Machine Vision Therapy: Multimodal Large Language Models Can Enhance Visual Robustness via Denoising In-Context Learning

Zhuo Huang [*1] **Chang Liu** [2] **Yinpeng Dong** [3] **Hang Su** [3] **Shibao Zheng** [2] **Tongliang Liu** [1]

Abstract

Although pre-trained models such as Contrastive Language-Image Pre-Training (CLIP) show impressive generalization results, their robustness is still limited under Out-of-Distribution (OOD) scenarios. Instead of undesirably leveraging human annotation as commonly done, it is possible to leverage the visual understanding power of Multi-modal Large Language Models (MLLMs). However, MLLMs struggle with vision problems due to task incompatibility, thus hindering their effectiveness. In this paper, we propose to effectively leverage MLLMs via Machine Vision Therapy which aims to rectify erroneous predictions of specific vision models. By supervising vision models using MLLM predictions, visual robustness can be boosted in a nearly unsupervised manner. Moreover, we propose a Denoising In-Context Learning (DICL) strategy to solve the incompatibility issue. Concretely, by examining the noise probability of each example through a transition matrix, we construct an instruction containing a correct exemplar and a probable erroneous one, which enables MLLMs to detect and rectify the incorrect predictions of vision models. Under mild assumptions, we theoretically show that our DICL method is guaranteed to find the ground truth. Through extensive experiments on various OOD datasets, our method demonstrates powerful capabilities for enhancing visual robustness under many OOD scenarios.

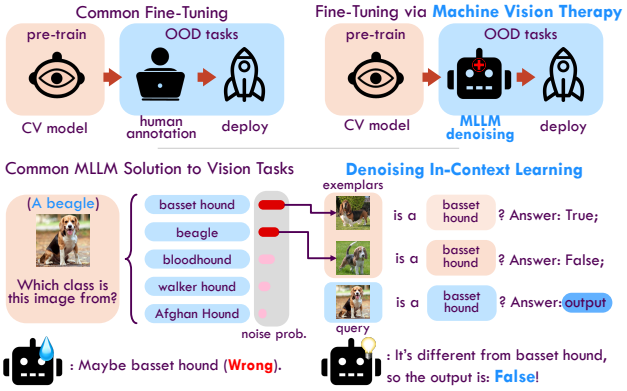

Figure 1. Illustration of our methodology: Upper row: Comparison between common fine-tuning process and fine-tuning via Machine Vision Therapy. Our method potentially eliminates the necessity for human-annotation by leveraging the knowledge from MLLMs. Lower row: Comparison between previous MLLM solution to vision tasks and Denoising In-Context Learning strategy. Instead of considering all classes, our method make predictions by presenting a pair of positive and negative exemplars.

1. Introduction

Pre-trained vision models such as Vision Transformers (ViT) (Dosovitskiy et al., 2020; Liu et al., 2021; Wang et al., 2021) with Contrastive Language-Image Pretraining (CLIP) (Chen et al., 2023a; Radford et al., 2021; Li et al., 2021; 2022; Wang et al., 2024b; Zheng et al., 2023a) have been widely used thanks to their strong generalization performance meanwhile effectively avoiding training vision models from scratch. But when deployed to Out-of-Distribution (OOD) scenarios (Dong et al., 2023; Kong et al., 2023; Hendrycks & Gimpel, 2016; Hong et al., 2024; Peng et al., 2023; Wang et al., 2019; Zhu et al., 2023b), their recognition performance could be seriously degraded (Shu et al., 2023). Downstream fine-tuning has been a common practice to regain the generalizability (Goyal et al., 2023; Wortsman et al., 2022), but it requires additional label acquisition through human labor, which is undesirable for large-scale applications.

Fortunately, the thriving Multi-modal Large Language Models (MLLMs) (Alayrac et al., 2022; Awadalla et al., 2023; Chen et al., 2024; Gong et al., 2023; Li et al., 2023c;b; Liu et al., 2023; Ye et al., 2023; Zhu et al., 2023a), which take advantage of the few-shot learning ability of

[*]Work done when visiting Tsinghua University, code is available at **this link**. [1]Sydney AI Centre, The University of Sydney, Sydney, Australia [2]Institute of Image Communication and Network Engineering, Shanghai JiaoTong University, Shanghai, China [3]Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua-Bosch Joint ML Center, THBI Lab, BNRist Center, Tsinghua University, Beijing, 100084, China. Correspondence to: Tongliang Liu <tongliang.liu@sydney.edu.au>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Large Language Models (LLM) (Brown et al., 2020; Chung et al., 2022; Floridi & Chiriatti, 2020; OpenAI, 2023; Scao et al., 2022; Touvron et al., 2023a;b; Zheng et al., 2023b), have manifested powerful capabilities on understanding visual information with language interpretations, and excelled at recognizing novel objects in multimodal tasks such as image captioning, visual question answering, visual reasoning, etc. Considering the vulnerability of vision models under OOD situations, here we hope to refine vision models by leveraging the knowledge of MLLMs, as shown in the upper row of Figure 1. However, due to the difficulty of aligning the text generation process with visual recognition tasks[1] (Alayrac et al., 2022; Wang et al., 2022), MLLMs struggle with generating correct answers that match the ground-truth class names, thus underperforming the current dominant contrastive paradigms, even when employing them as own vision encoders (Alayrac et al., 2022; Awadalla et al., 2023; Huang et al., 2023a; Wang et al., 2022; Zhai et al., 2023).

Focusing on enhancing the robustness of vision models, in this paper, we propose to effectively leverage MLLMs to conduct **Machine Vision Therapy (MVT)** which aims to diagnose and rectify the error predictions through a novel Denoising In-Context Learning (DICL) strategy. Then, we utilize the rectified supervision to guide the fine-tuning process in downstream OOD problems. Specifically, rather than giving a set of options to ask MLLMs for the exact answer (Alayrac et al., 2022; Huang et al., 2023a; Zhai et al., 2023), we show that it is sufficient to query for the ground truth by using only two exemplars, *i.e.*, 1) a correct one that demonstrates the exact match between a query class name with its image example and 2) an erroneous one that combines the same query class with an image from the most confusing category for the vision model. Since the erroneous predictions are essentially label noise, hence we draw inspiration from learning with noisy labels (Han et al., 2018; Liu & Tao, 2015; Lin et al., 2022; 2023b; Natarajan et al., 2013; Wu et al., 2024; 2023; Xia et al., 2020b;a; Yao et al., 2020; 2021; 2023; Yuan et al., 2024). Particularly, we can find the erroneous categories by estimating a transition matrix that captures the probability of one class being mistaken as another. By feeding the two exemplars, MLLMs can be instructed to leverage their few-shot learning power to distinguish the semantically similar images that are easily misclassified by vision models, as shown in the lower row of Figure 1. To process such instructions, we leverage the multi-modal in-context learning ability of several existing MLLMs (Chen et al., 2023b; Li et al., 2023a; Yasunaga et al., 2023; Zhao et al., 2023) to realize our methodology. After the error predictions are diagnosed and rectified, vision models can be further fine-tuned to enhance their OOD robustness on downstream data distribution. Through a comprehensive empirical study on many challenging datasets and their OOD variants, such as ImageNet (Deng et al., 2009), WILDS (Koh et al., 2021b), and DomainBed (Gulrajani & Lopez-Paz, 2021), we carefully validate the effectiveness of our method and demonstrate its superiority under various OOD scenarios on many well-known vision models.

To sum up, our contributions are three-fold:

- We design a novel Machine Vision Therapy paradigm to enhance computer vision models by effectively leveraging the knowledge of MLLMs without needing additional label information.
- We propose a Denoising In-Context Learning strategy to successfully align MLLMs with vision tasks.
- Through comprehensive quantitative and qualitative studies on many well-known datasets, we demonstrate that the proposed method can enhance: 1) generalization on both ID and OOD data, 2) robustness against domain shift, 3) robustness against common corruptions, 4) performance on recognizing fine-Grained attributes, 5) robustness against spurious correlations, 6) detection on prediction errors and OOD data.

2. Methodology

In this section, we carefully demonstrate the Machine Vision Therapy process which mainly contains three components, namely Transition Matrix Estimation, Denoising In-Context Learning, and Fine-Tuning of vision models. Next, we demonstrate problem setting and framework overview.

2.1. Problem Formulation and Overview

Generalizing to Out-of-Distribution tasks has been a challenging topic in computer vision problems, where we normally have a vision model parameterized by $\theta_{cv} \in \Theta_{cv}$ pre-trained on massive labeled in-distribution (ID) data $\mathcal{D}^{id} = \{x_i^{id}, y_i^{id}\}_{i=0}^{m} \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \mathbb{R}^C$. Here each ID example is sampled from a joint distribution, *i.e.*, $(X^{id}, Y^{id}) \sim p^{id}$, where X^{id} and Y^{id} stand for variables. After pretraining, we can assume the conditional distribution $P(Y^{id}|X^{id})$ can be perfectly captured by the inference function $\tilde{y}^{id} = f_{\theta_{cv}}(x^{id})$, where $\tilde{y}^{id}$ is the prediction. In OOD tasks, we are given a set of unlabeled examples $\mathcal{D}^{ood} = \{x_i^{ood}\}_{i=0}^{n}$ whose element $x^{ood} \in \mathcal{X}$ is drawn from an unknown data distribution p^{ood}. Due to the change of downstream task, some factors that affect the data generating process are shifted, causing a difference between p^{ood} and p^{id}, further hindering the label prediction, *i.e.*, $\tilde{y}^{ood} = f_{\theta_{cv}}(x^{ood}) \not\sim P(Y^{ood}|X^{ood})$, where Y^{ood} is the unknown ground truth. Fortunately, having been observed with extraordinary low-shot generalization capability, we leverage MLLM with parameters $\theta_{mllm} \in \Theta_{mllm}$ to enhance the OOD robustness of vision models.

Our framework is illustrated in Figure 2 and our problem can be formulated as follows:

[1] In this paper, we mainly focus on classification task.

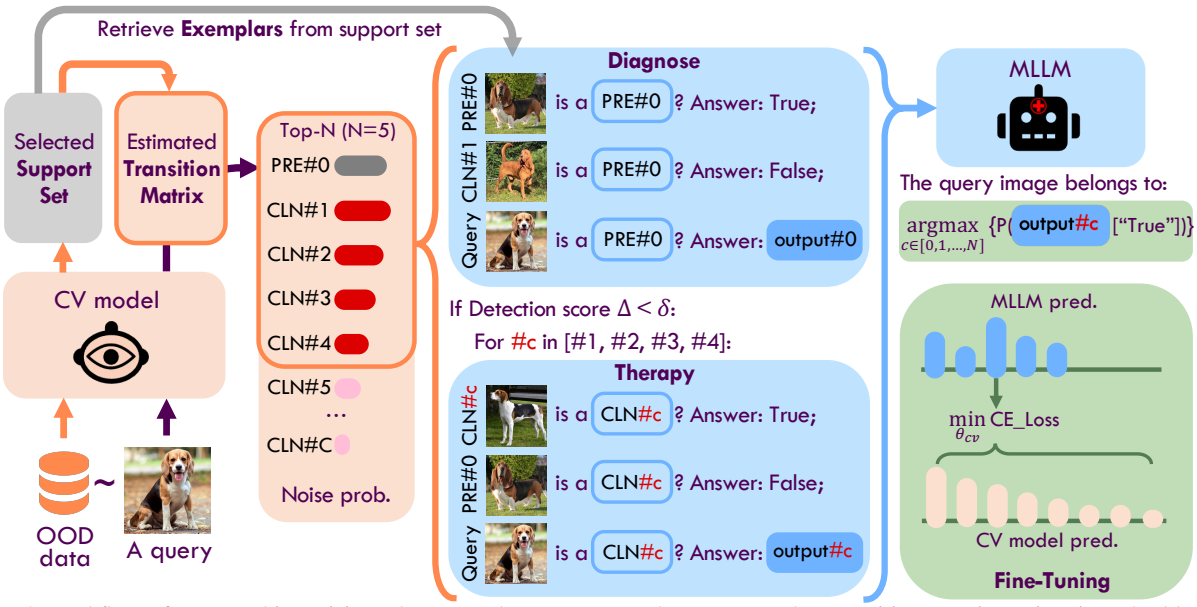

Figure 2. Workflow of our Machine Vision Therapy: The orange part demonstrates the Transition Matrix Estimation, the blue part indicates the Denoising In-Context Learning process, and the green part illustrates the Fine-Tuning of vision models.

$$\min_{\theta_{cv}} \mathcal{L}(f_{\theta_{cv}}, z);\ z = \left[\theta_{mllm}((X_c^+, Y_c^+); (X_c^-, Y_c^+); X_i)\right]_c^N;$$
$$Y_c = T[c; \text{argmax}[f_{\theta_{cv}}(X_i)]], \tag{1}$$

where X_i^+ and X_i^- denotes the positive and negative exemplars, respectively, X_i is the query image, and T is the transition matrix. Intuitively, when a distribution shift occurs, the emerging prediction errors are essentially label noises that can be captured by estimating a transition matrix. Hence, by focusing on calibrating the examples with high noise probabilities, the visual robustness of downstream tasks can be improved effectively. In particular, we feed all OOD data into the vision model to obtain the noisy prediction distribution $P(\tilde{Y}^{ood}|X^{ood})$, based on which we can effectively estimate T and provide exemplars to instruct MLLM[2]. Further, we conduct machine vision therapy to find the possible ground truth for X_i based on the MLLM output z. Finally, z is leveraged to minimize $\mathcal{L}$ to optimize θ_{cv}. Next, we explain the details of each process.

2.2. Transition Matrix Estimation

The distribution shift from OOD data x^{ood} leads to unreliable label prediction $\tilde{y}^{ood}$, which is highly unreliable due to instance-dependent feature noises (Li et al., 2024; Xia et al., 2020b) as shown in Section 3.3. Hence, in order to capture the relationship between $\tilde{Y}^{ood}$ and Y^{ood}, we leverage a transition matrix $T \in [0,1]^{C\times C}$ (Liu & Tao, 2015; Natarajan et al., 2013; Xia et al., 2019) which satisfies $P(Y^{ood}|X^{ood}) = T^\top P(\tilde{Y}^{ood}|X^{ood})$. However, estimating such a transition matrix is difficult without access to any noisy label supervision or strong assumption (Liu & Tao, 2015; Xia et al., 2019). Therefore, we propose a simple yet effective sample selection approach to construct a support set with clean labels. Specifically, we rank all OOD data within each class based on their prediction confidence, *i.e.*, $\max_c \left[f_{\theta_{cv}}(x^{ood})\right]_c$, where $[\cdot]_c$ denotes the value of the c-th entry. From the sorted dataset $\{x_1^{ood,c}, x_2^{ood,c}, \cdots, x_{\frac{n}{C}}^{ood,c}\}_{c=1}^C$, we uniformly sample ρ examples per class, where ρ is the labeling budget, *i.e.*, $\mathcal{D}^{supp} = \{\{x_{j\times\frac{n}{\rho C}}^{ood,c}\}_{j=1}^{\rho}\}_{c=1}^C$. In this way, we can effectively model the noisy posterior $P(\tilde{Y}^{ood}|X^{ood})$. Then, through an acceptable labeling process[3], we can obtain the clean label posterior $P(Y^{ood}|X^{ood})$, thus effectively estimating the transition matrix T. Finally, the noise transition probability $T\,[:; \arg\max[f_{\theta_{cv}}]]$ of a query image can be obtained by indexing T through its current prediction.

2.3. Denoising In-Context Learning

Thanks to the previously obtained noise probability list $T\,[:; \arg\max[f_{\theta_{cv}}]]$, we can further decide which one is the possible ground truth through DICL. In particular, we only consider the classes of the top-N noise probability as potential candidates. If the label prediction denoted by "PRE#0" is not in the candidates, we would fix it in the first

[2]Although some manual annotation is required, we show in later experiments that our strategy has an acceptable labeling workload and demonstrates superior performance to vanilla fine-tuning on the support set. Furthermore, the support set is **not used for parameter tuning** in our method, so our fine-tuning does not actually use any human annotation for training.

[3]We experimentally show that when there is a distribution shift between $\mathcal{D}^{supp}$ and $\mathcal{D}^{ood}$, the proposed method can still perform effectively. As a result, it is unnecessary to conduct the labeling process on each practical task. Instead, we can just use the existing support set to instruct most of OOD tasks.

place. Further, we conduct *Diagnosing* which decides the fidelity of the current prediction, and *Therapy* which finds the possible ground truth.

Diagnosing. Since the inference time of MLLMs is non-trivial, it is necessary to avoid redundant analysis on confident examples. Hence, to examine the fidelity of vision model predictions, our Diagnosing focuses on answering whether a query image belonging to class "PRE#0" is "True". Specifically, we retrieve from $\mathcal{D}^{supp}$ to obtain one exemplar image belonging to "PRE#0", and another exemplar image belonging to the class with the largest noise transition probability "CLN#1"[4]. Then, combined with the query image X_q, an in-context instruction is constructed:

> Question: This image <IMG_PRE#0> shows a photo of <PRE#0>, True or False? Answer: True;
>
> Question: This image <IMG_CLN#1> shows a photo of <PRE#0>, True or False? Answer: False;
>
> Question: This image <IMG_Query> shows a photo of <PRE#0>, True or False? Answer:

The symbols <IMG_PRE#0>, <IMG_CLN#1>, and <IMG_Query> are replace tokens for the image features of exemplars from "PRE#0" and "CLN#1", and X_q, respectively. The first exemplar acts as the positive one to show MLLMs the true image from class "PRE#0", and the second exemplar shows the negative one to show the highly probable false image from "CLN#1". Then, based on the X_q and "PRE#0", MLLMs can effectively judge the correctness by outputting z_0:

$$z_0 = \theta_{mllm}((X_{PRE\#0}, Y_{PRE\#0}); (X_{CLN\#1}, Y_{PRE\#0}); X_q). \tag{2}$$

To enable further quantitative analysis, we obtain the logits of "True" and "False" tokens from the MLLM output z_0 followed by a softmax function, *i.e.*, $z_0 := \text{softmax}([z_0[\text{True}], z_0[\text{False}]])$. Finally, we combine $z_0[\text{True}]$ and the prediction confidence of the vision model to obtain a detection score Δ:

$$\Delta = \frac{1}{2}(z_0[\text{True}] + \max_c [f_{\theta_{cv}}]_c (x^{ood})). \tag{3}$$

If Δ is larger than a threshold δ, we assume the current prediction "PRE#0" is correct[5], otherwise, we conduct the next Therapy process.

Therapy. During therapy, we continue to use the instruction template above and traverse across the rest clean class candidates. Particularly, for each iteration c in $N-1$ trials, we choose "CLN#c" as the positive class and "PRE#0" as the negative class, whose exemplars are correspondingly retrieved from $\mathcal{D}^{supp}$ to construct the prompt. Then, it is fed into MLLM to output whether the query image belongs to the class "CLN#c", *i.e.*, $z_c = \theta_{mllm}((X_{CLN\#c}, Y_{CLN\#c}); (X_{PRE\#0}, Y_{CLN\#c}); X_q)$, let $z_c := \text{softmax}([z_c[\text{True}], z_c[\text{False}]])$. As a result, we can decide the final prediction through:

$$y_{mllm} = \arg\max [z_c[\text{True}]]_{c=0}^{N}. \tag{4}$$

As shown in Section 3, the performance of MLLM prediction shows strong performance in many OOD scenarios. However, we still cannot directly employ MLLMs for inference, due to three main reasons: 1) Non-negligible inference time: Since current MLLMs cannot handle large-batch data, it would be unimaginably slower (*e.g.*, 1000×) when using MLLMs rather than vision models; 2) High requirements for computation: Inference through MLLM takes up huge memory of GPU. For MLLMs using large LLMs such as LLaMA-13B, it requires distributed inference on less advanced devices; 3) Model privacy issue: Many MLLMs are highly sensitive with limited accessibility, therefore. Hence, we propose to fine-tune vision models based on the prediction of MLLMs.

Algorithm 1 Machine Vision Therapy.

Input: Pre-trained vision model θ_{cv}, MLLM θ_{mllm}, OOD dataset $\mathcal{D}^{ood}$.

Uniformly sample ρC examples from confidence-sorted $\mathcal{D}^{ood}$ to construct support set $\mathcal{D}^{supp}$;
Estimate transition matrix T; {*Section 2.2*}
for $i \in 0, 1, \cdots, n$ **do**
  Based on the label prediction $\tilde{y}_i^{ood}$ to obtain the noisy transition probability $T[:, \arg\max[f_{\theta_{cv}}]]$;
  Conduct Diagnosing through Eq. 2 and compute detection score Δ through Eq. 3;
  if $\Delta > \delta$ **then**
    Accept current prediction;
  else
    Conduct Therapy and obtain MLLM prediction through Eq. 4; {*Section 2.3*}
    Based on the MLLM prediction, conduct fine-tuning through Eq. 5. {*Section 2.4*}
  end if
end for

2.4. Fine-Tuning of Vision Models

After obtaining the MLLM prediction y_{mllm}, we propose to optimize vision models through the following objective:

$$min_{\theta_{cv}} \mathcal{L}_{ce}(f_{\theta_{cv}}, y_{mllm}), \tag{5}$$

where $\mathcal{L}_{ce}(\cdot)$ denotes the cross-entropy loss. Here we summarize our methodology in Algorithm 1. Further, we can directly deploy the fine-tuned vision models to OOD tasks whose effectiveness is demonstrated in Section 3.

2.5. Theoretical Analysis

We denote the MLLM is pretrained over a distribution p defined by a latent concept $\phi \in \Phi$. During DICL, there are n examples to form a prompt S_n which are sampled from a prompt distribution p_{prompt} defined by concept $\phi^* \in \Phi$.

[4]The performance of retrieve strategy is carefully studied in Section 3.4.

[5]Detailed analysis is shown in Section 3.4.

To justify the proposed DICL strategy, based on the theoretical framework proposed by Xie et al. (2021), we show that when MLLM achieving the most probable z based on the given prompt S_n and query image-text pair x_q-y under a concept ϕ^*, the corresponding y is the same as the one found from p_{prompt}, which is y_q that matches with x_q.

Assumption 2.1 (Distribution consistency)**.** $\forall(x_q, y_q) \sim p_{prompt}, p(x_q, y_q) = p_{prompt}(x_q, y_q)$.

Moreover, the assumptions from Xie et al. (2021) also hold, then we have the following Theorems:

Theorem 2.2. *Assume that the above assumptions hold, if for all $\phi \in \Phi$, $\phi \neq \phi^*$, the concept ϕ^* satisfies the distinguishability condition: $\sum_{j=1}^{k} KL_j(\phi^* \| \phi) > \epsilon_{start}^{\phi} + \epsilon_{delim}^{\phi}$, then as $n \to \infty$, the prediction according to the pretraining distribution is*

$$\arg\max_y p(y|S_n, x_q, \phi^*) \to \arg\max_y p_{prompt}(y|x_q). \quad (6)$$

Thus, the in-context predictor f_n achieves the optimal $0-1$ risk: $\lim_{n\to\infty} \mathcal{L}_{0-1}(f_n) = \inf_f \mathcal{L}_{0-1}(f)$.

Lemma 2.3. *Under the same condiction of Theorem 2.2, the prediction z according to the pretraining distribution is*

$$\operatorname{argmax}_z p(z|S_n, x_q, y_q, \phi^*) \to \operatorname{argmax}_z p_{prompt}(z|x_q, y_q). \quad (7)$$

Theorem 2.4. *Assume that the above assumptions hold, as $n \to \infty$, when achieving the largest prediction probability of z given prompt under concept ϕ^*, the corresponding class description y follows the same y obtained from the prompt distribution:*

$$\operatorname{argmax}_y p(z|S_n, x_q, y, \phi^*) \to \operatorname{argmax}_y p_{prompt}(z|x_q, y). \quad (8)$$

Please see the **appendix** for proof. We can see that if n is large enough, the MLLM prediction z achieves the largest value when y_q is the exact match to x_q. As a result, we can justify that only when we feed the positive image-text pair to the MLLM, the prediction z is the largest among all other combinations between x_q and $y \in \mathcal{Y}, y \neq y_q$.

3. Experiments

In this section, we first provide our experimental details. Then we conduct quantitative comparisons with the state-of-the-art vision models. Finally, we conduct ablation studies and analyses to qualitatively validate our method.

3.1. Experimental Setup

Datasets. In our experiments, we use well-known ID datasets including ImageNet-1K (Deng et al., 2009) validation dataset, ImageNet-V2 (Recht et al., 2019), CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009) and MNIST (LeCun et al., 1998). We also evaluate OOD generalization on datasets that are commonly considered OOD ones, ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), mageNet-Sketch (Wang et al., 2019), ImageNet-V (Dong et al., 2022), iWildCam (Koh et al., 2021a), and DomainBed (Gulrajani & Lopez-Paz, 2020).

Models and baselines. For vision backbone, we employ CLIP models (Radford et al., 2021) and utilize ViT-L/14 and ViT-g (Zhai et al., 2022) from EVA (Fang et al., 2022) as the vision model to be enhanced. For the MLLM backbone, we consider two existing works MMICL (Zhao et al., 2023) and Otter (Li et al., 2023b) that possess multimodal ICL ability. We also assess the performance of CLIP variants on ResNet50, ResNet101, and ViT-B/32 as alternative vision encoders in the appendix. Additionally, we conduct Visual Question Answering (VQA) to directly ask MLLMs the class of query images. Moreover, we conduct vanilla fine-tuning (Vanilla FT) using only D^{supp} as a baseline. The performance of using MLLM prediction is denoted as MVT, and our fine-tuning result is denoted as FT.

Settings. For model evaluation, we randomly select 5000 images independently from the ImageNet validation set (IN-Val), ImageNet-V2 (IN-V2), ImageNet-A (IN-A), ImageNet-R (IN-R), ImageNet-Sketch (IN-SK), ImageNet-V (IN-V) and 10000 images independently from CIFAR10, CIFAR100, MNIST to constitute the test samples. Additionally, we select 3 images per category to construct a support set to provide in-context exemplars. In the main paper, we evaluate iWildCam from WILDS and VLCS, PACS, OfficeHome, and DomainNet from DomainBed. For the details of implementation, we choose the top-6 noisy classes to conduct MVT. Concretely, we set the threshold $\delta = 0.6$ to diagnose incorrect predictions, then we retrieve exemplars from the support set based on the most similar logit prediction to query images. For each round of DICL, we repeat the process for 3 times and average the model predictions. During fine-tuning, we optimize the vision models for 3 epochs using Adam and SGD optimizers for ViT-L and ViT-g, respectively. Other details and datasets are shown in **Appendix**.

3.2. Quantitative Comparison

First, we compare our MVT method with well-known vision models under both ID and OOD scenarios. As shown in Table 1, we can see that our method with fine-tuning denoted as "+FT" achieves better performance in most settings. Specifically, on "IN-V", our method with fine-tuning can significantly surpass both CLIP and EVA for 17% and 6%, respectively. Moreover, on "IN-A", our method achieves 4.3% and 2.8% performance improvement over the second-best method on both ViT-L and ViT-g backbone, respectively. We can also observe that even without fine-tuning, the prediction accuracy of MLLM denoted by "MVT" can still surpass all baselines on most scenarios, which denotes the strong performance enhancement of our MVT fine-tuning on vision models. Note that we did not provide fine-tuning on iWildCam because most of the predictions are incorrect. Though MVT can still achieve the best result, the vision encoders could be misled by erroneous decisions during the fine-tuning process.

Table 1. Classification accuracy (%) of baseline CLIP models and our method on 5 ID datasets and 5 OOD datasets. The baseline methods includes ViT-L from CLIP (Radford et al., 2021) and ViT-g from EVA (Fang et al., 2022), VQA, and Vanilla FT.

Arch	Method	ID					OOD				
		IN-Val	IN-V2	CIFAR10	CIFAR100	MNIST	IN-A	IN-R	IN-SK	IN-V	iWildCam
RN50		59.7	52.6	71.5	41.9	58.5	23.9	60.7	35.4	31.1	8.2
RN101	CLIP	61.7	56.2	80.8	48.8	51.6	30.2	66.7	40.9	35.4	12.3
ViT-B		62.9	56.1	89.9	65.0	47.9	32.2	67.9	41.9	30.5	10.9
ViT-L	CLIP	75.8	70.2	95.6	78.2	76.4	69.3	86.6	59.4	51.8	13.4
	VQA	64.9	59.9	97.6	**83.2**	56.7	66.0	87.3	56.9	56.2	13.3
	Vanilla FT	76.1	**70.8**	96.1	80.3	77.5	70.8	87.5	60.0	53.6	15.2
	MVT	75.2	**70.8**	**97.9**	78.9	53.0	71.2	88.1	59.0	62.1	**25.0**
	+FT	**76.9**	70.5	96.7	82.0	**79.2**	**75.1**	**89.5**	**61.4**	**68.8**	-
ViT-g	EVA	78.8	71.2	98.3	88.8	62.2	71.9	91.4	67.7	64.9	21.9
	VQA	64.3	59.6	97.9	84.5	55.7	64.6	87.4	58.2	59.2	19.7
	Vanilla FT	78.9	71.8	98.7	89.1	62.9	72.7	91.6	68.1	65.6	22.4
	MVT	**79.1**	71.6	98.1	89.0	63.2	73.2	91.4	67.9	66.3	**25.1**
	+FT	79.0	**72.2**	**98.9**	**91.2**	**65.7**	**75.5**	**92.8**	**68.6**	**70.6**	-

Table 2. Classification accuracy (%) of baseline CLIP models and our method on 4 subsets of DomainBed datasets. The baseline methods includes ViT-L from CLIP (Radford et al., 2021) and ViT-g from EVA (Fang et al., 2022), VQA, and Vanilla FT.

	Datasets	VLCS				PACS				OfficeHome				DomainNet					Avg
	method	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	4	
ViT-L	CLIP	74.9	83.5	80.3	74.5	97.8	97.4	97.5	99.4	87.7	92.7	85.7	85.6	61.1	62.1	60.2	78.4	51.1	80.6
	Vanilla FT	78.8	85.2	83.4	77.0	**98.0**	97.6	97.7	99.6	87.9	93.1	87.1	86.9	62.0	62.5	60.5	78.5	51.9	81.6
	MVT	83.8	89.0	87.2	80.3	97.6	97.5	**98.0**	99.4	87.7	93.4	89.0	88.5	61.3	62.1	60.4	78.7	53.4	82.8
	+FT	**84.2**	**89.8**	**87.9**	**82.5**	**98.0**	**98.2**	**98.0**	**99.8**	**90.9**	**95.0**	**90.9**	**90.8**	**62.5**	**63.8**	**62.4**	**80.1**	**54.0**	**84.0**
ViT-g	EVA	72.5	80.0	79.8	72.8	99.0	98.8	98.9	99.8	90.5	94.2	88.6	88.7	61.4	64.7	61.2	81.6	54.9	81.6
	Vanilla FT	75.5	82.3	82.1	75.6	98.9	98.7	98.9	99.8	90.6	94.5	89.2	89.0	61.5	64.9	61.3	81.8	54.8	82.3
	MVT	81.2	86.6	86.1	79.5	98.2	98.0	98.0	99.4	89.7	93.8	89.7	89.1	**62.2**	**65.0**	61.6	**82.3**	56.1	83.3
	+FT	**83.7**	**89.5**	**86.9**	**82.0**	**99.1**	**98.9**	**99.0**	**100.0**	**91.6**	**95.1**	**90.7**	**90.6**	61.9	64.8	**63.2**	81.9	**56.6**	**84.4**

Furthermore, we consider domain shift by leveraging DomainBed datasets. Specifically, for each dataset, we leave one domain out as a test dataset and fine-tune on rest domains. By comparing two state-of-the-art vision backbones ViT-L and ViT-g, we show the performance comparison in Table 2. As we can see, both MVT and MVT with fine-tuning can significantly surpass the baseline methods. For some scenarios such as the PACS dataset, our method can achieve nearly 100% performance. Moreover, in several scenarios in the VLCS dataset, both our MVT and fine-tuning can achieve almost 10% improvements. Additionally, we find that our method with fine-tuning largely surpasses vanilla fine-tuning baseline on both Tables 1 and 2. Hence, we can conclude that our learning strategy can indeed provide effective supervisions which enhances vision robustness under distribution shift.

3.3. Ablation Study

In this part, we conduct ablation studies to analyze each module of MVT by using ViT-L backbone vision model.

Ablation Study on Transition Matrix Estimation. To validate the performance of transition matrix estimation, we compare our confidence-based uniform sampling strategy to a random sampling baseline. The result on the ImageNet-V dataset is shown in Figure 3. To quantitatively show the superiority of our method, we compute the ℓ_2 norm of the difference between one estimation and ground truth which indicates the fidelity of the estimation. As a result, our estimation is much more accurate by achieving 3.83 norm, compared to 4.46 of random sampling.

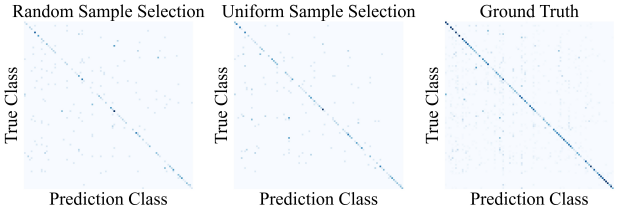

Figure 3. Ablation study on transition matrix estimation by comparing our method with random sampling and ground truth.

Ablation Study on Choosing Noisy Classes. Further, we justify the choice of using a transition matrix to obtain the noisy classes. As a comparison, we use the top-6 predictions as the therapy candidates and show the results in Table 3. We can see on all datasets, our method can outperform the opponent with non-trivial improvements. Therefore, leveraging the transition matrix to find the potential noisy classes is more effective than using prediction.

Table 3. Performance comparison between choosing noisy classes through transition matrix (MVT) and using Top-N predictions.

	IN-A	IN-SK	IN-Val	IN-R	IN-V2	IN-V
Top-N Pred.	60.3	58.4	74.2	85.3	67.7	58.3
MVT	65.5	59.0	75.1	86.0	70.7	61.6

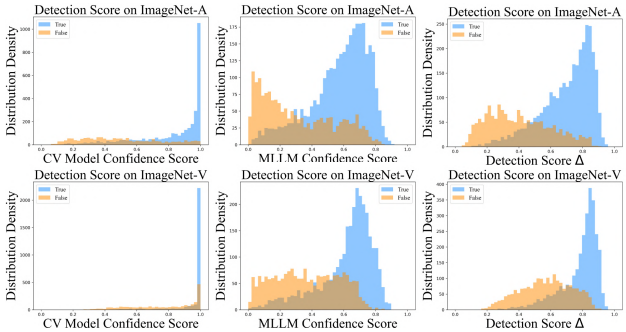

Figure 4. Ablation study on detection score distribution. Upper: ImageNet-A; Lower: ImageNet-V.

Table 4. Comparison of classification accuracy (%) on 5 OOD datasets with Otter (Li et al., 2023b) and MMICL (Zhao et al., 2023). We compare the performance on CLIP ViT-L (Radford et al., 2021) backbone.

MLLM	Method	IN-A	IN-R	IN-SK	IN-V	iWildCam
None	CLIP	69.3	86.6	59.4	51.8	13.4
Otter	MVT	64.1	85.2	59.5	51.9	**16.2**
	+FT	**73.5**	**88.7**	**60.0**	**55.7**	-
MMICL	MVT	71.2	88.1	59.0	62.1	**25.0**
	+FT	**75.1**	**89.5**	**61.4**	**68.8**	-

Ablation Study on Detection Score. To analyze the proposed detection score on conducting diagnosing, we show the distribution of prediction confidence provided by the vision model, MLLM, and our detection score Δ in Figure 4. Based on the results, we can justify our design of Δ: In the left column, we can see the confidence of correctly classified examples is very high, but the wrong ones show uniform distribution. Conversely, in the middle column, although MLLM poses slightly lower scores on correct ones, it significantly suppresses the confidence of wrong ones. As a result, we combine two scores to obtain Δ, which can produce clearly separable distributions to benefit the diagnosing process. Unless specified, we set the threshold $\delta = 0.6$ which works effectively in most scenarios.

Ablation Study on MLLM Backbone. To testify the effectiveness of MVT on different MLLM backbones, here we instantiate our method using Otter (Li et al., 2023a) and compare it to the previous realization on MIMIC (Zhao et al., 2023). The result is shown in Table 4. We can see that both the implementation on MMICL and Otter show superior performance to the employed vision encoder backbone. Although the performance slightly differs between Otter and MMICL, which could be due to the model capacity and their training strategy, we can generally conclude that our MVT method is applicable to different MLLMs backbones with ICL and could further benefit from more sophisticated MLLMs in the future.

3.4. Performance Analysis

Further, we conduct qualitative analysis to thoroughly validate the effectiveness of our MVT.

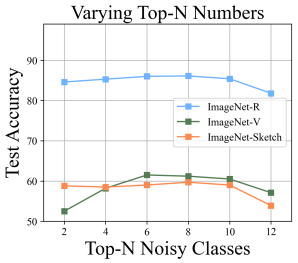

Figure 5. Performance analysis by varying the number of top-N chosen noisy classes.

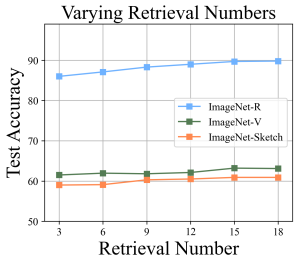

Figure 6. Performance analysis by varying the number of retrieved exemplars.

Choice of Top-N Noisy Classes. To study how a varied number of chosen noisy classes could affect the performance of our method, we change the top-N number from 2 to 12, and show the result on ImageNet-R, ImageNet-V, and ImageNet-Sketch datasets in Figure 5. We find a common phenomenon that either too small or too large a number of N could hurt the performance. This could be because that small N would ignore too many potential ground-truth classes. In contrast, large N includes too many choices that could interfere with the final prediction. Setting N to 6 could be an ideal choice for ImageNet-based datasets.

Effect of Retrieval Numbers. In our experiments, we retrieve exemplars for 3 times and average the predictions. To further investigate the effect of varied retrieval numbers, we change the number of retrievals from 3 to 18 and conduct experiments on the same OOD datasets as above. Specifically, we consider one positive and negative pair for a single DICL round as one retrieval. We repeat this process for R times and ensemble the MLLM predictions through $\frac{1}{R}\sum_r^R [z_c[\text{True}]^r, z_c[\text{False}]^r]$. In this way, it is possible that MLLM predictions would be more accurate. The result is shown in Figure 6. We observe that the performance steadily improves as the retrieval number increases, however, the performance gains vanish when the retrieval number becomes too large. Moreover, large retrieval numbers would multiply the computation cost. Therefore, it is suggested to set the number to a reasonably small value.

Performance of Different Retrieval Strategy. As shown by Alayrac et al. (Alayrac et al., 2022), Retrieval-based In-Context Example Selection (RICES) can significantly affect the ICL performance. Therefore, here we investigate its influence. Specifically, we propose two retrieval strategies, namely feature-based retrieval and logit-based retrieval. The former one is based on feature similarity and the latter one is based on the prediction logit. For each strategy, we conduct experiments on selecting the most similar examples and the least similar examples, which are denoted as "most" and "least", respectively. The

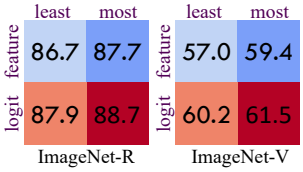

Figure 7. Performance analysis on different retrieval strategies.

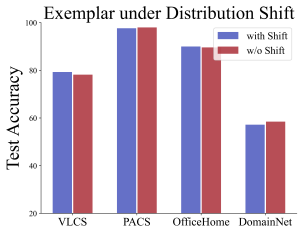

Figure 8. Performance analysis on in-context exemplars with distribution shift.

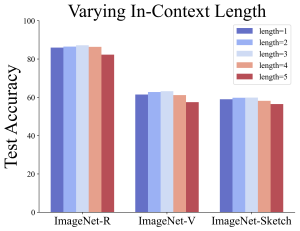

Figure 9. Performance analysis by varying the length of in-context exemplars.

results are shown in Figure 7. Apart from the intuitive finding that least-similar retrieval is inferior to selecting the most-similar one, we also observe that logit-based retrieval is more effective than feature-based one. We assume this is due to the image classification task is more related to logit value rather than feature similarity.

Effect of In-Context Exemplars with Distribution Shift. When the support set suffers from a distribution shift from the target OOD dataset, whether DICL can still perform robustly remains to be validated. Hence, we leave one domain out as our support set and leverage the rest domains as our target OOD dataset. In comparison, we choose a small hold-out data split as the support set which shares the same distribution as the OOD dataset. The results are shown in Figure 8. Surprisingly, we find that the performance is not influenced by the distribution shift, which demonstrates that our MVT can still be effective when exemplars are retrieved from different distributions.

Effect of Varying In-Context Length. Further, we analyze the effect of increasing exemplar length during inference. Particularly, we consider one positive and negative exemplar pair as length 1. Here we vary the length from 1 to 5 and show the results in Figure 9. We observe slight improvement when the length gradually increases which is consistent with the theoretical findings (Xie et al., 2021). However, when the length is longer than 4 the performance drops and the predictions of MLLM become unstable which could be other than "True" or "False". This might be due to the limited capacity of MLLMs on handling a certain amount of information, which is worth conducting studies on sophisticated MLLMs in the future.

Performance of OOD Detection. At last, we consider a more challenging scenario where data from open classes could exist in the target dataset. Here we simulate this situation by choosing 60% of the classes as closed classes and the rest are open classes. To detect such open-class data, *i.e.*, OOD detection (Hendrycks & Gimpel, 2016; Wang et al., 2024a)[6], we use the vision model prediction confidence as a baseline and compare it with the MLLM diagnosing confidence as well as the detection score Δ in Eq. 3. The result is shown in Figure 10. In the upper row, we observe the similar clearly distinguishable distributions using our score Δ as in Figure 4. In the lower row, we show the F1 score of each detection criterion under a threshold varied from 0 to 1 on three datasets. When a criterion produces confidence larger than the threshold, it would predict as close-class data, other as open-class ones. Based on the result, we find that MLLM achieves better detection performance when the threshold is small, but vision model confidence is relatively better when the threshold is large, *i.e.*, MLLM can effectively detect open classes while vision models are better at recognizing close classes. However, an effective detection should have a reasonable threshold value that is neither too large nor too small and meanwhile has a high F1 score. Hence, by combining them together, our detection score Δ can achieve the best F1 score when the threshold is around the middle range.

[6]Note that OOD detection here is different from the previous setting: here we focus on detecting open-class data, and previous one focuses on detecting prediction errors.

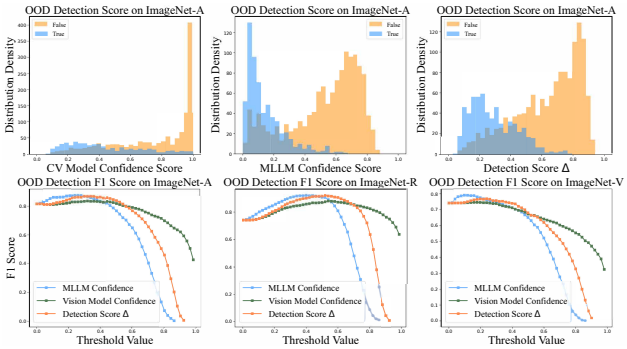

Figure 10. OOD detection analysis. Upper: Detection score distribution on ImageNet-A; Lower: F1 scores of vision model confidence, MLLM diagnosing confidence, and our Δ score in ImageNet-A, ImageNet-R, and ImageNet-V.

4. Conclusion

In this paper, we propose a novel paradigm of fine-tuning vision models via leveraging MLLMs to improve visual robustness on downstream OOD tasks. Specifically, we effectively estimate a transition matrix to help find the most probable noisy classes. By using a positive exemplar and a negative exemplar retrieved based on the noisy classes, we can conduct DICL to rectify incorrect vision model predictions through two stages dubbed diagnosing and therapy. Thanks to the rectified predictions, the robustness of vision models can be further improved through fine-tuning. We conduct detailed theoretical analysis and extensive quantitative and qualitative experiments to justify the proposed method. Our framework can significantly reduce the cost of training vision models and provide insights into many visual recognition problems such as OOD detection, OOD generalization, weakly-supervised learning, *etc*.

Acknowledgements

Yinpeng Dong is supported by NSFC project (No. 62276149). Hang Su is partially supported by NSFC projects (Nos. 92248303, 92370124, 62350080). Shibao Zheng is supported by NSFC projects (Nos. 62071292, U21B2013) and STCSM project (No. 22DZ2229005). Tongliang Liu is partially supported by the following Australian Research Council projects: FT220100318, DP220102121, LP220100527, LP220200949, and IC190100031.

Impact Statement

As the rapid development of MLLMs and LLMs has started influencing our lives, the proposed work could also show some potential broader impacts. For example, when highly powerful and accurate MLLMs occur in the future, human supervision for learning models will be no longer necessary. The learning paradigm could be large models guiding small models. However, it is vital that researchers make sure the knowledge interaction between large models and small models is ethically responsible, trustworthy, secure, robust, and explainable. Harmful knowledge or uncertain knowledge should not be freely transferred from one model to another. Hence, a rigorous code of ethics should be formulated to effectively regulate the process.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, pp. 23716–23736, 2022.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

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

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. Invariant rationalization. In *International Conference on Machine Learning*, pp. 1448–1458. PMLR, 2020.

Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y., and Wang, W. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7020–7030, 2023a.

Chen, R., Liu, Y., Kong, L., Chen, N., Zhu, X., Ma, Y., Liu, T., and Wang, W. Towards label-free scene understanding by vision foundation models. *Advances in Neural Information Processing Systems*, 36, 2024.

Chen, Y., Zhang, S., Han, B., and Jia, J. Lightweight in-context tuning for multimodal unified models. *arXiv preprint arXiv:2310.05109*, 2023b.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

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

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Dong, Y., Ruan, S., Su, H., Kang, C., Wei, X., and Zhu, J. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. *Advances in Neural Information Processing Systems*, 35:36789–36803, 2022.

Dong, Y., Kang, C., Zhang, J., Zhu, Z., Wang, Y., Yang, X., Su, H., Wei, X., and Zhu, J. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1022–1032, 2023.

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

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

Floridi, L. and Chiriatti, M. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., and Chen, K. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

Goyal, S., Kumar, A., Garg, S., Kolter, Z., and Raghunathan, A. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, pp. 19338–19347, 2023.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *ICLR*, 2021.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, volume 31, 2018.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2016.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *CVPR*, pp. 15262–15271, 2021b.

Hong, Z., Wang, Z., Shen, L., Yao, Y., Huang, Z., Chen, S., Yang, C., Gong, M., and Liu, T. Improving nontransferable representation learning by harnessing content and style. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FYKVPOHCpE.

Hou, W., Chen, S., Chen, S., Hong, Z., Wang, Y., Feng, X., Khan, S., Khan, F. S., and You, X. Visual-augmented dynamic semantic prototype for generative zero-shot learning. *arXiv preprint arXiv:2404.14808*, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Liu, Q., et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023a.

Huang, Z., Li, M., Shen, L., Yu, J., Gong, C., Han, B., and Liu, T. Winning prize comes from losing tickets: Improve invariant learning by exploring variant parameters for out-of-distribution generalization. *arXiv preprint arXiv:2310.16391*, 2023b.

Huang, Z., Xia, X., Shen, L., Han, B., Gong, M., Gong, C., and Liu, T. Harnessing out-of-distribution examples via augmenting content and style. In *The Eleventh International Conference on Learning Representations*, 2023c. URL https://openreview.net/forum?id=boNyg20-JDm.

Huang, Z., Zhu, M., Xia, X., Shen, L., Yu, J., Gong, C., Han, B., Du, B., and Liu, T. Robust generalization against photon-limited corruptions via worst-case sharpness minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16175–16185, 2023d.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021a.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, pp. 5637–5664. PMLR, 2021b.

Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., and Liu, Z. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19994–20006, 2023.

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., and Liu, Z. Mimic-it: Multi-modal in-context instruction tuning. 2023a.

Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., and Liu, Z. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023b.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023c.

Li, M., Wu, R., Liu, H., Yu, J., Yang, X., Han, B., and Liu, T. Instant: Semi-supervised learning with instance-dependent thresholds. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

Lin, R., Yu, C., Han, B., and Liu, T. On the over-memorization during natural, robust and catastrophic overfitting. In *The Twelfth International Conference on Learning Representations*, 2023a.

Lin, R., Yu, C., and Liu, T. Eliminating catastrophic overfitting via abnormal adversarial examples regularization. *Advances in Neural Information Processing Systems*, 36, 2024.

Lin, Y., Yao, Y., Du, Y., Yu, J., Han, B., Gong, M., and Liu, T. Do we need to penalize variance of losses for learning with label noise? *arXiv preprint arXiv:2201.12739*, 2022.

Lin, Y., Yao, Y., Shi, X., Gong, M., Shen, X., Xu, D., and Liu, T. Cs-isolate: Extracting hard confident examples by content and style isolation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. In *NeurIPS*, volume 33, pp. 21464–21475, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021.

Lynch, A., Dovonon, G. J., Kaddour, J., and Silva, R. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching. In *International Conference on Machine Learning*, pp. 7313–7324. PMLR, 2021.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *NeurIPS*, volume 26, 2013.

OpenAI. Gpt-4 technical report, 2023.

Peng, X., Chen, R., Qiao, F., Kong, L., Liu, Y., Wang, T., Zhu, X., and Ma, Y. Sam-guided unsupervised domain adaptation for 3d segmentation. *arXiv preprint arXiv:2310.08820*, 2023.

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

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Shu, Y., Guo, X., Wu, J., Wang, X., Wang, J., and Long, M. Clipood: Generalizing clip to out-of-distributions. In *ICML*. PMLR, 2023.

Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., and Cai, D. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, volume 32, 2019.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

Wang, Q., Fang, Z., Zhang, Y., Liu, F., Li, Y., and Han, B. Learning to augment distributions for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 36, 2024a.

Wang, Q., Lin, Y., Chen, Y., Schmidt, L., Han, B., and Zhang, T. Do clips always generalize better than imagenet models? *arXiv preprint arXiv:2403.11497*, 2024b.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pp. 568–578, 2021.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In *CVPR*, pp. 7959–7971, 2022.

Wu, Y., Xia, X., Yu, J., Han, B., Niu, G., Sugiyama, M., and Liu, T. Making binary classification from multiple unlabeled datasets almost free of supervision. *arXiv preprint arXiv:2306.07036*, 2023.

Wu, Y., Yao, J., Xia, X., Yu, J., Wang, R., Han, B., and Liu, T. Mitigating label noise on graph via topological sample selection. *arXiv preprint arXiv:2403.01942*, 2024.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, volume 32, 2019.

Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., and Chang, Y. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020a.

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. Part-dependent label noise: Towards instance-dependent label noise. In *NeurIPS*, volume 33, pp. 7597–7610, 2020b.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., and Sugiyama, M. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33:7260–7271, 2020.

Yao, Y., Liu, T., Gong, M., Han, B., Niu, G., and Zhang, K. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34:4409–4420, 2021.

Yao, Y., Gong, M., Du, Y., Yu, J., Han, B., Zhang, K., and Liu, T. Which is better for learning with noisy labels: the semi-supervised method or modeling label noise? In *International Conference on Machine Learning*, pp. 39660–39673. PMLR, 2023.

Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., and Yih, W.-t. Retrieval-augmented multimodal language modeling. In *ICML*, 2023.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yuan, S., Feng, L., and Liu, T. Early stopping against label noise without validation data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CMzF2aOfqp.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., and Ma, Y. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., and Chang, B. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023.

Zheng, J., Yao, Y., Han, B., Wang, D., and Liu, T. Enhancing contrastive learning for ordinal regression via ordinal content preserved data augmentation. In *The Twelfth International Conference on Learning Representations*, 2023a.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023b.

Zhou, D., Liu, T., Han, B., Wang, N., Peng, C., and Gao, X. Towards defending against adversarial examples via attack-invariant features. In *International conference on machine learning*, pp. 12835–12845. PMLR, 2021.

Zhou, D., Wang, N., Gao, X., Han, B., Wang, X., Zhan, Y., and Liu, T. Improving adversarial robustness via mutual information estimation. In *International Conference on Machine Learning*, pp. 27338–27352. PMLR, 2022.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023a.

Zhu, Z., Zhang, Y., Chen, H., Dong, Y., Zhao, S., Ding, W., Zhong, J., and Zheng, S. Understanding the robustness of 3d object detection with bird's-eye-view representations in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21600–21610, 2023b.

Supplementary Material for "Machine Vision Therapy: Multimodal Large Language Models Can Enhance Visual Robustness via Denoising In-Context Learning"

In this supplementary material, we provide extensive quantitative and qualitative studies on a wide range of datasets to thoroughly understand the essence of the proposed framework. First, we discuss some related works on OOD generalization and Multimodal Large Language Models in Section A. Second, we theoretically analyze and justify the proposed DICL strategy in Section B. Then, we elucidate the additional details of our experimental setting and implementation in Section C. Further, we provide additional quantitative comparisons on different MLLM and CV backbone models, and different OOD types in Section D. Finally, we perform additional analysis to further explore the effectiveness of our method in Section F. Finally, we discuss the limitation of this work and the broader impact in Section G

A. Related Work

In this section, we provide a brief discussion of the OOD generalization problem and multimodal large-language models.

A.1. OOD Generalization

OOD data refers to those with different distributions from training data. OOD generalization aims at improving the performance of deep models to unseen test environments. Researchers attempted to tackle the problem from different perspectives, such as data augmentation, OOD detection, invariant causal mechanisms (Huang et al., 2023c;b; Zhou et al., 2021; 2022), and so on. Data augmentation is effective in improving model generalization. Typical methods involve Cutout (DeVries & Taylor, 2017), which randomly occludes parts of an input image; CutMix (Yun et al., 2019), which replaces a part of the target image with a different image; Mixup (Zhang et al., 2017), which produces a convex combination of two images; DeepAugment (Hendrycks et al., 2021a), which passes a clean image through an image-to-image network and introduces several perturbations during the forward pass. Some methods conduct OOD detection to separate OOD data. Typical methods include softmax confidence score (Hendrycks & Gimpel, 2016; Huang et al., 2023d), which is a baseline for OOD detection; Outlier Exposure (OE) (Hendrycks et al., 2018), which uses unlabeled data as auxiliary OOD training data. Energy scores are shown to be better for distinguishing OOD samples from IID ones (Liu et al., 2020). Some work resort to causality to study the OOD generalization problem. Typical methods include MatchDG (Mahajan et al., 2021), which proposes matching-based algorithms when base objects are observed and approximate the objective when objects are not observed; INVRAT (Chang et al., 2020), which leveraged some conditional independence relationships induced by the common causal mechanism assumption.

A.2. Multimodal Large Language Models

The field of vision-language models has witnessed significant advancements in recent years, driven by the growing synergy between computer vision and natural language processing. Notably, this synergy has led to the exceptional zero-shot performance (Hou et al., 2024) of CLIP (Radford et al., 2021), a model that employs a two-tower contrastive pretraining approach to align image and text information. In the rapidly evolving landscape of LLMs, exemplified by GPTs (Brown et al., 2020), LLaMA (Touvron et al., 2023a), and Vicuna (Chiang et al., 2023), it has become evident that LLMs possess the capacity to process information from diverse domains. BLIP-2 (Li et al., 2023c), for instance, serves as a foundational model, aligning visual features and text features using a Querying Transformer (Q-former) and utilizing OPT (Zhang et al., 2022) and FLAN (Chung et al., 2022) as language models. Building upon BLIP-2, Instruct-BLIP (Dai et al., 2023) has enhanced instruction-following capabilities. To further bolster the instruction-following proficiency of multi-modal models, LLaVA (Liu et al., 2023) and Mini-GPT4 (Zhu et al., 2023a) have introduced meticulously constructed instruction sets, which have found widespread application in various multi-modal models. mPLUG-Owl (Ye et al., 2023) introduces a two-stage learning paradigm, first fine-tuning the visual encoder and then refining the language model with LoRA (Hu et al., 2021). This approach effectively fuses image and text features. Some models consider additional modalities, such as ImageBind (Girdhar et al., 2023), which simultaneously incorporates data from six modalities without the need for explicit supervision, and PandaGPT (Su et al., 2023) which enhances its instruction-following capabilities. Several multi-modal

models prioritize the in-context learning abilities of LLMs. Flamingo (Alayrac et al., 2022), in one of the pioneering efforts, integrates a gated cross-attention module to align with the spaces of images and text. Otter (Li et al., 2023b) refines OpenFlamingo (Awadalla et al., 2023), an open-source version of Flamingo, by introducing instruction-tuning datasets to improve instruction-following abilities. Multi-Modal In-Context Learning (MMICL) (Zhao et al., 2023) is a comprehensive vision-language model that incorporates Instruct-BLIP, enabling the analysis and comprehension of multiple images, as well as the execution of instructions. MLLMs possess the remarkable capacity to capture intricate details and engage in reasoning when presented with an image. Nevertheless, it remains uncertain about how to enhance visual perception by harnessing the knowledge embedded within LLMs.

B. Theoretical Framework

Pretraining distribution formulation. We based on the in-context learning framework proposed by Xie et al. (2021). In this framework, a latent concept ϕ from a concept space Φ defines a pretraining distribution p over prompt tokens o observed from a vocabulary space O. To generate the desired content, we first sample a concept from a prior $p(\phi)$ and then sample the tokens conditioned on the concept. We denote the total length of the pretraining examples is T:

$$p(o_1, \ldots, o_T) = \int_{\phi \in \Phi} p(o_1, \ldots, o_T|\phi)p(\phi)d\phi. \tag{9}$$

The conditional probability $p(o_1, \ldots, o_T|\phi)$ is defined by a Hidden Markov Model (HMM). Based on the concept ϕ, a transition matrix of the HMM hidden states $h_1, \ldots, h_T$ from a hidden state space $\mathcal{H}$ can be found.

Prompt distribution formulation. During the in-context learning process, we sample a prompt from a new distribution p_{prompt}, which contains n independent exemplars and 1 query example, which are all conditioned on a shared prompt concept ϕ^*. The goal is to predict the next token based on the exemplars and the query example. Specifically, the i-th exemplar O_i consists of a tokenized image feature $x_i = O_i[1 : k_x]$, a text description to claim the class of the image $y_i = O_i[k_x : k_x + k_y]$, and a binary prediction to judge whether the claim of the image is "True" or "False", which is denoted by $z_i = O_i[k_x + k_y : k_x + k_y + 1]$ at the end of each exemplar. The generating process of the i-th exemplar is as follows:

1. First generate a start hidden state h_i^{start} from prompt start distribution p_{prompt};
2. Given h_i^{start}, generate the exemplar sequence $O_i = [x_i, y_i, z_i]$ from $p(O_i|h_i^{start}, \phi^*)$, the generate exemplars are conditioned on a given prompt concept ϕ^*.

The query example Q is sampled similarly without the binary prediction z_q, *i.e.* $Q = [x_q, y_q]$. Between each exemplar and the query example, there is a special delimiter token o^{delim} denoting the end of each exemplar sequence. The prompt can be formulated as follows:

$$[S_n, Q] = [x_1, y_1, z_1, o^{delim}, x_2, y_2, z_2, o^{delim}, \ldots, x_n, y_n, z_n, o^{delim}, x_q, y_q] \sim p_{prompt}, \tag{10}$$

where S_n denotes the n independent exemplars for in-context demonstration.

Denoising In-context learning task. In our denoising in-context learning, the output prediction z for each image and text pair $[x, y]$ is sampled based on the prompt distribution $p_{prompt}(z|x, y)$:

$$z_q \sim p_{prompt}(z|x, y) = \mathbb{E}_{h_q^{start} \sim p_{prompt}(h_q^{start}|Q)}[p(z|Q, h_q^{start}, \phi^*)], \tag{11}$$

where h_q^{start} denotes the hidden state corresponding to the first token of Q, *i.e.,* the first token of x_q.

Our goal is to analyze the in-context predictor $f_n(x_q, y_q) = \arg\max_z p(z|S_n, x_q, y_q)$ which outputs the most likely prediction over the pretraining distribution p conditioned on the exemplars S_n sampled from the prompt distribution p_{prompt}. We assume the in-context predictor is trained by $0-1$ error with n training examples $\mathcal{L}_{0-1}(f_n) = \mathbb{E}_{x_q, y_q} \sim p_{prompt}[\mathbf{1}[f_n(x_q) \neq y_q]]$ and $\mathcal{L}_{0-1}(f_n) = \mathbb{E}_{x_q, y_q, z_q} \sim p_{prompt}[\mathbf{1}[f_n(x_q, y_q) \neq z_q]]$..

One major difference of our denoising in-context learning strategy is that we not only use positive exemplars that show exact image-text match, *i.e.,* $(x, y) \sim p(x, y) = p_{prompt}(x, y)$, we also have negative exemplars where image and text are

not corresponding to each other. To construct such a prompt, we have to first select the ideal y, based on the matching result of x and y, we can further obtain the prediction z. Therefore, in the following theoretical proof, we propose to conduct two-step analyses on y and z, respectively.

B.1. Assumptions

Our theoretical framework is built upon Xie et al. (2021), whose assumptions are also applied to our analysis.

Assumption B.1 (Delimiter hidden states)**.** Let the delimiter hidden states $\mathcal{D}$ be a subset of $\mathcal{H}$. For any $h^{delim} \in \mathcal{D}$ and $\phi \in \Phi$, $p(o^{delim}|h^{delim}, \phi) = 1$ and for any $h \notin \mathcal{D}$, $p(o^{delim}|h, \phi) = 0$.

Assumption B.2 (Bound on delimiter transitions)**.** For any delimiter state $h^{delim} \in \mathcal{D}$ and any hidden state $h \in \mathcal{H}$, the probability of transitioning to a delimiter hidden state under ϕ is upper bounded $p(h^{delim}|h, \phi) < c_2$ for any $\phi \in \Phi \setminus \{\phi^*\}$, and is lower bounded $p(h^{delim}|h, \phi^*) > c_1 > 0$ for ϕ^*. Additionally, the start hidden state distribution for delimiter hidden states is bounded as $p(h^{delim}|\phi) \in [c_3, c_4]$.

Assumption B.3 (Distribution shift from prompt start distribution)**.** We assume that the prompt start distribution p_{prompt} is close in TV distance to all hidden transition distributions (under ϕ^*) starting from a delimiter hidden state: $\max_{h^{delim} \in \mathcal{D}} TV(p_{prompt}(h) \| p(h|h^{delim}, \phi^*)) < \tau/4$. Here, $\tau = p_{prompt}(y_{max}|Q) - \max_{y \neq y_{max}} p_{prompt}(y|Q)$ is the margin between the most likely label $y_{max} = \arg\max_y p_{prompt}(y|Q)$ and the second most likely label.

Assumption B.4 (Well-specification)**.** The prompt concept ϕ^* is in Φ.

Assumption B.5 (Regularity)**.** The pretraining distribution p satisfies: 1) Lower bound on transition probability for the prompt concept ϕ^*: for any pair of hidden states $h, h' \in \mathcal{H}$, $p(h|h', \phi^*) > c_5 > 0$. 2) Start hidden state is lower bounded: for any $h \in \mathcal{H}$, $p(h|\phi^*) \geq c_8 > 0$. 3) All tokens can be emitted: for every symbol o, there is some hidden state $h \in \mathcal{H}$ such that $p(o|h, \phi^*) > c_6 > 0$, 4) The prior $p(\phi)$ has support over the entire concept family Φ and is bounded above everywhere.

Except from the above five adapted assumptions from Xie et al. (2021), we have an another mild assumption:

Assumption B.6 (Distribution consistency)**.** The pretraining distribution p and prompt distribution p_{prompt} satisfy $\forall (x_q, y_q) \sim p_{prompt}, p(x_q, y_q) = p_{prompt}(x_q, y_q)$.

This assumption indicates that the chosen prompt distribution is a sub-distribution of the pretraining distribution and the joint distribution of x_q and y_q is consistent across p and p_{prompt}. This assumption avoids the situations where there are concept shifts between p and p_{prompt}, *i.e.*, all $y \sim p_{prompt}$ are known in p and can find an exact match for each x_q in p.

B.2. Theoretical Proof

We first show that given a query image x_q, when conditioned on a concept ϕ^* and prompt S_n, the most probable text output token for y_q is the same as the class in the prompt distribution p_{prompt} with maximum probability. Then, we show that: in our denoising in-context learning, when achieving the most likely prediction z output by the MLLM predictor, the class text y_q in the pretraining distribution p is the same as the one found by the prompt distribution p_{prompt}, which is the exact match for the give image x_q.

Before we start analyzing the binary prediction z, we first investigate the most probable class y given prompt and query image $\arg\max_y p(y|S_n, x_q)$.

Theorem B.7. *Assume that the above assumptions hold, if for all $\phi \in \Phi$, $\phi \neq \phi^*$, the concept ϕ^* satisfies the distinguishability condition: $\sum_{j=1}^{k} KL_j(\phi^* \| \phi) > \epsilon^{\phi}_{start} + \epsilon^{\phi}_{delim}$, then as $n \to \infty$, the prediction y according to the pretraining distribution is*

$$\arg\max_y p(y|S_n, x_q, \phi^*) \to \arg\max_y p_{prompt}(y|x_q). \tag{12}$$

Thus, the in-context predictor f_n achieves the optimal $0-1$ risk: $\lim_{n\to\infty} \mathcal{L}_{0-1}(f_n) = \inf_f \mathcal{L}_{0-1}(f)$.

The detailed proof of this theorem is similar to Xie et al. (2021), please refer to the Section D of the original paper.

Under this assumption, the in-context predictor is guaranteed to have the highest probability of generating the class description y that exactly matches the query image x_q. In another way, when x_q does not belong to y, the probability $p(y|S_n, x_q)$ is less than the optimal value.

Lemma B.8. *Under the same condiction of Theorem B.7, the prediction z according to the pretraining distribution is*

$$\arg\max_z p(z|S_n, x_q, y_q, \phi^*) \to \arg\max_z p_{prompt}(z|x_q, y_q). \tag{13}$$

Lemma B.8 can be easily derived based on Theorem B.7 by considering y as a fixed prompt in Q.

Theorem B.9. *Assume that the above assumptions hold, as $n \to \infty$, when achieving the largest prediction probability of z given prompt under concept ϕ^*, the corresponding class description y follows the same y obtained from the prompt distribution:*

$$\arg\max_y p(z|S_n, x_q, y, \phi^*) \to \arg\max_y p_{prompt}(z|x_q, y) \tag{14}$$

Proof. Since we already have Theorem B.7, if we can prove that $\arg\max_y p(y|S_n, x_q, \phi^*) = \arg\max_y p(z|S_n, x_q, y, \phi^*)$ and $\arg\max_y p_{prompt}(y|x_q) = \arg\max_y p_{prompt}(z|x_q, y)$, then we can complete the justification.

$$p(z|S_n, x_q, y, \phi^*) = \sum_{h_q^{start} \in \mathcal{H}} p(z|h_q^{start}) p(h_q^{start}|S_n, x_q, y, \phi^*). \tag{15}$$

By expanding the last term, we have:

$$p(h_q^{start}|S_n, x_q, y, \phi^*) = \frac{p(x_q, y|h_q^{start}, S_n, \phi^*) p(h_q^{start})}{p(x_q, y)} \tag{16}$$

$$\propto \frac{p(x_q, y|h_q^{start}, S_n, \phi^*)}{p(x_q, y)} \tag{17}$$

where $p(h_q^{start})$ is considered as a constant. Moreover, based on Assumption B.6, the joint distribution $p(x_q, y)$ is predefined by the pretraining distribution, which does not affect the marginal distribution of z, thus we can have

$$\frac{p(x_q, y|h_q^{start}, S_n, \phi^*)}{p(x_q, y)} = \frac{p(y|S_n, x_q, h_q^{start}, \phi^*) p(x_q|h_q^{start})}{p(x_q, y)} \tag{18}$$

$$\propto p(y|S_n, x_q, h_q^{start}, \phi^*) p(x_q|h_q^{start}). \tag{19}$$

Since the change of y does not affect the quantity of $p(z|h_q^{start})$, therefore, applying argmax on both sides of the equation holds for finding the optimal y:

$$\arg\max_y p(z|S_n, x_q, y, \phi^*) = \arg\max_y \sum_{h_q^{start} \in \mathcal{H}} p(z|h_q^{start}) p(y|S_n, x_q, h_q^{start}, \phi^*) \tag{20}$$

$$= \arg\max_y p(y|S_n, x_q, h_q^{start}, \phi^*). \tag{21}$$

Similarly,

$$p_{prompt}(z|x_q, y) = \sum_{h_q^{start} \in \mathcal{H}} p(z|h_q^{start}, \phi^*) p_{prompt}(h_q^{start}|x_q, y), \tag{22}$$

$$p_{prompt}(h_q^{start}|x_q, y) = \frac{p_{prompt}(x_q, y|h_q^{start}) p_{prompt}(h_q^{start})}{p_{prompt}(x_q, y)} \tag{23}$$

$$\propto p_{prompt}(x_q, y|h_q^{start}) \tag{24}$$

$$\propto p_{prompt}(y|x_q, h_q^{start}) p_{prompt}(x_q|h_q^{start}), \tag{25}$$

$$\arg\max_y p_{prompt}(z|x_q, y) = \arg\max_y \sum_{h_q^{start} \in \mathcal{H}} p_{prompt}(z|h_q^{start}, \phi^*) p_{prompt}(y|x_q, h_q^{start}) \tag{26}$$

$$= \arg\max_y p_{prompt}(y|x_q, h_q^{start}), \tag{27}$$

where the change of y still does not affect the quantity of $p_{prompt}(z|h_q^{start}, \phi^*)$. Since

$$p(y|S_n, x_q, \phi^*) = \sum_{h_q^{start} \in \mathcal{H}} p(y|h_q^{start}, S_n, x_q, \phi^*) p(h_q^{start}|S_n, x_q, \phi^*), \tag{28}$$

$$p_{prompt}(y|x_q) = \sum_{h_q^{start} \in \mathcal{H}} p_{prompt}(y|h_q^{start}, x_q) p_{prompt}(h_q^{start}|x_q), \tag{29}$$

it is easy to find that

$$\arg\max_y p_{prompt}(y|x_q, h_q^{start}) = \arg\max_y p_{prompt}(y|x_q), \tag{30}$$

$$\arg\max_y p(y|S_n, x_q, h_q^{start}, \phi^*) = \arg\max_y p(y|S_n, x_q, \phi^*). \tag{31}$$

Thus, we have that as $n \to \infty$,

$$\arg\max_y p(z|S_n, x_q, y, \phi^*) \to \arg\max_y p_{prompt}(z|x_q, y). \tag{32}$$

□

Lemma B.8 and Theorem B.9 together show that when given a query image x_q, if the chosen query class description y_q is the true class of x_q, then under the given assumptions, the binary prediction z for judging the correctness of the image-text pair would be the maximum value compared to all other class descriptions $y \neq y_q, y \in \mathcal{Y}$. Therefore, we can justify that using an in-context predictor can help identify the true class label of a given image.

C. Additional Details

We run all our experiments on 8 NVIDIA 3090 GPUs using the Pytorch framework. During MVT, we freeze the MLLM backbone model to stop generating gradients that might cause additional computational costs. Then, for the vision encoder, we use `model.eval()` to produce vision predictions. Additionally, the predictions are evaluated and corrected. Based on the rectified predictions, we use `torch.optim.Adam()` or `torch.optim.SGD()` optimizer to fine-tune the vision model for 3 epochs. Note that we conduct fair comparisons in each experiment by using the same optimizer. Due to the memory of ViT-g is larger, thus we use `torch.optim.SGD()` to optimize ViT-g model and `torch.optim.Adam()` for other vision models. The vision encoder is trained with `torch.float32` precision to prevent overflow. The batch size for ViT-L vision encoder is 16 and the batch size for ViT-g is 8 with 2 accumulation steps. The learning rate of the training process is $5e-7$ and the cosine scheduler for ViT-L with `torch.optim.Adam()`. Due to limited GPU memory, we fine-tune ViT-g with `torch.optim.SGD()` of learning rate $1e-4$ and 0 momentum. Besides, experiments in Sec. D.2 in the Appendix do not follow the settings mentioned above. Because the vision encoders to be fine-tuned have a large capacity gap with the vision encoders in MLLMs. We need to fine-tune the vision encoder to match the performance of MLLMs. Therefore, the learning rate is adjusted to $1e-4$ and the training epoch is adjusted to 20. Note that all the fine-tuned data from the evaluation dataset are the chosen ones for therapy. Then we test the performance of all baseline methods on a split-out test set.

In DICL, our prompt for multi-class classification tasks is as follows:

```
This image {replace_token} shows a photo of <#text>, True or False; Answer:
```

where the `{replace_token}` is further replaced by the image feature, and `<#text>` is further replaced by the class name. The MMICL and Otter model we use can be found at https://github.com/haozhezhao/mic and https://github.com/Luodian/Otter, respectively. All our fine-tuned vision models can be directly found in Openai CLIP models: https://huggingface.co/openai.

Table 5. Classification accuracy (%) of baseline CLIP models and our method with MMICL (Zhao et al., 2023) and Otter (Li et al., 2023b) as the VLMs on 5 ID datasets and 5 OOD datasets. We compare the performance of our method, and the fine-tuned models supervised by our method with the baseline models, i.e., ViT-L from CLIP (Radford et al., 2021). Fine-tuning with both MMICL and Otter improves the classification accuracy.

MLLM	Method	ID					OOD				
		IN-Val	IN-V2	CIFAR10	CIFAR100	MNIST	IN-A	IN-R	IN-SK	IN-V	iWildCam
None	CLIP	75.8	70.2	95.6	78.2	76.4	69.3	86.6	59.4	51.8	13.4
MMICL	MVT	75.2	**70.8**	**97.9**	78.9	53.0	71.2	88.1	59.0	62.1	**25.0**
	+FT	**76.9**	70.5	96.7	**82.0**	79.2	**75.1**	**89.5**	**61.4**	**68.8**	-
Otter	MVT	74.2	67.4	94.7	70.1	52.0	64.1	85.2	59.5	51.9	16.2
	+FT	76.3	70.1	96.6	81.8	**81.3**	73.5	88.7	60.0	55.7	-

Table 6. Classification accuracy (%) of baseline CLIP models and our method with MMICL (Zhao et al., 2023) and Otter (Li et al., 2023b) as the VLMs on 4 subsets of DomainBed datasets, including VLCS, PACS, OfficeHome, and DomainNet. We compare the performance of our method and the fine-tuned models supervised by our method with the baseline models, i.e., ViT-L from CLIP (Radford et al., 2021). Fine-tuning with both MMICL and Otter improves the classification accuracy.

MLLM	Datasets method	VLCS				PACS				OfficeHome				DomainNet					Avg
		0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	4	
None	CLIP	74.9	83.5	80.3	74.5	97.8	97.4	97.5	99.4	87.7	92.7	85.7	85.6	61.1	62.1	60.2	78.4	51.1	80.6
MMICL	MVT	83.8	89.0	87.2	80.3	97.6	97.5	**98.0**	99.4	87.7	93.4	89.0	88.5	61.3	62.1	60.4	78.7	53.4	82.8
	+FT	**84.2**	**89.8**	**87.9**	**82.5**	**98.0**	**98.2**	**98.0**	**99.8**	**90.9**	**95.0**	**90.9**	**90.8**	**62.5**	**63.8**	**62.4**	**80.1**	**54.0**	**84.0**
Otter	MVT	67.5	77.4	73.7	66.6	97.0	96.3	96.5	99.0	85.6	89.9	83.6	83.3	56.5	58.6	56.3	74.1	46.5	77.0
	+FT	76.8	87.7	82.3	77.4	**98.0**	97.7	**98.0**	**99.8**	88.7	93.4	87.7	87.1	62.0	63.0	61.3	79.7	53.4	82.0

D. Additional Quantitative Comparisons

Here we provide extensive quantitative comparisons of different MLLM and vision models, in various robustness settings.

D.1. Quantitative Comparison using Otter

First, similar to Table 1 in the main paper, here we conduct additional experiments on various ImageNet-based datasets and DomainBed datasets using ViT-L but a different MLLM backbone: Otter (Li et al., 2023b). The results are shown in Tables 5 and 6. We find that the performance of MVT is dependent on the MLLM backbone: when using Otter as the backbone model for MVT, the OOD performance would slightly degrade from the performance of MMICL, which could be due to the capability of MLLM to conduct ICL. However, the rectified predictions can still contain useful information to boost the performance of vision models. In several cases in ImageNet-Val, MNIST, and ImageNet-R, Otter with fine-tuning can still improve the visual robustness to the best or second-best results.

Table 7. Classification accuracy (%) of baseline CLIP models and our method on 5 ID datasets and 5 OOD datasets. We compare the performance of our method, and the fine-tuned models supervised by our method with the baseline models, including ResNet-50 and ViT-B/32. The supervisor MLLM is MMICL (Zhao et al., 2023).

Arch	Method	ID					OOD				
		IN-Val	IN-V2	CIFAR10	CIFAR100	MNIST	IN-A	IN-R	IN-SK	IN-V	iWildCam
RN50	CLIP	59.7	52.6	71.5	41.9	**58.5**	23.9	60.7	35.4	31.1	8.2
	MVT	**76.2**	**70.8**	**80.2**	**49.7**	50.8	**47.5**	**72.9**	**41.6**	**54.1**	**14.5**
	+FT	66.3	65.7	75.1	46.9	47.3	32.1	64.4	36.5	38.2	-
ViT-B	CLIP	62.9	56.1	89.9	**65.0**	47.9	32.2	67.9	41.9	30.5	10.9
	MVT	**77.5**	**71.0**	**92.5**	60.4	**51.5**	**60.6**	**83.0**	**47.8**	**53.1**	**19.3**
	+FT	66.3	66.0	90.1	59.5	46.6	38.8	68.7	43.1	37.6	-

D.2. MVT on Additional Vision Models

Then, we conduct MVT using MMICL but using different vision backbone models such as ViT-B and ResNet-50 on ImageNet and DomainBed datasets. The results are shown in Tables 7 and 8. We can see that the performance of MVT is quite strong compared to other vision models which shows over 10% and 4% improvements in ImageNet datasets and Do-

Table 8. Classification accuracy (%) of baseline CLIP models and our method on 4 subsets of DomainBed datasets, including VLCS, PACS, OfficeHome, and DomainNet. We compare the performance of our method and the fine-tuned models supervised by our method with the baseline models, including ResNet-50 and ViT-B/32. The supervisor MLLM is MMICL (Zhao et al., 2023).

Arch	Datasets / method	VLCS				PACS				OfficeHome				DomainNet					Avg
		0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	4	
RN50	CLIP	75.0	82.3	81.3	75.0	91.3	90.3	90.0	96.2	71.7	80.9	69.4	67.8	47.2	46.8	44.9	64.0	32.9	71.0
	MVT	**84.3**	**88.0**	**88.7**	**81.8**	**96.0**	**96.1**	**95.4**	**98.8**	**77.2**	**85.3**	**77.5**	**75.4**	**46.1**	**46.3**	**43.4**	**61.7**	**33.2**	**75.0**
	+FT	83.7	87.3	88.1	81.3	95.6	95.7	95.1	98.6	75.9	85.0	75.8	74.7	45.3	45.6	43.0	60.4	32.6	74.3
ViT-B	CLIP	74.0	82.0	79.6	74.4	93.6	92.8	93.0	98.2	79.2	86.4	77.4	76.3	49.7	54.3	51.0	68.7	40.7	74.8
	MVT	**84.2**	**87.3**	**88.4**	**82.8**	**96.7**	**96.4**	**96.6**	**98.8**	**84.0**	**89.3**	**82.9**	**81.5**	**49.5**	**53.1**	**51.5**	**69.9**	**41.7**	**78.5**
	+FT	76.0	84.8	81.3	81.6	92.9	88.8	89.4	93.3	81.1	88.3	80.8	77.7	47.2	52.5	48.7	66.9	40.5	74.8

mainBed datasets, respectively. Especially on ImageNet-V2, ImageNet-A, ImageNet-R, and ImageNet-V, the performance improvement of MVT are encouragingly over 15%, 24%, 12%, and 23%, respectively. After fine-tuning, the performance can be improved in most cases, such as ResNet-50 is further improved by 13.1% and 3.3% correspondingly on ImageNet-V2 and DomainBed thanks to MMICL.

Table 9. Classification accuracy (%) of baseline CLIP models and our method with MMICL (Zhao et al., 2023) as the VLM on 15 corruptions and 5 severities of ImageNet-C datasets. We compare the performance of our method and the fine-tuned models supervised by our method with the baseline models, i.e., ViT-L from CLIP (Radford et al., 2021). The fine-tuned models with our MVT method have the best performance.

Arch	Datasets / method	Gaussian Noise						Shot Noise						Impulse Noise					
		1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
MMICL	CLIP	69.8	66.7	59.7	46.9	30.6	54.7	70.5	64.9	57.7	43.6	32.1	53.8	65.7	60.2	55.9	45.0	32.7	51.9
	MVT	70.1	67.5	61.2	**49.8**	**33.6**	56.4	70.8	66.8	59.2	**46.1**	**35.6**	55.7	66.3	61.9	58.4	47.5	35.6	53.9
	+FT	**71.0**	**67.9**	**61.3**	48.7	33.5	**56.5**	**72.0**	**67.1**	**60.1**	**46.1**	35.2	**56.1**	**68.5**	**64.2**	**59.8**	**48.9**	**35.8**	**55.4**
		Defocus Blur						Glass Blur						Motion Blur					
		1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
	CLIP	66.1	62.4	53.0	43.4	35.0	52.0	65.5	59.3	40.5	33.8	25.4	44.9	70.9	66.8	59.9	49.5	41.8	57.8
	MVT	67.1	63.3	55.8	**47.6**	**38.8**	54.5	67.1	61.3	42.8	36.0	29.4	47.3	71.9	67.7	60.9	51.5	43.2	59.0
	+FT	**68.8**	**64.1**	**56.3**	47.4	38.4	**55.0**	**68.9**	**64.8**	**45.2**	**37.6**	**30.2**	**49.3**	**72.8**	**69.1**	**62.1**	**52.7**	**45.3**	**60.4**
		Zoom Blur						Snow						Frost					
		1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
	CLIP	62.2	55.9	49.8	43.9	37.3	49.8	68.3	61.2	61.9	56.1	52.6	60.0	68.5	61.2	53.8	51.1	46.6	56.2
	MVT	64.1	57.3	52.0	45.7	38.7	51.6	69.2	61.5	62.9	57.1	54.0	60.9	69.5	61.5	54.2	52.7	47.4	57.1
	+FT	**65.2**	**59.2**	**54.2**	**48.8**	**41.4**	**53.8**	**70.6**	**63.9**	**64.6**	**59.2**	**55.6**	**62.8**	**71.9**	**65.2**	**57.9**	**56.4**	**51.5**	**60.6**
		Fog						Brightness						Contrast					
		1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
	CLIP	69.8	67.9	65.0	61.3	52.0	63.2	74.3	74.0	72.8	70.6	68.1	72.0	70.6	69.3	64.8	52.4	35.1	58.4
	MVT	70.7	69.2	66.5	62.6	53.8	64.6	74.7	74.1	72.6	71.1	68.8	72.3	70.9	69.9	65.1	52.9	36.9	59.1
	+FT	**72.5**	**71.3**	**69.5**	**67.1**	**60.3**	**68.1**	**76.0**	**75.1**	**74.3**	**73.1**	**71.1**	**73.9**	**73.5**	**73.5**	**70.2**	**59.2**	**42.7**	**63.8**
		Elastic						Pixelate						JPEG					
		1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
	CLIP	69.2	50.6	64.1	53.1	30.4	53.5	71.0	70.4	66.2	60.1	54.6	64.5	70.8	67.7	65.1	58.0	45.3	61.4
	MVT	70.0	51.1	65.8	55.2	**32.7**	55.0	71.7	70.7	66.5	61.9	57.3	65.6	**72.5**	**69.6**	**67.5**	**60.5**	47.7	**63.6**
	+FT	**70.7**	**53.6**	**67.7**	**58.5**	32.2	**56.5**	**72.8**	**71.8**	**69.1**	**62.9**	**57.7**	**66.9**	71.2	68.9	65.8	60.0	**48.8**	62.9

D.3. Robustness against Visual Corruptions

Further, we consider the visual robustness against image corruptions by evaluating our method on a robustness benchmark: ImageNet-C (Hendrycks & Dietterich, 2019). Specifically, there are 15 different types of corruption with different corruption severities varied from 1 to 5. Here we cover all scenarios to evaluate our method using MMICL as a backbone model and a baseline method CLIP ViT-L. The results are shown in Table 9. We can see that our method shows very strong performance in all scenarios. Compared to CLIP, using MVT can improve the performance by over 2%, and through fine-tuning, the performance is further boosted by over 4%. The encouraging results again demonstrate the effectiveness of our method.

(a) One-To-One (O2O)

(b) Many-To-Many (M2M)

Figure 11. Figures are from Lynch et al. (Lynch et al., 2023), the letters on each images denote a certain background. There are two spurious correlation types in the Spawrious dataset, namely O2O and M2M. In the O2O setting, each dog class is correlated to one certain background type and different distributions have different correlation probabilities as shown by the bar below the O2O figure. As for the M2M setting, multiple classes and backgrounds are correlated together and the correlation changes to different groups of classes and backgrounds during testing.

D.4. Robustness against Spurious Correlation

Moreover, we consider a common distribution shift scenario where the training dataset and test dataset have different foreground and background correlation, *i.e.*, spurious correlation. Specifically, as shown in Figure 11 standing for the Spawrious dataset that we use, there are two different settings: One-To-One (O2O) correlation where each class is correlated to one background type with a certain probability. The foreground objects in the training dataset and test dataset have different probabilities to be combined with a certain background. For the Many-To-Many (M2M) setting, the foregrounds and backgrounds are split into subgroups that contain multiple classes and background types. When different subgroups are correlated together between training and test datasets, the M2M spurious correlation is formed and brings more complexity. In the Spawrious dataset, there are three levels of hardness based on correlation probability difference between training and test datasets, namely easy, medium, and hard. Here, we consider all scenarios and show the results in Table 10. We can see that the MVT method can outperform the ViT-L and ViT-g baseline methods in all scenarios, which leads to the conclusion that our method is robust to spurious correlations and can identify the class of interests despite the changing backgrounds.

Table 10. Performance comparison between MVT and CLIP on robustness against spurious correlation using Spawrious dataset.

Type	O2O_easy	O2O_medium	O2O_hard	M2M_easy	M2M_medium	M2M_hard	Avg.
ViT-L	94.1	95.4	93.3	96.7	95.0	92.5	94.5
MVT	**95.8**	**96.3**	**93.6**	**96.8**	**95.8**	**92.9**	**95.2**
ViT-g	94.6	97.0	92.6	96.7	95.6	94.8	95.2
MVT	**95.3**	**97.4**	**92.8**	**96.8**	**96.6**	**95.4**	**95.7**

Figure 12. Examples of celebA photos with different attributes.

D.5. Performance on Recognizing Fine-grained Attributes

Additionally, here we further explore the capability of recognizing subtle attributes based on the CelebA dataset (Liu et al., 2015). Particularly, we consider 12 face attributes, as shown in Figure 12. For each attribute, we testify whether a learning model could correctly identify the attribute in a given image. Here we compare our MVT method with CLIP ViT-L and ViT-g, and the performance of MVT produced by conducting therapy on ViT-L and ViT-g models.

Particularly, since CelebA is a binary classification task, here we design different prompts for vision models and our MLLM. For CLIP models, we use `The person in this image is <#classname>` as text input, where `<#classname>` of each attribute is shown in Table 11. For our method, we still designed one positive prompt and one negative prompt for each ICL round. Specifically, for "Male" attribute, our in-context instruction is as follows:

> Question: Is the person in this image {replace_roken} a male? Answer: True;
>
> Question: Is the person in this image {replace_roken} a female? Answer: False;
>
> Question: Is the person in this image {replace_roken} a male? Answer:

in which is first exemplar demonstrates an image of a male positively described as male, the second exemplar shows an image of a male negatively described as female, and finally, we ask whether the input image is a male and use the output of MLLM as the prediction.

The results on CelebA are shown in Table 12, we observe that our method is quite effective in recognizing fine-grained attributes and its performance significantly surpasses ViT-L and ViT-g with a large margin. Especially in attributes such as "Blond_Hair", "Mustache", and "Wearing_Necktie", the performance improvements are even over 10% on both two CLIP models, and the final averaged results on all 12 attributes, the total improvements are 8.1% and 3.4% for ViT-L and ViT-g,

Table 11. Class names for 12 chosen attributes.

Attribute	-1	+1
Male	a woman	a man
Wear_Hat	not wearing a hat	wearing a hat
Smiling	not smiling	smiling
Eyeglasses	not wearing eye glasses	wearing eye glasses
Blond_Hair	not having blond hair	having blind hair
Mustache	not having mustache	having mustache
Attractive	not attractive	attractive
Wearing_Lipstick	not wearing lipstick	wearing lipstick
Wearing_Necklace	not wearing necklace	wearing necklace
Wearing_Necktie	not wearing necktie	wearing necktie
Young	not young	young
Bald	not bald	bald

Table 12. Performance comparison between MVT and CLIP on recognizing fine-grained attributes using CelebA dataset.

Attr.	Male	Wearing_Hat	Smiling	Eyeglasses	Blond_Hair	Mustache	Attractive	Wearing_Lipstick	Wearing_Necklace	Wearing_Necktie	Young	Bald	Avg.
ViT-L	63.0	60.8	64.5	75.8	36.2	29.0	42.0	30.8	38.0	37.5	66.6	86.3	52.5
MVT	**74.0**	**67.0**	**65.4**	**76.1**	**53.0**	**55.8**	**42.4**	**39.4**	**38.6**	**53.9**	**73.5**	**88.1**	**60.6**
ViT-g	98.5	75.5	70.4	83.8	46.0	66.6	58.2	72.5	43.5	28.4	54.1	91.3	65.7
MVT	**98.9**	**77.2**	**71.0**	**84.1**	**58.3**	**74.9**	**59.0**	**73.2**	**43.6**	**41.2**	**56.1**	**91.8**	**69.1**

respectively. Therefore, it is reasonable to conclude that our method can be effectively conducted on fine-grained attribute recognition and significantly outperforms several powerful vision models.

E. Complementary Experimental Results

Due to the non-negligible inference time of MLLMs, we cannot conduct performance evaluation of VQA and MVT on full test set of ImageNet variants and CIFAR variants. Therefore, some results in Table 1 are conducted under a test set split. To further fully validate our performance under the original test set, we conduct MVT on a vision model which successfully distills the knowledge and enables efficient inference. The results are shown in Table 13. We can see that the effectiveness of MVT is again validated on all datasets. We can also observe a similar effect to the main paper: the performance improvement on OOD data is much more significant than that on ID data, which still validates the effectiveness of MVT on enhancing visual robustness.

Table 13. Performance comparison on full test set evaluation.

Arch	Method	IN-Val	IN-V2	CIFAR10	CIFAR100	MNIST	IN-A	IN-R	IN-SK	IN-V
ViT-L	CLIP	75.5	69.8	95.5	78.3	76.4	70.8	87.8	59.6	51.5
	MVT+FT	77.1	70.4	98.1	81.8	79.3	75.3	89.9	61.9	68.6
ViT-g	EVA	80.1	73.6	98.3	88.7	62.5	69.4	92.2	68.9	64.9
	MVT+FT	81.2	74.9	98.6	90.9	65.6	75.0	93.9	71.1	70.6

F. Additional Performance Analysis

In this section, we carefully conduct additional performance analysis to further validate the effectiveness of our MVT.

F.1. Analysis on MLLM Guided Fine-Tuning

We find that our MLLM-guided fine-tuning is quite effective in further improving the prediction accuracy based on MVT corrections. To investigate why such a fine-tuning process can help the learning performance, here we compare the prediction logits of image examples from ImageNet and its variant datasets before and after the fine-tuning process. As we already shown in Figure 4 in the main paper, our MVT framework can effectively find the examples misclassified by the vision model, thus we randomly select some images that are further processed with our therapy and fine-tuning. The logits are shown in Figure 13, we can observe that after fine-tuning, the predictions of the previously incorrect examples are finally rectified which shows that the proposed MVT method is indeed helpful for visual recognition. Moreover, we also observe that the ground truth logit values of some examples are quite small which means that the vision models are very confident in their incorrect predictions. Thanks to our MVT, we can not only successfully find the ground truth classes, but also help produce a less confident prediction for the previously incorrect examples. We hypothesize such an effect could help mitigate the overfitting issue (Lin et al., 2023a; 2024) and thus generalize better on unseen OOD test sets.

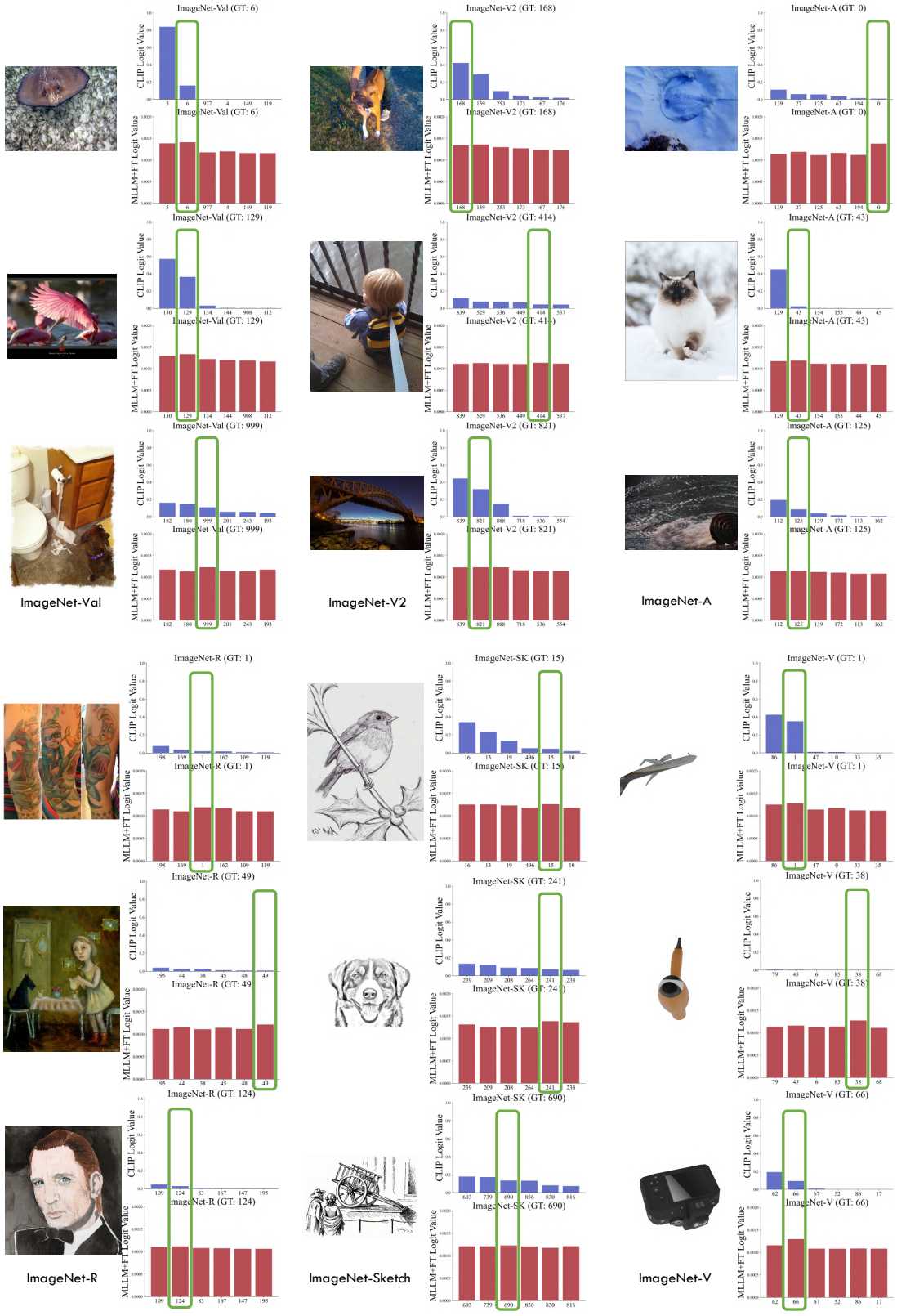

Figure 13. Images with the prediction logits before and after fine-tuning. Examples are randomly chosen from ImageNet-Val, ImageNet-V2, ImageNet-A, ImageNet-R, ImageNet-Sketch, and ImageNet-V. The green boxes highlight the ground truth class logits.

F.2. Analysis on Various DICL Designs

To justify why using just one positive-negative exemplar pair can effectively conduct vision tasks, here we provide an analysis of using three different DICL designs. Specifically, we consider using two positive exemplars, two negative exemplars, and two incorrect exemplars as baseline DICL designs and compare them to the original MVT results in the main paper. For example, the three types of instructions are shown as:

Two positive exemplars:

> Question: This image <IMG_PRE#0> shows a photo of <PRE#0>, True or False? Answer: True;
> Question: This image <IMG_CLN#c> shows a photo of <CLN#c>, True or False? Answer: True;
> Question: This image <IMG_Query> shows a photo of <PRE#0>, True or False? Answer:

Two negative exemplars:

> Question: This image <IMG_CLN#c> shows a photo of <PRE#0>, True or False? Answer: False;
> Question: This image <IMG_CLN#c+1> shows a photo of <PRE#0>, True or False? Answer: False;
> Question: This image <IMG_Query> shows a photo of <PRE#0>, True or False? Answer:

Two incorrect exemplars:

> Question: This image <IMG_CLN#c> shows a photo of <PRE#0>, True or False? Answer: True;
> Question: This image <IMG_PRE#0> shows a photo of <PRE#0>, True or False? Answer: False;
> Question: This image <IMG_Query> shows a photo of <PRE#0>, True or False? Answer:

Table 14. Performance of using different DICL prompt designs. We also compare them with the original MVT using ViT-L and fine-tuning performance using the MMICL backbone.

MLLM	Method	ID					OOD				
		IN-Val	IN-V2	CIFAR10	CIFAR100	MNIST	IN-A	IN-R	IN-SK	IN-V	iWildCam
None	CLIP	75.8	70.2	95.6	78.2	76.4	69.3	86.6	59.4	51.8	13.4
MMICL	Two incorrect	73.2	68.3	93.2	77.6	54.6	68.7	86.8	58.2	58.3	22.3
	Two positive	74.5	69.5	94.1	77.9	56.2	69.0	87.4	57.5	60.9	23.7
	Two negative	75.0	69.0	96.6	78.1	55.7	69.3	87.8	58.1	61.2	24.3
	MVT	75.2	**70.8**	**97.9**	78.9	53.0	71.2	88.1	59.0	62.1	**25.0**
	+FT	**76.9**	70.5	96.7	82.0	**79.2**	**75.1**	**89.5**	**61.4**	**68.8**	-

The results are shown in Table 14. We find that all three designs are inferior to our method MVT, thus we know that using one positive and negative exemplar pair is the most effective instruction strategy. Moreover, we find that using Two negative exemplars is slightly better than the other two, which manifests that negative exemplars are quite important in deciding the effectiveness of MVT, further justifying that using a noise transition matrix to find the most probable negative classes is essential to our method. Furthermore, we also find that using two incorrect exemplars achieves the worst performance, even inferior to CLIP ViT-L. This could be because that feed incorrect instructions could indeed mislead the learning performance, thus showing degradation.

F.3. Analysis on OOD Robustness

To further investigate the performance on facing OOD data with varied strengths, here we use ImageNet-C to show how increasing the corruption severity could affect the prediction of the vision model and MLLM (MMICL). For illustration, we randomly choose four examples from ImageNet-C and plot their top-6 prediction logits from the vision model. Moreover, for clear comparison, we also choose the top-6 prediction classes as our therapy candidates (which is different from the settings of MVT) to show the MLLM prediction results. As shown in Figure 14, we can see that as the severity increases, the prediction of CLIP logit values is highly unstable. When the severity is large, the final top-1 prediction could be

incorrect. However, the prediction of MLLM remains consistent through all severities. Even when CLIP prediction is incorrect, MLLM can still correctly find the ground truth classes. Only when there is no ground truth in top-6 predictions, MLLM can be uncertain about the final prediction. Therefore, the robustness of MLLM against corruption could justify that our MVT is quite effective on OOD tasks compared to vision models.

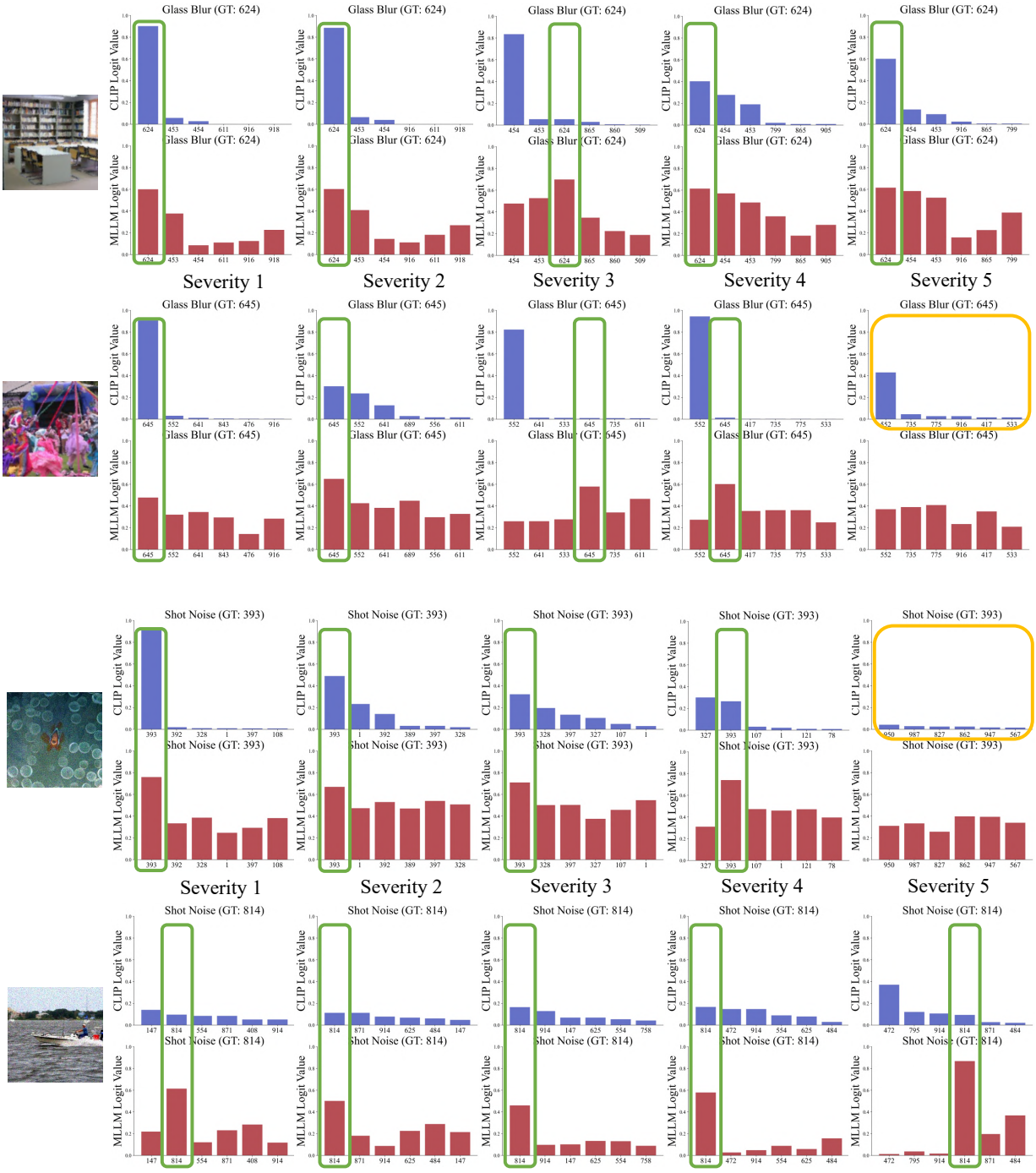

Figure 14. Robustness analysis using ImageNet-C. The first two rows are examples of Glass Blur corruption, and the last two rows are examples of Shot Noise corruption. The green boxes highlight the logits of ground truth classes, and the orange boxes denote there are no ground truth predictions in the CLIP top-6 predictions.

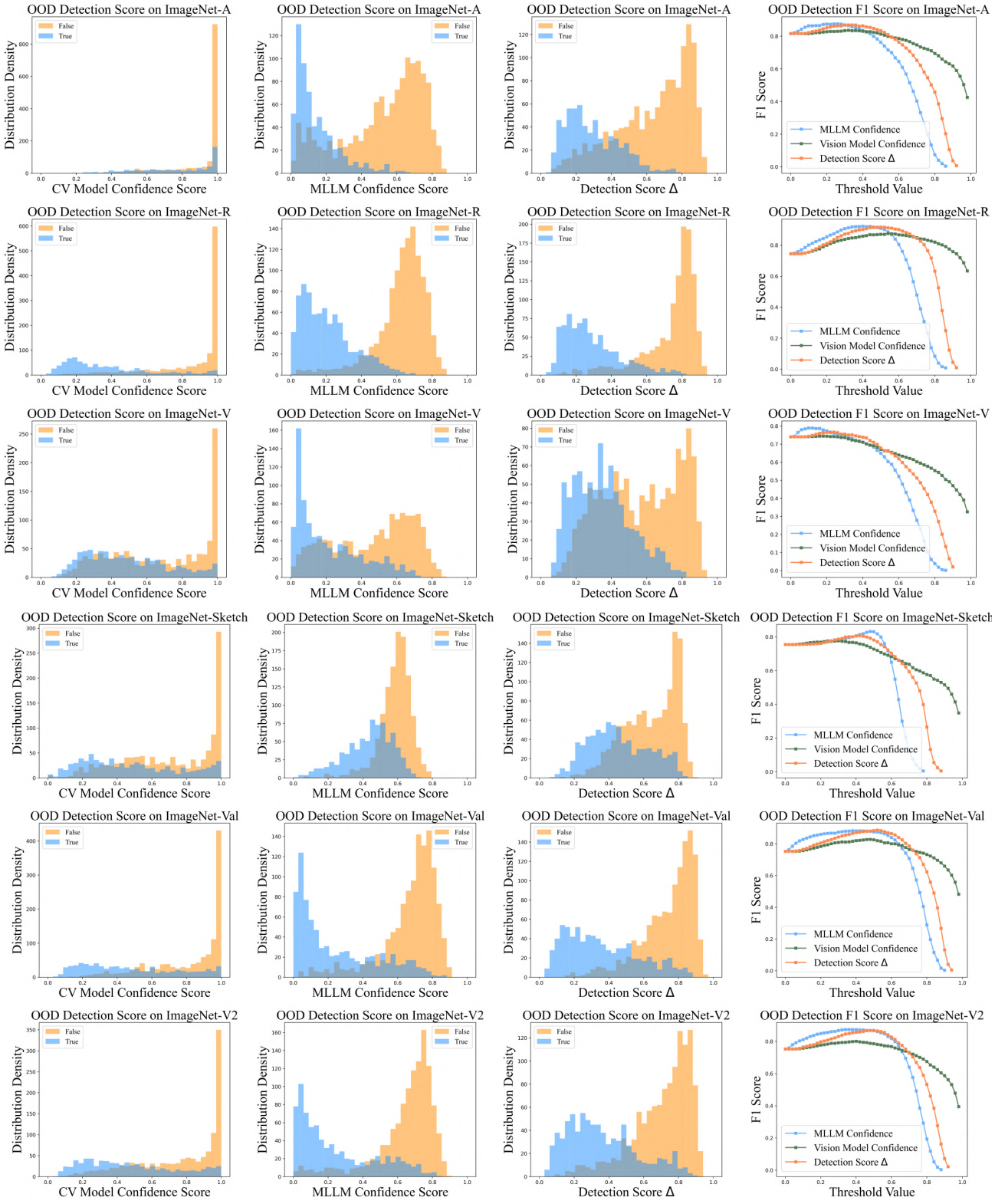

Figure 15. OOD detection performance on ImageNet-A, ImageNet-R, ImageNet-V, ImageNet-Sketch, ImageNet-Val, and ImageNet-V2 datasets using ViT-L. The first three columns: Density distribution of different OOD detection scores; The last column: F1 score values by varying the OOD detection threshold δ.

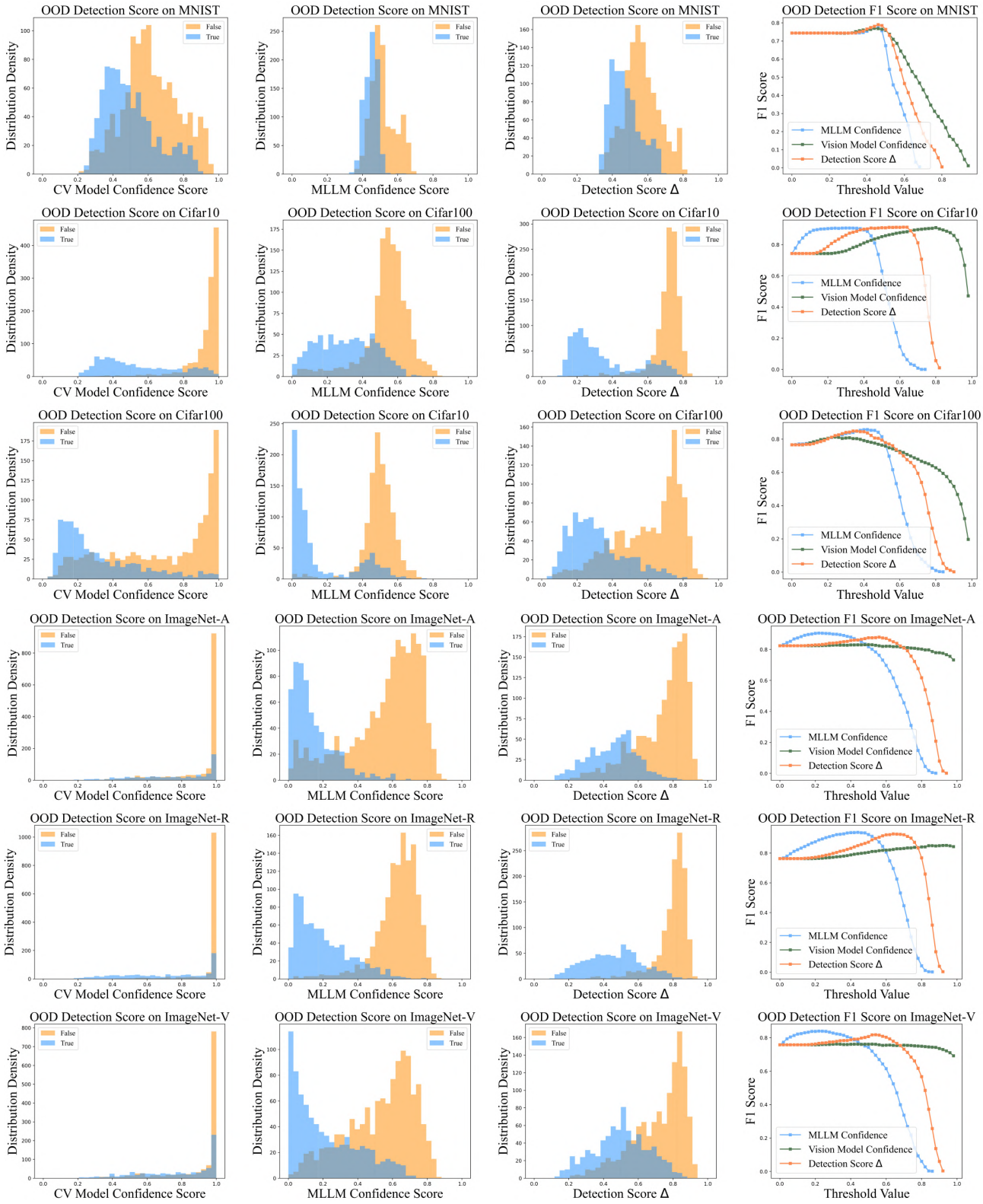

Figure 16. First three rows: OOD detection performance on MNIST, CIFAR10, and CIFAR100 datasets using ViT-L. Last three rows: OOD detection performance on ImageNet-A, ImageNet-R, and ImageNet-V datasets using ViT-g. The first three columns: Density distribution of different OOD detection scores; The last column: F1 score values by varying the OOD detection threshold δ.

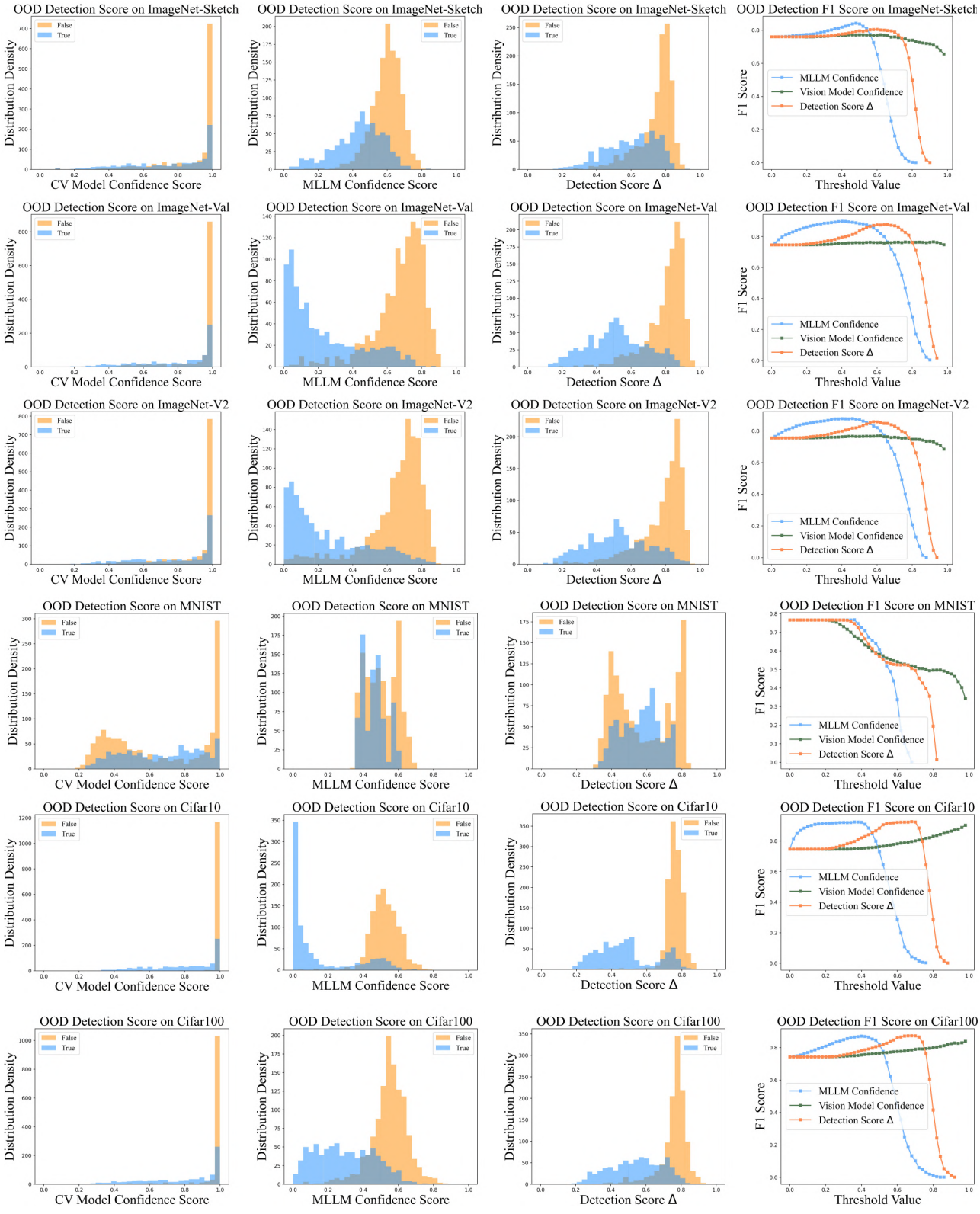

Figure 17. OOD detection performance on ImageNet-Sketch, ImageNet-Val, ImageNet-V2, MNIST, CIFAR10, and CIFAR100 datasets using ViT-g. The first three columns: Density distribution of different OOD detection scores; The last column: F1 score values by varying the OOD detection threshold δ.

F.4. Analysis on OOD Detection

As an extension of the main paper, here we further provide more results on OOD detection. We consider ImageNet-A, ImageNet-R, ImageNet-V, ImageNet-Sketch, ImageNet-Val, ImageNet-V2, MNIST, CIFAR10, and CIFAR100 datasets using both ViT-L and ViT-g models. The results are shown in Figures 15, 16, 17. We can observe similar phenomenons as in the main paper: MLLM can effectively identify open-class data and vision models can identify close-class data. By combining the prediction confidence of MLLM and vision models, our detection score Δ can be successfully leveraged to conduct OOD detection, and the F1 score of Δ can achieve the largest value with a reasonable detection threshold δ. Still, some datasets are relatively challenging compared to other datasets: ImageNet-Sketch and MNIST. This could be due to that the classification performance on these two datasets is not outstanding, about 50% to 68% for both vision models and MLLMs. Moreover, the patterns of such two datasets are quite simple: both of them are handwritten lines without complex natural features, which could further hinder the extraction of useful knowledge, thus leading to sub-optimal detection performance. Overall, the OOD detection performance of MVT is effective on most datasets and the capability of recognizing unknown examples could provide insight to the OOD detection field. We plan to further investigate its potential in future works.

G. Limitation and Broader Impact

In this paper, we proposed an effective framework that aims to enhance the visual robustness of vision models by exploiting the knowledge of MLLMs instead of requiring additional human annotations. Based on our proposed DICL strategy, the paradigm of MLLMs can be perfectly aligned to vision tasks and achieve encouraging results. However, we found that the performance of our MVT is highly related to the ICL capability of MLLMs. Moreover, since there are only two existing MLLMs that possess multimodal ICL power, the potential of our MVT framework could be further improved when sophisticated MLLMs with multimodal ICL abilities are developed in the future. We hope our work could bring insight into the multimodal learning field with our DICL to achieve alignment between MLLMs and vision learning tasks. Based on our work, we believe many traditional fields that are related to visual recognition such as weakly-supervised learning, OOD detection, and fine-grained image classification could be further advanced by effectively leveraging the knowledge of MLLMs.