Concept Matching with Agent for Out-of-Distribution Detection

Yuxiao Lee[1], Xiaofeng Cao[1*], Jingcai Guo[2], Wei Ye[3], Qing Guo[4], Yi Chang[1, 5]
[1]School of Artificial Intelligence, Jilin University, China
[2]The Hong Kong Polytechnic University
[3]College of Electronic and Information Engineering, Tongji University, China
[4]CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore
[5]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, Ministry of Education, China
yuxiao9922@mails.jlu.edu.cn, xiaofengcao@jlu.edu.cn,
jc-jingcai.guo@polyu.edu.hk, yew@tongji.edu.cn, tsingqguo@ieee.org, yichang@jlu.edu.cn

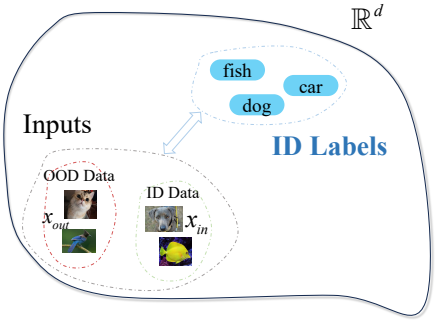
Abstract

The remarkable achievements of Large Language Models (LLMs) have captivated the attention of both academia and industry, transcending their initial role in dialogue generation. To expand the usage scenarios of LLM, some works enhance the effectiveness and capabilities of the model by introducing more external information, which is called the agent paradigm. Based on this idea, we propose a new method that integrates the agent paradigm into out-of-distribution (OOD) detection task, aiming to improve its robustness and adaptability. Our proposed method, Concept Matching with Agent (CMA), employs neutral prompts as agents to augment the CLIP-based OOD detection process. These agents function as dynamic observers and communication hubs, interacting with both In-distribution (ID) labels and data inputs to form vector triangle relationships. This triangular framework offers a more nuanced approach than the traditional binary relationship, allowing for better separation and identification of ID and OOD inputs. Our extensive experimental results showcase the superior performance of CMA over both zero-shot and training-required methods in a diverse array of real-world scenarios.

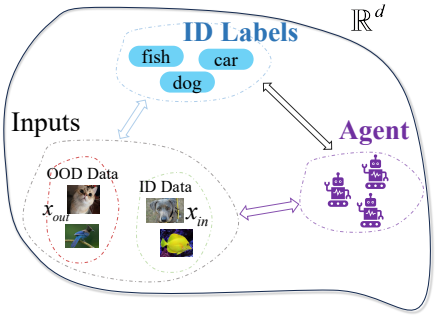
Code — https://github.com/yuxiaoLeeMarks/CMA

1 Introduction

The emergence and development of Large Language Models (LLMs) (Chang et al. 2024; Zhao et al. 2023; Brown et al. 2020; Achiam et al. 2023) have significantly reshaped the landscape of Artificial Intelligence (AI), marking a pivotal breakthrough in both academic research and practical applications. These models have not only revolutionized the way we generate conversations but also demonstrated their capacity as intermediary agents with more nuanced roles, facilitating the accomplishment of myriad tasks with unprecedented efficiency and adaptability (Wang et al. 2024; Xi et al. 2023; Zeng et al. 2023). The Agent paradigm has been extensively applied across multiple domains and tasks, playing a profound role (Qian et al. 2023a,b; Hong et al. 2023). The core of this paradigm lies in introducing **external information to change the input distribution of the model**, thereby enhancing the model's performance.



(A) Binary Relationship



(B) Vector Triangle Relationship

Figure 1: In OOD detection, the Vector Triangle Relationship alters the traditional Binary Relationship by introducing Agents, thereby more effectively processing and distinguishing between ID data and OOD data.

An important and challenging task within the field of machine learning is to enhance the robustness of models across diverse scenarios. When an artificial intelligence system encounters data that significantly deviates from its training data distribution, OOD detection becomes crucial for ensuring its reliability and robustness. In the past, most OOD detection methods employed single-modal learning (Yang et al. 2021; Liu et al. 2021; Hendrycks and Gimpel 2016). As CLIP (Radford et al. 2021) has demonstrated astonishing performance across various downstream tasks, an increasing number of CLIP-based methods for out-of-distribution (OOD) detection have emerged (Ramesh et al. 2022; Wang et al.

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2022a; Crowson et al. 2022; Wang, Xing, and Liu 2021; Gao et al. 2024).

Question. However, previous OOD detection methods, whether single-modal learning or CLIP-based approaches, typically rely on binary relationship to differentiate between in-distribution (ID) data and OOD data (Figure 1). These methods include solely using ID data to construct boundaries or employing a combination of ID and OOD data to demarcate their respective domains. While these methods are effective to some extent, they lack the flexibility and adaptability needed to handle the dynamic complexity of real-world data distributions.

Motivation. In the realm of OOD detection, the predominant methodology consists of binary segmentation between ID and OOD data, typically facilitated by scoring functions. The effectiveness of this prototype is closely tied to the sophisticated design of these scoring functions. Inspired by the remarkable success of the agent-based paradigm, introducing external information as agents in CLIP-based OOD detection framework could reshape the distribution of ID and OOD inputs. Structurally, this paradigm shift from binary segmentation to a vector triangle relationship relational framework holds substantial potential for uncovering deeper insights, potentially transforming the interplay between ID labels and data inputs (Figure 1).

Our scheme. In this paper, we propose the **C**oncept **M**atching with **A**gent (**CMA**) methodology, which integrates ***neutral textual concept prompts in natural language*** as ***Agents*** within the CLIP-based OOD detection framework. Our framework is illustrated in Figure 2. These agents serve dual roles: as observers and as intermediate hubs that facilitate the interaction between ID Labels and Data inputs. By doing so, we aim to establish a vector triangle relationship among the ID labels, data inputs, and the agents themselves. In this triangular vector relationship (Figure 1), the score of OOD data is diminished due to the collision effect of the Agents, thereby widening the gap between the scores of ID data and OOD data. In other words, images closer to the ID class are less likely to be influenced by neutral text concepts, whereas images from the OOD class are more susceptible to these concepts, resulting in lower scores. The method to achieve our idea is derived from our profound insights into the Language-Vision representation (See Section 3). Building on these insights, we formulate the entire Concept Matching with Agent (CMA) framework.

In summary, our proposed method, CMA, possesses three distinct advantages. (1) Our approach obviates the need for training data, enabling zero-shot OOD detection while leveraging the collision between Agent and both ID and OOD data to effectively widen the gap between the two, ensuring robust performance. This stands in stark contrast to traditional OOD detection methods, which rely on extensive external data for intricate training (Yang et al. 2021; Liu et al. 2021). (2) Our method exhibits remarkable scalability, allowing for the tailoring of specialized Agents to suit various scenarios, thereby further enhancing performance (See Section 5). This is facilitated by the flexibility of our triangular vector relationship, which can enhance the impact of certain OOD images through specific Agents, thereby reducing their scores. (3) It is noteworthy that the CMA maintains robustness against both hard OOD inputs, encompassing both semantic hard OODs (Winkens et al. 2020) and spurious OODs (Ming, Yin, and Li 2022). This makes our approach a truly practical and viable option. The contributions of our study are summarized as follows:

- Drawing upon the concept of Agents in LLMs, we propose the incorporation of agent-based observation into OOD detection. By facilitating interactions among agents, ID labels, and data inputs, we establish a vector triangle relationship for them. This structured shift diverges from the conventional binary frameworks, offering enhanced flexibility in application scenarios and providing a novel analytical perspective on OOD detection.
- We propose a novel CLIP-based OOD Detection framework. Compared to previous methods, our approach more effectively achieves the objective of OOD Detection: it widens the gap between ID and OOD, and possesses enhanced versatility and practicality.
- We conducted experiments on various datasets with distinct ID scenarios and demonstrated that CMA achieves superior performance across a wide range of real-world tasks. Compared to most existing OOD detection methods, CMA brings substantial improvements to the large-scale ImageNet OOD benchmark.

2 Preliminaries

Contrastive Vision-Language Models. In comparison to traditional CNN architectures, the ViT (Dosovitskiy et al. 2020) leverages the Transformer Encoder framework (Vaswani et al. 2017) to accomplish the task of image classification, realizing the possibility of utilizing language model architectures for visual tasks. This provides insights into the field of vision-language representation learning, with CLIP (Radford et al. 2021) being a notable representative. CLIP employs self-supervised contrastive objectives to embed images and their corresponding textual descriptions into a shared feature space, achieving alignment between the two. Structurally, CLIP, which utilizes a dual-stream architecture, comprises an image encoder $\mathcal{I} : x \rightarrow \mathbb{R}^d$ and a text encoder $\mathcal{T} : t \rightarrow \mathbb{R}^d$. After pretraining on a dataset of 400 million text-image pairs, the joint visual-language embedding of CLIP associates objects in various patterns. Due to the robust performance of CLIP, several OOD detection methods based on it have emerged. Nevertheless, the challenge of how to better utilize Language-Vision representation for OOD detection remains a difficult yet significant issue.

Zero-shot OOD Detection. For traditional OOD detection frameworks (Hendrycks and Gimpel 2016), a common assumption is made of a typical real-world scenario wherein classifier f are trained on ID data categorized as $\mathcal{Y}_{in} = \{1, ..., C\}$ and subsequently deployed in an environment containing samples from unknown classes $y \notin \mathcal{Y}_{in}$, outside the distribution of the ID. The classifier f is then tasked with

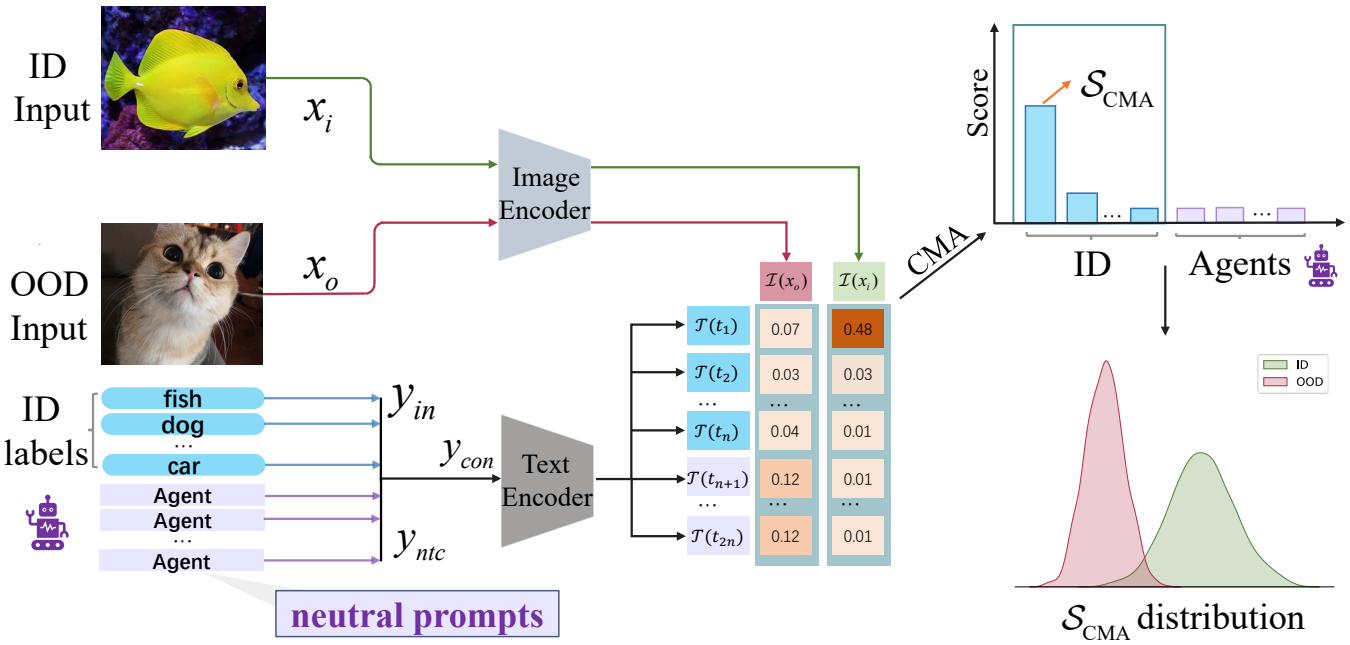


Figure 2: **Overview of Concept Matching with Agent (CMA) framework.** The input image x undergoes Image Encoder $\mathcal{I}$ to produce an image embedding. The concatenation of the ID labels $\mathcal{Y}_{in}$ and Agents $\mathcal{Y}_{ntc}$ is then subjected to Text Encoder $\mathcal{T}$ to generate a text embedding. The similarity between the image and text embeddings is computed, with a higher result indicating a greater degree of similarity (darker shading denotes higher similarity). This is followed by the CMA operation, which computes the $\mathcal{S}_{\text{CMA}}$ for each image as the ultimate discriminative metric. Further details are provided in Section 3.

determining membership to the ID. In contrast, zero-shot OOD detection (Ming et al. 2022; Miyai et al. 2023; Fort, Ren, and Lakshminarayanan 2021), in line with the current trend of deep learning, leverages pre-trained models on open datasets, eliminating the need for additional training of the model. The approach determines membership to the ID by calculating results through mapping the data into a common space $\mathbb{R}^d$. Compared to traditional OOD detection frameworks that necessitate training, the zero-shot OOD detection framework is notably more versatile and practical.

CLIP-based OOD Detection The CLIP model aligns image features with text features describing the image in a high-dimensional space by simultaneously training image and text encoders on a large dataset, thereby learning rich visual-language joint representations. When applied to OOD detection tasks, CLIP only requires class names and does not require training on specific ID data, allowing it to attempt to classify or determine whether an input image belongs to a known class. It is worth noting that the ID classes in CLIP-based OOD detection refer to the classes used in downstream classification tasks, which are different from the pre-trained classes in the upstream. The OOD classes are those that do not belong to any of the ID classes in the downstream tasks.

For current CLIP-based OOD detection, MCM (Ming et al. 2022) has become a basic paradigm. Its core idea is to treat text embeddings as "concept prototypes" and evaluate their in-distribution or out-of-distribution properties by measuring the similarity between the input image features and these concept prototypes. Specifically, given a set of ID categories $\mathcal{Y}_{in}$ with a corresponding text description for each category, we first use a pre-trained text encoder $\mathcal{T}$ to convert these text descriptions into d dimensional vectors $\mathbf{c}_i = \mathcal{T}(t_i) \in \mathbb{R}^d$, where $i \in \{1, 2, ..., N\}$ and N is the number of ID categories. For any input image $\mathbf{x}'$ whose visual features are extracted by an image encoder $\mathcal{I}$ as $\mathbf{v}' = \mathcal{I}(\mathbf{x}') \in \mathbb{R}^d$, the MCM score is defined as the cosine similarity between the visual features and the closest concept prototype, which is scaled by an appropriate softmax to enhance the separability of ID and OOD samples:

$$\mathcal{S}_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{in}, \mathcal{T}, \mathcal{I}) = \frac{\exp(\text{sim}(\mathbf{v}', \mathbf{c}_{\hat{y}})/\tau)}{\sum_{i=1}^{N} \exp(\text{sim}(\mathbf{v}', \mathbf{c}_i)/\tau)},$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, $\hat{y} = \arg\max_i \text{sim}(\mathbf{v}', \mathbf{c}_i)$ denotes the most matching concept index, and τ is a temperature parameter used to adjust the distribution of similarity.

3 Method

3.1 Out-of-Distribution Detection

Out-of-Distribution (OOD) detection pertains to the task of discerning whether a given input sample originates from the same distribution as the training data (in-distribution, ID) or from a different distribution (out-of-distribution, OOD). Formally, let X denote the input space, with P_{in} denoting the probability distribution of ID data and P_{out} denoting the

distribution of OOD data. The objective of OOD detection is to classify an input sample $x \in X$ as either belonging to P_{in} or P_{out}. This problem can be framed as a binary classification challenge, where the model output, denoted as $f(x;\theta)$, yields a probability score p indicating the confidence that x is in-distribution: $p(x) = P(y = 1|x;\theta)$, with $y = 1$ signifying that x is in-distribution. A threshold τ is established to facilitate classification, whereby if $p(x) > \lambda$, the sample is classified as in-distribution, and if $p(x) \leq \lambda$, it is classified as out-of-distribution. Thus, OOD detection aims to enhance the robustness and reliability of machine learning models by effectively mitigating the risks posed by unseen or anomalous data.

3.2 How To Construct Vector Triangle Relationships Using Agents

Although the employment of the Agent paradigm in the realm pertaining to LLMs has become a matter of course, the challenge lies in its adaptation to the domain of Out-of-distribution Detection. To address this issue, we have conducted a systematic examination of linguistic visual representations and, based on this, conducted targeted experiments, identifying three primary phenomena.Figure 3 shows two basic examples. These laws provide a new perspective for us to understand and optimize multimodal learning models:

- **The length of prompt will affect the prediction.** In similar text descriptions, the longer the prompt is within a certain range, the higher the score.
  Let L denote the length of the prompt, and $S(L)$ denote the matching score. This relationship could be described as follows:

$$S(L) \propto L \quad \text{for } L \in [a, b],$$

  where a and b denote the lower and upper bounds of the prompt length, respectively.
- **Different words have different weights in the text description.** The more important the word is in describing the overall image, the higher its weight. This means that keywords such as color and shape have a significant impact on the matching score.
  Let w_i denote the weight of the i-th word, and let P denote the set of all words in the textual description. The overall weight of the description can be expressed as follows:

$$W(P) = \sum_{i=1}^{n} w_i \cdot f_i(P)$$

  where $f_i(P)$ is a function associated with the i-th word, encompassing attributes such as word frequency, significance, and so forth, and n denotes the total number of words in the description.
- **Neutral prompts have little impact on images in the ID category.** Conversely, for images in the OOD category, neutral prompts can significantly reduce the score of the ID textual description.
  Let ID denote the set of images in the ID category, OOD denote the set of images in the OOD category, N denote the set of neutral prompts, and $\Delta S(i,p)$ denote the change in score for an image i when a prompt p is applied. The following relationships hold:

$$\forall i \in ID, \forall n \in N : |\Delta S(i,n)| \approx 0$$
$$\forall i \in OOD, \forall n \in N : \Delta S(i,n) \ll 0$$

  where $|\Delta S(i,n)| \approx 0$ indicates that the change in score for images in the ID category remains approximately zero when neutral prompts are applied, and $\Delta S(i,n) \ll 0$ indicates that the change in score for images in the OOD category is significantly negative when neutral prompts are applied.

The causes of these phenomena can be contemplated through the training methods and mechanisms of the CLIP-like models.Given that the common text input for pre-trained models during training is long sentences, they possess higher confidence in inferring long sentences. Additionally, CLIP-liked models undergo contrastive learning training, which enables the model to autonomously focus on the differences between various text descriptions, thereby giving higher weights to key words and skewing the predicted outcomes towards the positive class without being influenced by other words. We conducted a more detailed discussion and statistical verification in the **Appendix**.

Drawing on these significant observations and analyses, we propose the inclusion of neutral prompts unrelated to the ID category as Agents. By engaging Agents in a 'collision' with OOD and ID data, we aim to distance the OOD data from the ID domain.

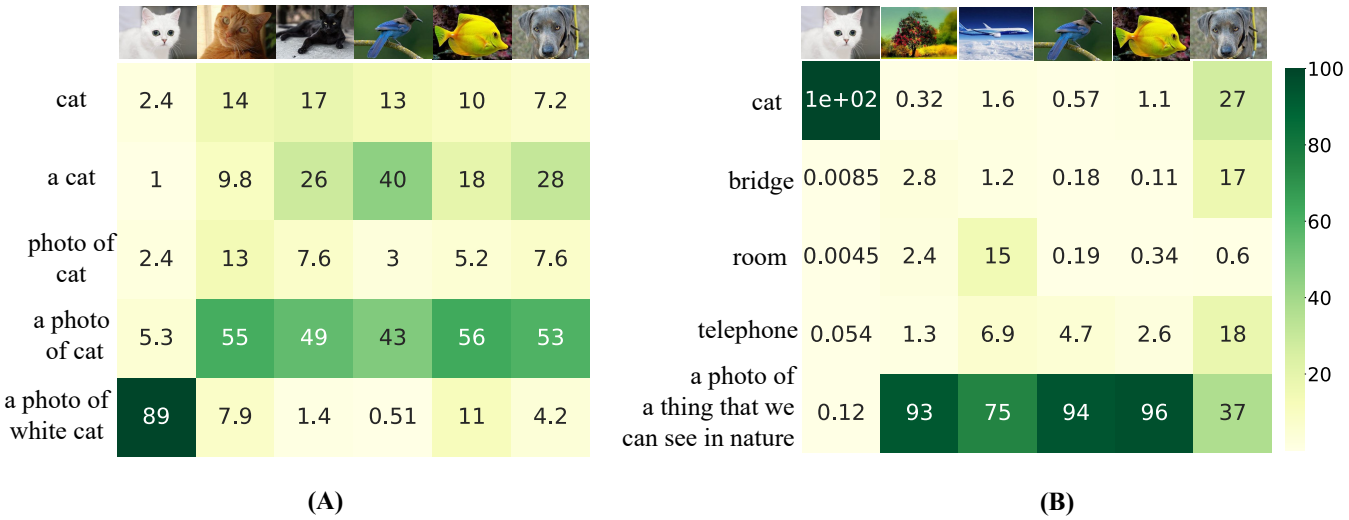
3.3 Proposed Approach

Based on the aforementioned analysis, Concept Matching with Agent (CMA) employs neutral prompts as Agents to widen the gap between ID and OOD by constructing a vector triangle relationship. Essentially, CMA employs external agents to minimize the maximum score, which is fundamentally distinct from traditional learning methods. Therefore, it does not require additional data or training. Specifically, for a set of textual descriptions t_i, where $i \in \{1, 2, ..., N\}$ corresponding to ID categories $\mathcal{Y}_{in}$, we select a certain number of neutral prompts $\mathcal{Y}_{ntc}$ to concatenate with them. The number is generally the same as the number of ID categories N, and the concatenated textual concepts $\mathcal{Y}_{con}$ are used as the input of the text encoder $\mathcal{T}$ to output text features $\mathbf{c}_i = \mathcal{T}(t_i)$, where $i \in \{1, 2, ..., 2N\}$, $t_i \in \mathcal{Y}_{con} = \{\mathcal{Y}_{in}, \mathcal{Y}_{ntc}\}$. Then, it is calculated with the image features output $\mathbf{v}' = \mathcal{I}(\mathbf{x}')$ by the image encoder $\mathcal{I}$.

Formally, we define the **CMA** score as:

$$\mathcal{S}_{\text{CMA}}(\mathbf{x}'; \mathcal{Y}_{in}, \mathcal{Y}_{ntc}, \mathcal{T}, \mathcal{I}) = \frac{\exp(\text{sim}(\mathbf{v}', \mathbf{c}_{\hat{y}})/\tau)}{\sum_{i=1}^{2N} \exp(\text{sim}(\mathbf{v}', \mathbf{c}_i)/\tau)},$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, $\hat{y} = \underset{1 \leq i \leq N}{\arg\max}\, \text{sim}(\mathbf{v}', \mathbf{c}_i)$ denotes the most matching concept



(A) **(B)**

Figure 3: Heatmaps depicting the cosine similarity between image inputs and ID concept vectors. In Figure (A), the ID concept vectors consist of sentences containing the word "cat" of varying lengths. It is observed that images tend to align with longer sentences regardless of whether there is a matching ID concept. Concurrently, keywords such as "white" significantly influence image matching. In Figure (B), aside from "cat", no ID concept can be precisely matched with the given images. However, other than cat images, all images exhibit a preference for aligning with a long sentence devoid of tangible objects. Notably, cat image remains unaffected, aligning solely with the ID concept "cat". All the data in the figure is obtained from the practical use of CLIP (https://github.com/openai/CLIP).

index in ID categories, and τ is a temperature parameter used to adjust the distribution of similarity.

It is particularly noteworthy that although our calculation method is similar to MCM, there are significant differences in detail: (1) When calculating the highest score, the scores corresponding to neutral prompts should be excluded from the calculation, but when performing softmax scaling, the scores corresponding to neutral prompts should be retained; (2) For text descriptions of ID categories, we do not apply prompts for text enhancement, but only use the corresponding category names. This is to further differentiate between ID and neutral prompts, making the model more focused on whether the image truly corresponds to the ID category.

Finally, our OOD detection function can be formally formulated as:

$$g\left(\mathbf{x}';\mathcal{Y}_{\text{in}},\mathcal{Y}_{ntc},\mathcal{T},\mathcal{I}\right)=\begin{cases}1, & S_{\text{CMA}}\geq\lambda\\ 0, & otherwise\end{cases},$$

where 1 indicates ID and 0 indicates OOD. λ is the threshold, and examples below λ are considered OOD inputs.

4 Experiments

4.1 Setup

Datasets We conducted a comprehensive evaluation of the performance of our method across various dimensions and compared it with widely employed OOD detection algorithms. (1) We assessed our approach on the ImageNet-1k OOD benchmark. This benchmark utilizes the large-scale visual dataset ImageNet-1k (Deng et al. 2009) as the ID data and four OOD datasets (including subsets of iNaturalist (Van Horn et al. 2018), SUN (Xiao et al. 2010), Places(Zhou et al. 2017), and Textures (Cimpoi et al. 2014), which are same as Sun *et al.* (Sun et al. 2022)) to fully evaluate the method's performance across various semantic and scenario contexts. (2) We evaluated our method on various small-scale datasets. Specifically, we considered the following ID datasets: FashionMNIST (Xiao, Rasul, and Vollgraf 2017), STL10 (Coates, Ng, and Lee 2011), OxfordIIIPet (Parkhi et al. 2012), Food-101 (Bossard, Guillaumin, and Van Gool 2014), CUB-200 (Wah et al. 2011), PlantVillage (Hughes, Salathé et al. 2015), LFW (Huang et al. 2012), Stanford-dogs (Khosla et al. 2011), FGVC-Aircraft (Maji et al. 2013), Grocery Store (Klasson, Zhang, and Kjellström 2019), and CIFAR-10 (Krizhevsky, Hinton et al. 2009). (3) We assessed our method on hard OOD tasks (Winkens et al. 2020; Ming, Yin, and Li 2022). Following the standards of the MCM (Ming et al. 2022), we evaluated using subsets of ImageNet-1k, namely ImageNet-10 and ImageNet-20, which have similar classes (e.g., dog (ID) vs. wolf (OOD)). During the experiments, we ensured that each OOD dataset did not overlap with the ID dataset in terms of classes.

Model In our experiment, all algorithms uniformly employ CLIP (Radford et al. 2021) as the pre-trained model, which is one of the most prevalent and publicly available visual-linguistic models. Specifically, we utilize CLIP-B/16 as the foundational evaluation model, consisting of a ViT-B/16 transformer (Dosovitskiy et al. 2020) serving as the image encoder and a masked self-attention Transformer (Vaswani et al. 2017) as the text encoder. Additionally, unless otherwise specified, the temperature coefficient is uniformly set to 1 across all algorithms.

Method	iNaturalist		SUN		Places		Texture		**Average**	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Requires training (or w. fine-tuning)										
MSP (Hendrycks and Gimpel 2016)	40.89	88.63	65.81	81.24	67.90	80.14	64.96	78.16	59.89	82.04
Energy (Liu et al. 2020)	21.59	95.99	34.28	93.15	36.64	91.82	51.18	88.09	35.92	92.26
ODIN (Liang, Li, and Srikant 2017)	30.22	94.65	54.04	87.17	55.06	85.54	51.67	87.85	47.75	88.80
ViM (Wang et al. 2022b)	32.19	93.16	54.01	87.19	60.67	83.75	53.94	87.18	50.20	87.82
KNN (Sun et al. 2022)	29.17	94.52	35.62	92.67	39.61	91.02	64.35	85.67	42.19	90.97
NPOS (Tao et al. 2022)	16.58	96.19	43.77	90.44	45.27	89.44	**46.12**	88.80	37.93	91.22
CoOp (Zhou et al. 2022)	43.38	91.26	38.53	91.95	46.68	89.09	50.64	87.83	44.81	90.03
LoCoOp (Miyai et al. 2024)	38.49	92.49	33.27	93.67	39.23	91.07	49.25	**89.13**	40.17	91.53
Zero-shot (no training required)										
MCM (Ming et al. 2022)	30.94	94.61	37.67	92.56	44.76	89.76	57.91	86.10	42.82	90.76
GL-MCM (Miyai et al. 2023)	**15.18**	96.71	30.42	93.09	38.85	89.90	57.93	83.63	35.47	90.83
CMA(Ours)	23.84	**96.89**	**30.11**	**93.69**	**29.86**	**93.17**	47.35	88.47	**32.79**	**93.05**

Table 1: **Comparison results on ImageNet-1k OOD benchmarks.** We use ImageNet-1k as ID dataset. All methods use CLIP-B/16 as a backbone. Bold values represent the highest performance.

ID datasets	iNaturalist		SUN		Places		Texture		**Average**	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
FashionMNIST (Xiao, Rasul, and Vollgraf 2017)	0.00	100.00	0.00	100.00	0.00	100.00	13.61	93.60	3.40	98.40
STL10 (Coates, Ng, and Lee 2011)	0.00	100.00	2.56	99.01	0.00	100.00	0.00	100.00	0.64	99.75
OxfordIIIPet (Parkhi et al. 2012)	0.00	99.89	0.37	99.89	0.96	99.71	0.35	99.85	0.42	99.84
Food101 (Bossard, Guillaumin, and Van Gool 2014)	0.25	99.90	0.49	99.86	1.79	99.40	2.99	99.33	1.38	99.62
CUB-200 (Wah et al. 2011)	0.00	100.00	0.00	100.00	0.00	99.99	0.00	100.00	0.00	100.00
PlantVillage (Hughes, Salathé et al. 2015)	2.95	97.83	0.18	98.11	1.12	98.41	4.37	97.56	2.16	97.98
LFW (Huang et al. 2012)	2.89	99.14	2.24	99.52	8.04	98.25	18.88	95.34	8.01	98.06
Stanford-dogs (Khosla et al. 2011)	0.11	99.92	0.16	99.91	0.58	99.73	0.67	99.72	0.38	99.82
FGVC-Aircraft (Maji et al. 2013)	0.00	99.99	0.67	99.87	1.02	99.69	0.00	99.99	0.42	99.89
Grocery Store (Klasson, Zhang, and Kjellström 2019)	0.03	99.89	0.15	99.97	0.69	99.82	0.79	99.76	0.42	99.86
CIFAR10 (Krizhevsky, Hinton et al. 2009)	0.00	100.00	5.12	98.29	2.56	99.67	0.00	99.88	1.92	99.46
CIFAR100 (Krizhevsky, Hinton et al. 2009)	12.80	97.15	17.92	95.66	15.36	96.19	4.53	98.42	12.65	96.86

Table 2: Zero-shot OOD detection with CMA based on CLIP-B/16 with various ID datasets.

Metrics For evaluation, we use the following metrics: (1) the false positive rate (FPR95) of OOD samples when the true positive rate of in-distribution samples is at 95%, (2) the area under the receiver operating characteristic curve (AUROC). All evaluation outcomes for our method are derived from the average of three experiments, and (3) ID classification accuracy (ID ACC).

4.2 Main Results

OOD detection on Large-scale datasets. The benchmarking of large-scale OOD datasets demonstrates the feasibility of the method for real-world applications and holds significant value. Typically, we employ ImageNet-1k as the ID dataset for Large-scale OOD detection. Table 1 presents the performance of our method in comparison with other approaches under this benchmark. Overall, our method surpasses other methods, achieving superior performance. When juxtaposed against the average performance of Zero-shot methods, CMA demonstrates enhancements of **2.22%** in terms of AUROC and **2.68%** in terms of FPR95. Similarly, when juxtaposed against the average performance of Require training methods, CMA also demonstrates improvements, achieving enhancements of **0.79%** in AUROC and **3.13%** in FPR95, thereby showcasing its exceptional performance.

OOD detection on small-scale datasets. In contrast to benchmarks based on large-scale OOD datasets, small-scale OOD detection often features fewer distinct ID categories. Demonstrating robust performance across these categories is indicative of a method's scalability. Table **??** illustrates the efficacy of our approach across various ID datasets. A notable outcome is that our method achieves impressive performance across these datasets, especially when no specific training was tailored to each dataset.

5 Discussion

The number of Agents. In Section 3, we propose the CMA score formulation $\mathcal{S}_{\text{CMA}}(\mathbf{x}')$ with N agents corresponding to ID categories. Through systematic experiments with $k = M/N$ (where M denotes agent count), we derive the optimized scoring function: $\mathcal{S}'_{\text{CMA}}(\mathbf{x}') = \frac{\exp(\text{sim}(\mathbf{v}', \mathbf{c}_{\hat{y}})/\tau)}{\sum_{i=1}^{N}\exp(\text{sim}(\mathbf{v}', \mathbf{c}_i)/\tau) + \sum_{j=1}^{M}\exp(\text{sim}(\mathbf{v}', \mathbf{c}_{j+N})/\tau)}$ where $\hat{y} = \arg\max_{1\le i\le N} \text{sim}(\mathbf{v}', \mathbf{c}_i)$. As shown in Figure 4, empirical analysis on ImageNet-1k reveals optimal performance at $k = 1$. Notably, insufficient agents (N too small) disrupt the vector triangle relationship, causing structural instability that degrades OOD detection efficacy. Beyond a critical threshold, additional agents yield diminishing returns, confirming the framework's structural saturation effect.

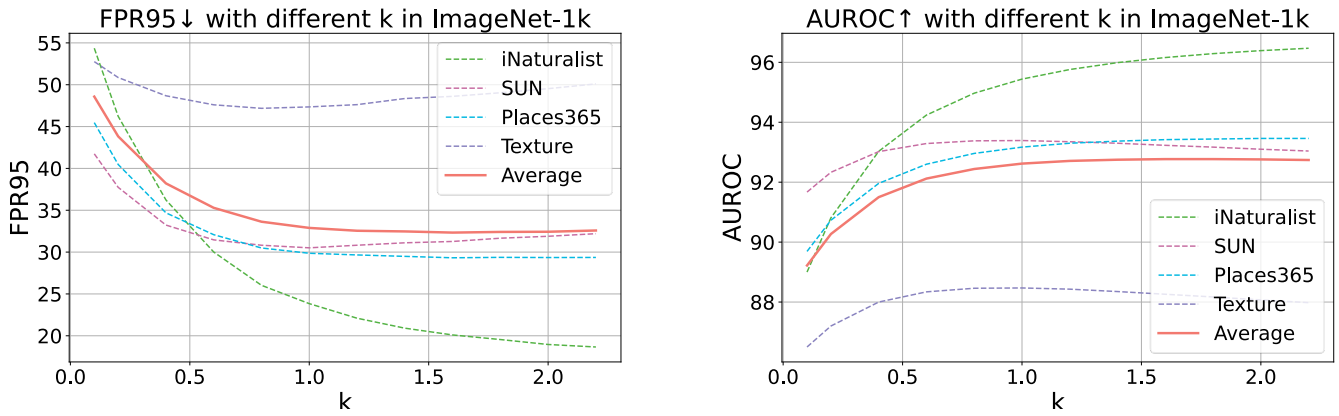


Figure 4: **Impact curve of $k = \frac{\text{number of Agents}}{\text{number of ID Labels}}$ on the performance of CMA.** *Left* shows that as k gradually increases, the FPR95 on various datasets generally decreases, with the fastest decline occurring in the range of k less than 0.5, followed by a gradual slowdown, which is more evident on the average curve. *Right* shows that as k gradually increases, the AUROC gradually increases, also with a rapid rise followed by a gradual slowdown.

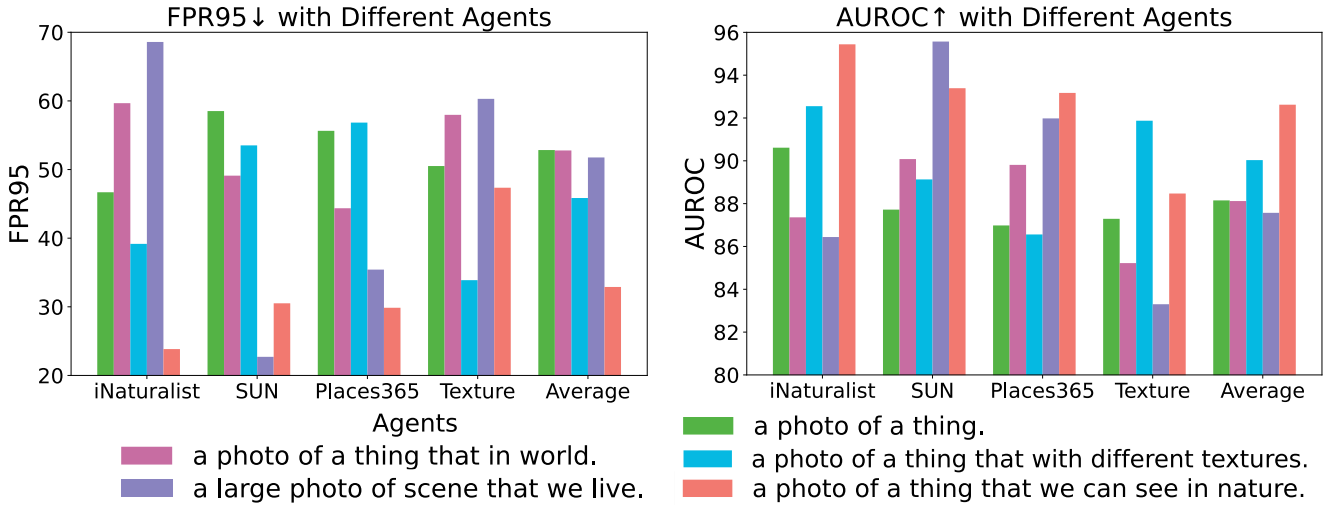


Figure 5: **Comparison with different Agents.** *Left* shows the performance of different agents on various datasets in terms of FPR95, while *Right* shows the performance in terms of AUROC. Clearly, the agent that performs best on average across all datasets does not perform best on every dataset. Moreover, even agents that perform poorly on average can still show decent performance on certain datasets.

Performance of different Agents. Figure 5 demonstrates significant performance variance (average **19.95%** FPR95 and **4.50%** AUROC gaps) across agent implementations. Context-specific specialization emerges notably in environmental descriptors like "a large photo of the scene we live in", which achieves peak performance on SUN dataset (Xiao et al. 2010) despite suboptimal cross-domain generalization. This duality reveals both CMA's sensitivity to agent selection and its adaptive potential through environment-specific agent optimization.

6 Conclusion

This paper proposes a novel zero-shot OOD detection framework, Concept Matching with Agent (CMA). By introducing the concept of Agents into the OOD detection task, a vector triangular relationship consisting of ID labels, data inputs, and Agents is constructed, offering a fresh perspective on OOD detection. Beginning with a language-vision representation, we demonstrate the impact of neutral words on CLIP-liked models, thereby proposing the innovative idea of treating neutral words as Agents. Building on this, we propose a novel score computation method based on CMA. By incorporating the relationships between Agents and data inputs, this method enables unique interactions between ID data and OOD data with Agents, thereby facilitating a better separation between ID and OOD. We investigate the effectiveness of CMA across various scenarios, including large-scale datasets, small-scale datasets, and hard OOD detection, achieving outstanding performance across a wide range of tasks. Finally, we delve deeper into CMA, highlighting its flexibility and scalability. We hope that our work will inspire future exploration of new paradigms for OOD detection.

Acknowledgements

This work was supported by National Natural Science Foundation of China, Grant Number: 62476109, 62206108, and the Natural Science Foundation of Jilin Province, Grant Number: 20240101373JC, and Jilin Province Budgetary Capital Construction Fund Plan, Grant Number: 2024C008-5, and Research Project of Jilin Provincial Education Department, Grant Number: JJKH20241285KJ.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461. Springer.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3): 1–45.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 215–223. JMLR Workshop and Conference Proceedings.

Crowson, K.; Biderman, S.; Kornis, D.; Stander, D.; Hallahan, E.; Castricato, L.; and Raff, E. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, 88–105. Springer.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

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

Fort, S.; Ren, J.; and Lakshminarayanan, B. 2021. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34: 7068–7081.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Hendrycks, D.; and Gimpel, K. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; et al. 2023. MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations*.

Huang, G. B.; Mattar, M.; Lee, H.; and Learned-Miller, E. 2012. Learning to Align from Scratch. In *NIPS*.

Hughes, D.; Salathé, M.; et al. 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Fei-Fei, L. 2011. Novel Dataset for Fine-Grained Image Categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO.

Klasson, M.; Zhang, C.; and Kjellström, H. 2019. A hierarchical grocery store image dataset with visual and semantic labels. In *2019 IEEE winter conference on applications of computer vision (WACV)*, 491–500. IEEE.

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

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

Liu, J.; Shen, Z.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

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

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

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

Ming, Y.; Yin, H.; and Li, Y. 2022. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 10051–10059.

Miyai, A.; Yu, Q.; Irie, G.; and Aizawa, K. 2023. Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521*.

Miyai, A.; Yu, Q.; Irie, G.; and Aizawa, K. 2024. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Qian, C.; Cong, X.; Yang, C.; Chen, W.; Su, Y.; Xu, J.; Liu, Z.; and Sun, M. 2023a. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Qian, C.; Dang, Y.; Li, J.; Liu, W.; Chen, W.; Yang, C.; Liu, Z.; and Sun, M. 2023b. Experiential co-learning of software-developing agents. *arXiv preprint arXiv:2312.17025*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

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

Tao, L.; Du, X.; Zhu, J.; and Li, Y. 2022. Non-parametric Outlier Synthesis. In *The Eleventh International Conference on Learning Representations*.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022a. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3835–3844.

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

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 1–26.

Wang, M.; Xing, J.; and Liu, Y. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Winkens, J.; Bunel, R.; Roy, A. G.; Stanforth, R.; Natarajan, V.; Ledsam, J. R.; MacWilliams, P.; Kohli, P.; Karthikesalingam, A.; Kohl, S.; et al. 2020. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

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

Zeng, A.; Liu, M.; Lu, R.; Wang, B.; Liu, X.; Dong, Y.; and Tang, J. 2023. Agenttuning: Enabling generalized agent abilities for llms. *arXiv preprint arXiv:2310.12823*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.