space as illustrated in Figure 3.

4.3 Hateful Image Detection

In this task, the hateful data that contains offensive content against Muslims and Jews is treated as in-distribution data, whereas innocuous data void of such content is treated as out-distribution data. Consequently, embeddings of distinct phrases from a collection of offensive and hateful phrases, provided as part of the Hate dataset, are utilized as task embeddings. The entire set of offensive and hateful phrases is employed as in-distribution textual data.

The Mahalanobis, MSP, KNN, and NPOS methods necessitate the construction of an in-distribution image dataset. Therefore, they should not be used as methods for unethical image detection tasks such as hate image detection, as doing so would require the construction of unethical image datasets, leading to numerous ethical problems such as direct or indirect leakage of sensitive information. In contrast, HFTT requires no usage of any image, thus it can be applied to any unethical image detection task without ethical concerns. To highlight the differences between traditional OOD detection methods and HFTT, we include two additional baselines [49, 24] and one extra dataset [3].

In Table 2, we can observe that most OOD detection methods show significantly lower performance compared to HFTT. These results arise because OOD detection methods assume a classification problem with clear distinctions between classes. In tasks dealing with abstract concepts, the boundaries between data clusters within the in-distribution are ambiguous, which results in the underperformance of existing OOD detection methods. NegLabel shows different results compared to traditional OOD detection methods but still falls short of our proposed approach. We provide a further comparison of our method to CLIPN [49] and NegLabel [24] in Appendix B.

To further study the generalizability of HFTT, we observe the effectiveness of HFTT in low-quality image detection [17] and within the medical image domain [6, 16, 48]. The findings demonstrate the potential extension of HFTT's applicability beyond conventional OOD detection tasks. The results of these experiments and an ablation study on hyper-parameters are provided in Appendices B and C.

5 Conclusion and Limitation

In this paper, we proposed a novel methodology for identifying undesirable content hidden within visual datasets. Close theoretical scrutiny of the joint embedding space of VLMs led to the development of HFTT, an efficient framework for training detectors to automatically identify unwanted visual content by leveraging solely textual data together with pre-trained VLMs. HFTT is comprised of a creative objective function that markedly diminishes human involvement in data annotation and the textual data synthesis technique in HFTT that can simulate the usage of unknown visual data distributions in the training process without additional cost. The distinctive attributes of HFTT broaden its applicability from a clearly-scoped OOD detection task to a far more general set of tasks that are more abstract. Because HFTT requires some type of VLM as the base model, its capabilities are bounded by the representative capacity of pre-trained VLMs. This dependency on pre-trained VLMs makes the use of HFTT in tasks that VLMs struggle with challenging.

Impact Statements. This paper contributes to the growing field of data curation and selection research. As datasets for training large AI models expand without adequate safeguards, identifying unwanted data points, such as biased or offensive content, from training datasets is becoming crucial. We believe our work will make a positive contribution to this area, opening up new possibilities for the effortless removal of unwanted visual data. While our method could potentially be misused for content censorship, we believe the positive impact it provides significantly outweighs these concerns.

Acknowledgments and Disclosure of Funding

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No.RS-2021-II211343, 2022-0-00959, RS-2022-II220959, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A3B1077720, 2022R1A5A708390811), and the BK21 FOUR program of the Education and the Research Program for Future ICT Pioneers, Seoul National University in 2024.

References

[1] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[3] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023. URL `https://proceedings.mlr.press/v202/bitterwolf23a.html`.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

[11] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576, 2022.

[12] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=j5NrN8ffXC`.

[13] Felipe González-Pizarro and Savvas Zannettou. Understanding and detecting hateful content using contrastive learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 257–268, 2023.

[14] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2808–2817, June 2023.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[16] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[18] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Hkg4TI9xl`.

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

[20] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HyxCxhRcY7`.

[21] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8759–8773. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/hendrycks22a.html`.

[22] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.

[24] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. *arXiv preprint arXiv:2403.20078*, 2024.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[26] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[27] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/abdeb6f575ac5c6676b747bca8d09cc2-Paper.pdf`.

[28] Kyubum Lee, Maria Livia Famiglietti, Aoife McMahon, Chih-Hsuan Wei, Jacqueline Ann Langdon MacArthur, Sylvain Poux, Lionel Breuza, Alan Bridge, Fiona Cunningham, Ioannis Xenarios, et al. Scaling up data curation using deep learning: an application to literature triage in genomic variation resources. *PLoS computational biology*, 14(8):e1006390, 2018.

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[31] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding CLIP latents for zero-shot captioning via text-only training. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Lt8bMlhiwx2`.

[32] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=H1VGkIxRZ`.

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[34] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

[35] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[36] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, 2022.

[37] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, 2022.

[38] Muhammad Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Mateusz Kozinski, Rogerio Feris, and Horst Bischof. LaFTer: Label-free tuning of zero-shot classifier using language and unlabeled image collections. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=elPtHcfjpH`.

[39] Sangha Park, Jisoo Mok, Dahuin Jung, Saehyung Lee, and Sungroh Yoon. On the powerfulness of textual outlier exposure for visual ood detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[41] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.

[42] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[44] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[45] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=JHklpEZqduQ`.

[46] Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. *arXiv preprint arXiv:2303.02966*, 2023.

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

[48] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018.

[49] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023.

[50] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

[51] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2021.

[52] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

A Proofs

Theorem 1. For the quadratic loss function $L\left(u, y; \theta\right) = \left(1 - y\theta^\top u\right)^2$, the optimal cosine-similarity classifier $\theta^\star$ that classifies sets U_{-1} and U_{+1} is

$$
\begin{aligned}
&\arg\min_{\theta} \mathop{\mathbb{E}}_{u \in U_{-1}} \left[L\left(u, -1; \theta\right)\right] + \mathop{\mathbb{E}}_{u \in U_{+1}} \left[L\left(u, +1; \theta\right)\right] \\
&= \arg\min_{\theta} \left(1 + \theta^\top u_{-1}\right)^2 + \left(1 - \theta^\top u_{+1}\right)^2 = \frac{u_{+1} - u_{-1}}{\|u_{+1} - u_{-1}\|}.
\end{aligned}
$$

Proof. Our assumption validates the following equations:

$$
\theta^\top \theta = u^\top u = v^\top v = 1, \Sigma_{+1} + \Sigma_{-1} = \epsilon \mathbb{I}, \|u_{+1}\| = \|u_{-1}\|,
$$

where Σ_y and $\mathbb{I}$ denote the covariance matrix of U_y and the identity matrix, respectively, and $\epsilon > 0$ is a constant. Then,

$$
\begin{aligned}
&\arg\min_{\theta} \mathop{\mathbb{E}}_{u \in U_{-1}} \left[L\left(u, -1; \theta\right)\right] + \mathop{\mathbb{E}}_{u \in U_{+1}} \left[L\left(u, +1; \theta\right)\right] \\
&\quad= \arg\min_{\theta} \mathop{\mathbb{E}}_{u \in U_{-1}} \left[\left(1 + \theta^\top u\right)^2\right] + \mathop{\mathbb{E}}_{u \in U_{+1}} \left[\left(1 - \theta^\top u\right)^2\right] \\
&\quad= \arg\min_{\theta} \left(\mathop{\mathbb{E}}_{u \in U_{-1}} \left[1 + \theta^\top u\right]\right)^2 + \left(\mathop{\mathbb{E}}_{u \in U_{+1}} \left[1 - \theta^\top u\right]\right)^2 + \theta^\top \Sigma_{+1} \theta + \theta^\top \Sigma_{-1} \theta \\
&\quad= \arg\min_{\theta} \left(\mathop{\mathbb{E}}_{u \in U_{-1}} \left[1 + \theta^\top u\right]\right)^2 + \left(\mathop{\mathbb{E}}_{u \in U_{+1}} \left[1 - \theta^\top u\right]\right)^2 + 2\epsilon \\
&\quad= \arg\min_{\theta} \left(1 + \theta^\top u_{-1}\right)^2 + \left(1 - \theta^\top u_{+1}\right)^2.
\end{aligned}
$$

The gradient of the objective function with respect to θ is

$$
-2\left(1 - \theta^\top u_{-1}\right) u_{-1} + 2\left(1 + \theta^\top u_{+1}\right) u_{+1}.
$$

Therefore, the optimal cosine-similarity classifier $\theta^\star$ satisfies the following equation:

$$
\left(1 - \theta^{\star\top} u_{-1}\right) u_{-1} = \left(1 + \theta^{\star\top} u_{+1}\right) u_{+1}.
$$

Then,

$$
1 - \theta^{\star\top} u_{-1} = 1 + \theta^{\star\top} u_{+1} \text{ or } 1 - \theta^{\star\top} u_{-1} = -1 - \theta^{\star\top} u_{+1}.
$$

The second equation is not true for any u_{-1} and u_{+1}. Hence,

$$
\theta^{\star\top} = \frac{u_{+1} - u_{-1}}{\|u_{+1} - u_{-1}\|}.
$$

□

Corollary 1. The classifier $\theta^\star$, with respect to V_{-1} and V_{+1}, satisfies the double inequalities of

$$
\mathop{\mathbb{E}}_{v \in V_{-1}} \left[\theta^{\star\top} v\right] < 0 < \mathop{\mathbb{E}}_{v \in V_{+1}} \left[\theta^{\star\top} v\right].
$$

Proof. Based on the inequalities $u_{+1}^\top v_{+1} > u_{+1}^\top v_{-1}$ and $u_{-1}^\top v_{+1} < u_{-1}^\top v_{-1}$,

$$
\mathop{\mathbb{E}}_{v \in V_{-1}} \left[\theta^{\star\top} v\right] = \frac{u_{+1}^\top v_{-1} - u_{-1}^\top v_{-1}}{\|u_{+1} - u_{-1}\|} < 0 < \mathop{\mathbb{E}}_{v \in V_{+1}} \left[\theta^{\star\top} v\right] = \frac{u_{+1}^\top v_{+1} - u_{-1}^\top v_{+1}}{\|u_{+1} - u_{-1}\|}.
$$

□

B Additional Experiments

Low-quality image detection. We additionally demonstrate the applicability of our method, HFTT, in detecting low-quality images, which are commonly unwanted visual data beyond OOD and hate images. Specifically, we assume the task of detecting corrupted images lurking within a raw visual dataset consisting of 1000 ImageNet classes. For this experiment, we employ ImageNet and ImageNet-C as in-distribution and out-distribution data, respectively. As shown in Table 3, HFTT consistently surpasses existing methods in the detection of corrupted images.

Table 3: Comparison of HFTT with baselines on the low-quality image detection.

Method	FPR	AUROC
MSP	64.17	83.94
Energy	99.99	09.16
MaxLogit	78.01	68.47
MCM	51.54	89.06
HFTT (ours)	**42.13**	**92.81**

Table 4: Comparison of HFTT with MCM on the medical image datasets.

Method	OOD Dataset			
	PVQA		PCAM	
	FPR	AUROC	FPR	AUROC
MCM	95.44	47.10	71.49	68.74
MCM + description	86.84	60.39	84.50	43.44
HFTT	22.58	93.60	8.07	96.94
HFTT + description	13.72	97.05	4.95	98.35
HFTT + description + corpus engineering	6.24	98.69	4.33	98.73

Medical image domain. We compare the performance of HFTT and MCM in the medical image domain. Specifically, we treat the ISIC-18 skin lesion diagnosis dataset [6] as in-distribution and the PathVQA [16] and PatchCamelyon [48] datasets as out-distribution. The ISIC-18 skin lesion diagnosis dataset is an image classification benchmark for seven skin disease categories. We apply MCM and HFTT to CLIP on the seven disease categories. Table 4 reveals a significantly low detection performance of MCM, attributed to the limited medical domain knowledge of CLIP. Even appending descriptions (generated by GPT-4) to the disease names does not yield favorable results for MCM (+ description). In contrast, our proposed method achieves significantly better results by leveraging the model knowledge and the medical-related information in the corpus. The utilization of medical-related information within the corpus by HFTT is evidenced by the fact that further improvements can be achieved by modifying the corpus to align with the medical domain (+ corpus engineering).

The experimental results on other pre-trained models. As discussed in Section 3.1 of our paper, our method assumes that text and image are well-aligned through contrastive learning, similar to CLIP. Therefore, if CLIP is used as the vision encoder, the text encoder must also be CLIP. If text and image are well-aligned, our method can be applied to models other than CLIP. Here, we provide the results for CLIP-L/14, BLIP-B/16, and BLIP-L/16 in addition to the CLIP-B/16 used in our study. Table 5 demonstrates that our method is effective across various vision-language models.

HFTT vs. CLIPN and NegLabel. NegLabel [24] constructs an OOD corpus by selecting texts distant from the in-distribution texts from a predefined corpus, then compares the distances between the input image and those texts in the CLIP embedding space to detect OOD. While NegLabel shows high OOD detection performance on ImageNet (see Table 7), it has the following limitations compared to our method: 1) Although NegLabel does not require training additional parameters, it must compute the embeddings of all texts in the corpus and measure their similarity to the in-distribution texts to find the optimal OOD corpus for a given in-distribution. Our training method also requires nearly the same cost as obtaining the embeddings of all texts within a predefined corpus and calculating the similarities between those embeddings and the task+trainable embeddings, as discussed in Section 4.1. Thus, NegLabel and our method require the same level of optimization cost; 2) Since NegLabel uses the embeddings of the determined OOD corpus as they are, it falls behind our method, which has trainable parameters, in terms of generalization. To demonstrate this, we further compare our method and NegLabel in the medical image domain. Specifically, we treat the ISIC-18 skin lesion diagnosis dataset [1] as in-distribution and the PathVQA [16] and PatchCamelyon [48] datasets as out-of-distribution. The ISIC-18 skin lesion diagnosis dataset is an image classification benchmark for seven skin disease categories. Table 6 illustrates the limitations of NegLabel in terms of generalization. While NegLabel fails to construct an effective OOD corpus for the medical image

Table 5: Comparison of HFTT and competitive baselines on the ImageNet-1K dataset. The best result in each column is in bold. Our method outperforms all baselines on both variants of CLIP and BLIP, demonstrating that it can be used to improve the OOD detection performance of various VLMs.

Model	Method	OOD Dataset									
		iNaturalist		SUN		Places		Texture		NINCO	
		FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
CLIP-B	MSP	34.63	91.35	32.06	91.86	47.62	88.86	49.78	87.65	72.83	71.98
	Energy	34.70	90.55	32.33	90.58	40.29	89.32	51.24	72.36	70.06	73.85
	MaxLogit	35.03	89.46	32.86	90.33	41.15	89.60	68.17	75.63	68.96	74.24
	MCM	34.33	91.36	32.27	91.86	47.48	88.68	50.90	87.52	73.26	71.98
	HFTT (ours)	**27.32**	**93.28**	**19.68**	**95.20**	**43.24**	**90.32**	**43.26**	**88.20**	**70.08**	**74.61**
CLIP-L	MSP	26.66	94.20	22.37	94.37	36.82	92.45	52.83	86.57	67.27	78.70
	Energy	30.84	91.25	25.94	94.10	32.94	92.30	64.33	79.26	63.49	79.72
	MaxLogit	32.76	90.96	26.48	92.96	**31.88**	92.39	72.08	73.85	**60.67**	**81.07**
	MCM	26.96	94.19	22.77	94.37	36.74	92.44	52.66	86.56	68.16	78.65
	HFTT (ours)	**24.10**	**94.58**	**17.80**	**95.39**	33.83	**93.09**	**52.06**	**86.58**	69.19	78.98
BLIP-B	MSP	64.70	82.22	30.38	91.06	71.40	78.82	76.99	81.30	**71.47**	72.07
	Energy	67.15	79.30	45.21	89.07	70.28	77.49	91.24	75.38	80.29	**77.20**
	MaxLogit	69.57	75.44	69.57	71.19	69.86	76.26	93.55	60.31	88.58	56.37
	MCM	64.41	**82.29**	30.21	91.05	70.53	79.32	75.84	81.55	71.56	72.02
	HFTT (ours)	**63.28**	82.22	**19.16**	**95.12**	**68.48**	**79.50**	**63.74**	**84.53**	72.12	73.86
BLIP-L	MSP	51.20	87.91	22.37	93.86	61.63	84.68	64.85	85.28	65.96	78.29
	Energy	45.63	87.23	33.94	90.29	55.73	85.91	72.38	82.16	71.23	77.49
	MaxLogit	44.59	86.94	35.56	86.45	**50.96**	**86.46**	86.38	71.22	79.78	67.59
	MCM	50.75	88.03	22.34	93.88	60.88	85.38	64.71	**85.39**	66.04	78.32
	HFTT (ours)	**44.24**	**89.88**	**6.81**	**98.40**	62.20	84.16	**63.35**	83.39	**64.82**	**80.46**

Table 6: OOD detection in the medical image domain.

Method	PVQA		PCAM	
	FPR	AUROC	FPR	AUROC
CLIPN	35.47	84.64	3.10	98.76
NegLabel	37.44	94.11	48.07	94.86
HFTT (ours)	13.72	97.05	4.95	98.35

dataset, our method achieves significantly higher performance by learning optimal embeddings for detection.

CLIPN utilizes an additional "no" text encoder alongside CLIP. This additional text encoder predicts the probability that a given object is not present in an image. Thus, CLIPN predicts whether a given image is in-distribution or out-distribution by using the original CLIP text encoder and the "no" text encoder to estimate the probabilities, respectively. Images with a low probability of being in-distribution and a high probability of being out-distribution are identified as OOD.

Although CLIPN achieves high OOD detection performance on ImageNet (see Table 7), it has the following limitations compared to our method:

1) CLIPN requires significantly higher inference costs due to the use of an additional text encoder; 2) While our method requires lightweight training that does not involve images, CLIPN demands extensive and expensive training of the "no" text encoder on large vision-language datasets; 3) CLIPN can only be applied to tasks where the distinction between in-distribution and out-distribution is clear and straightforward, such as classification datasets. This is because all training images must be classified as either "yes" or "no" images. Therefore, it is unsuitable for tasks dealing with abstract concepts, such as hateful image detection, as discussed in Section 4.3 of our paper; 4) Our method can be easily applied to any detection task defined in natural language, whereas CLIPN shows significantly degraded performance for in-distribution tasks that fall outside the training distribution of the "no" text encoder. In terms of applicability, our proposed method surpasses CLIPN. To demonstrate this,

Table 7: OOD detection performance on ImageNet in-distribution (average for Texture, Places, SUN, and iNaturalist).

Method	FPR	AUROC
CLIPN	31.10	93.10
NegLabel	25.40	94.21
HFTT (ours)	33.33	91.76

Table 8: Results of using different textual data synthesis methods. HFTT outperforms other methods that are more complex.

					OOD Dataset							
Method	iNaturalist		SUN		Places		Texture		NINCO		Average	
	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
WTO	29.26	92.23	23.31	93.93	40.18	90.77	42.96	88.01	69.07	76.54	40.96	88.30
CTO	27.53	91.66	38.41	86.78	43.31	89.31	60.13	79.64	76.25	69.08	49.13	83.29
DTO	29.32	92.32	24.57	94.04	43.37	89.63	42.26	89.34	70.36	74.76	41.98	88.02
Caption	54.07	79.49	57.15	74.23	63.44	77.04	41.21	91.12	89.30	56.42	47.65	86.28
Dedupl.	28.17	93.03	21.08	94.75	43.67	90.10	42.69	88.54	68.86	75.22	40.89	88.33
Ours	27.44	93.27	19.24	95.28	43.54	90.26	43.08	88.23	70.15	74.48	40.69	88.30

we further compare our method with CLIPN in the medical image domain. Table A illustrates the limitations of CLIPN in terms of generalization. While CLIPN effectively detects PCAM, it exhibits very low detection performance on PVQA. In contrast, our method achieves high performance on both OOD tasks.

C Ablation Study

In this section, we analyze how different components affect the performance of HFTT.

Textual data synthesis method. While we propose a textual data synthesis method that incurs no additional cost, alternative approaches beyond this can also be explored. We apply the following methods in conjunction with HFTT and list their results in Table 8:

1. WTO, CTO, DTO: Recently, Park *et al.* proposed a method that utilizes textual outliers instead of visual outliers in outlier exposure [19]. For our experiments, we use word-level textual outliers (WTO) generated using in-distribution images along with CLIP and BERT [8], caption-level textual outliers (CTO) generated by an image captioning model [30], and description-level textual outliers (DTO) created using a large language model.
2. Caption: We can consider the extensive use of image captions from LAION-400M [42] to substitute for the entire visual data distribution.
3. Deduplication: To experimentally compare Eq. 2 and 3, we applied HFTT after removing words from our word set that have meanings identical to ImageNet classes as much as possible.

Table 8 reveals that textual outliers generated using additional models and in-distribution images show comparable or inferior performance to our textual data synthesis method. Furthermore, the captions results suggest that heavily relying on image captions does not effectively enhance the average detection performance for various OOD data. Lastly, there appears to be no discernible performance difference between the application of Eq. 1 and 4. This indicates that our proposed loss minimizes the need for human labor by eliminating the process of selecting out-of-distribution data without sacrificing performance.

The focal loss hyper-parameter. HFTT incorporates the concept of focal loss to shape the decision boundary of detectors near the in-distribution. We observe its effect by incrementally increasing the focal loss hyper-parameter γ from zero. Table 9 demonstrates that using the focal loss ($\gamma > 0$) generally leads to better performance compared to not using it ($\gamma = 0$).

Table 9: Results of using different values of γ for the focal loss. The performance of HFTT appears to be relatively robust to changes in the choice of γ, with the adoption of focal loss with $\gamma > 0$ generally leading to improved results.

					OOD Dataset							
γ	iNaturalist		SUN		Places		Texture		NINCO		Average	
	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
0	27.67	93.08	20.62	94.98	43.97	90.12	44.42	87.57	69.23	75.22	41.18	88.19
1	27.44	93.27	19.24	95.28	43.54	90.26	43.08	88.23	70.15	74.48	40.69	88.30
2	27.10	93.32	19.48	95.20	43.17	90.32	42.95	88.33	70.19	74.40	40.58	88.31
3	27.03	93.32	19.56	95.17	43.21	90.32	42.82	88.38	70.42	74.26	40.61	88.29

Table 10: Results of changing the temperature of the final Softmax layer.

					OOD Dataset							
Temp.	iNaturalist		SUN		Places		Texture		NINCO		Average	
	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
1.0	94.97	46.92	76.17	48.52	93.00	55.90	99.80	23.96	94.96	56.42	91.78	46.34
0.1	93.52	50.17	95.22	51.30	91.67	58.43	99.75	25.37	94.62	56.93	94.96	48.44
0.01	27.44	93.27	19.24	95.28	43.54	90.26	43.08	88.23	70.15	74.48	40.69	88.30

Temperature. In HFTT, a temperature parameter is utilized for computing $p\left(x\right)$. CLIP learns a temperature parameter during their pre-training phase, and we employ these same learned temperature values in all of our experiments. Table 10 illustrates that modifying the temperature value may reduce the efficacy of HFTT.

The number of trainable embeddings. If the dimensionality of the data manifold in the joint embedding space of VLMs is low, HFTT can be effective even with a small number of trainable embeddings. To validate this, we present the OOD detection performance of HFTT in Table 11, illustrating how it varies with the number of trainable embeddings (N). Remarkably, HFTT can improve OOD detection performance on ImageNet, which has 1,000 classes, even with a very limited number of trainable embeddings. This is possible due to the inherently low dimension of data in the actual model output space [1].

Table 11: Results of shrinking or expanding the number of trainable embeddings (N).

					OOD Dataset							
N	iNaturalist		SUN		Places		Texture		NINCO		Average	
	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
5	27.50	93.26	19.96	95.07	43.55	90.25	43.07	88.23	70.15	74.38	40.85	88.24
10	27.44	93.27	19.24	95.28	43.54	90.26	43.08	88.23	70.15	74.48	40.69	88.30
100	27.40	93.30	19.63	95.14	43.67	90.28	42.93	88.27	70.14	74.53	40.75	88.30
2000	27.32	93.28	19.68	95.20	43.24	90.32	43.26	88.20	70.08	74.61	40.72	88.32

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Abstract and Section 1

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 5

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

   Justification: Section 3 and Appendix A

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: URL in Abstract and Section 4

   Guidelines:

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

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

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

   Answer: [Yes]

   Justification: URL in Abstract

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: URL in Abstract and Section 4

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [No]

   Justification: We report the results over five runs. The variances of the results are close to zero so reporting them would be redundant.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: Section 4.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: Our submission conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: Section 5

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: This paper does not release any data or model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Section 4 and Appendix

    Guidelines:

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

13. **New Assets**

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

    Answer: [Yes]

    Justification: URL in abstract and Section 4

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [NA]
Justification: This paper does not involve crowdsourcing nor research with human subjects.
Guidelines:

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