MEGACOIN: Enhancing Medium-Grained Color Perception for Vision-Language Models

Ming-Chang Chiu[1 2 3], Shicheng Wen[3], Pin-Yu Chen[4], Xuezhe Ma[3]
[1]APAL
[2]Cornell University
[3]University of Southern California
[4]IBM Research
mingchangc@cornell.edu

Abstract

In vision-language models (VLMs), the ability to perceive and interpret color and physical environment is crucial for achieving contextually accurate understanding and interaction. However, despite advances in multimodal modeling, there remains a significant lack of specialized datasets that rigorously evaluate a model's capacity to discern subtle color variations and spatial context—critical elements for situational comprehension and reliable deployment across real-world applications. Toward that goal, we curate MegaCoin , a high-quality, human-labeled dataset based on real images with various contextual attributes. MegaCoin consists of two parts: MegaCoin-Instruct, which serves as a supervised fine-tuning (SFT) dataset for VLMs; and MegaCoin-Bench, an annotated test set that can be used as a stand-alone QA dataset. MegaCoin provides three annotated features for 220,000 real images: foreground color, background color, and description of an object's physical environment, constituting 660k human annotations. In addition, MegaCoin can be applied to benchmark domain generalization (DG) algorithms. We explore benchmarking DG methods in the linear probing setup for VLM and show some new insights. Last but not least, we show that VLMs, including GPT-4o, have subpar color recognition capabilities, and fine-tuning with MegaCoin can result in improved performance on visual evaluation tasks. In certain cases, MegaCoin fine-tuned small-scale opensource models such as LLaVA and Bunny can outperform closed-source GPT-4o. We hope the utilities of MegaCoin can shed light on the directions VLMs can improve and provide a more complex platform for domain generalization algorithms.

Multimodal Instruction Tuning — https://github.com/charismaticchiu/MegaCOIN

Domain Generalization — https://github.com/charismaticchiu/MegaCoin-DomainBed

Extended version — https://arxiv.org/pdf/2412.03927

Introduction

The rapid advancement of vision-language models (VLMs) has brought about an increasing demand of high-quality datasets that can facilitate improved training and evaluation of these models. A critical factor influencing VLM performance is the quality and granularity of the training data. While existing datasets have substantially contributed to progress in this domain, there remains exploration in datasets that provide medium-grained annotations, which can also capture the complex relationships between visual content and textual descriptions. In this work, we introduce MEGACOIN, a meticulously curated dataset comprised of high-quality, human-annotated images with various contextual attributes. MEGACOIN is dually motivated by certain deficits of recognition capabilities in computer vision models such as color vision (Chiu, Chen, and Ma 2023; Chiu et al. 2022) and the need for contextual information in VLM training and evaluation, addressing a key limitation in existing datasets that often lack human-labeled contextual elements. By "medium granularity," we refer to a level of detail that goes beyond simple object labels but stops short of bounding boxes or pixel-level segmentation. This granularity allows for rich and synthesizable descriptions based on human-annotated labels, striking a balance between detail and generalizability.

MEGACOIN leverages *real* images and provides a realistic and challenging setting for training and evaluating VLMs. Our dataset comprises 220,000 real images, each annotated with three human-provided attributes: foreground color (FGD), background color (BGD), and the physical environment (ENV) in which the object is situated. This results in a total of 660,000 annotations. This rich annotation scheme allows for a deeper image understanding that goes beyond simple image labels, enabling more nuanced SFT for VLMs. MEGACOIN is designed to serve three primary purposes:

- *Multimodal Instruction Tuning*: MEGACOIN can serve as an SFT dataset for VLMs. This resource enhances VLMs' ability to understand and generate accurate textual descriptions of visual content, with a particular focus on contextual elements often overlooked in existing datasets. By incorporating medium-grained color and environmental annotations, MEGACOIN enables VLMs to develop a more nuanced understanding of visual scenes.
- *Evaluation Benchmark*: The annotated test set of MEGACOIN can be extracted and used as a standalone QA dataset, MEGACOIN-Bench. This new benchmark provides a robust tool for evaluating the performance of VLMs on basic visual perception tasks, with a specific

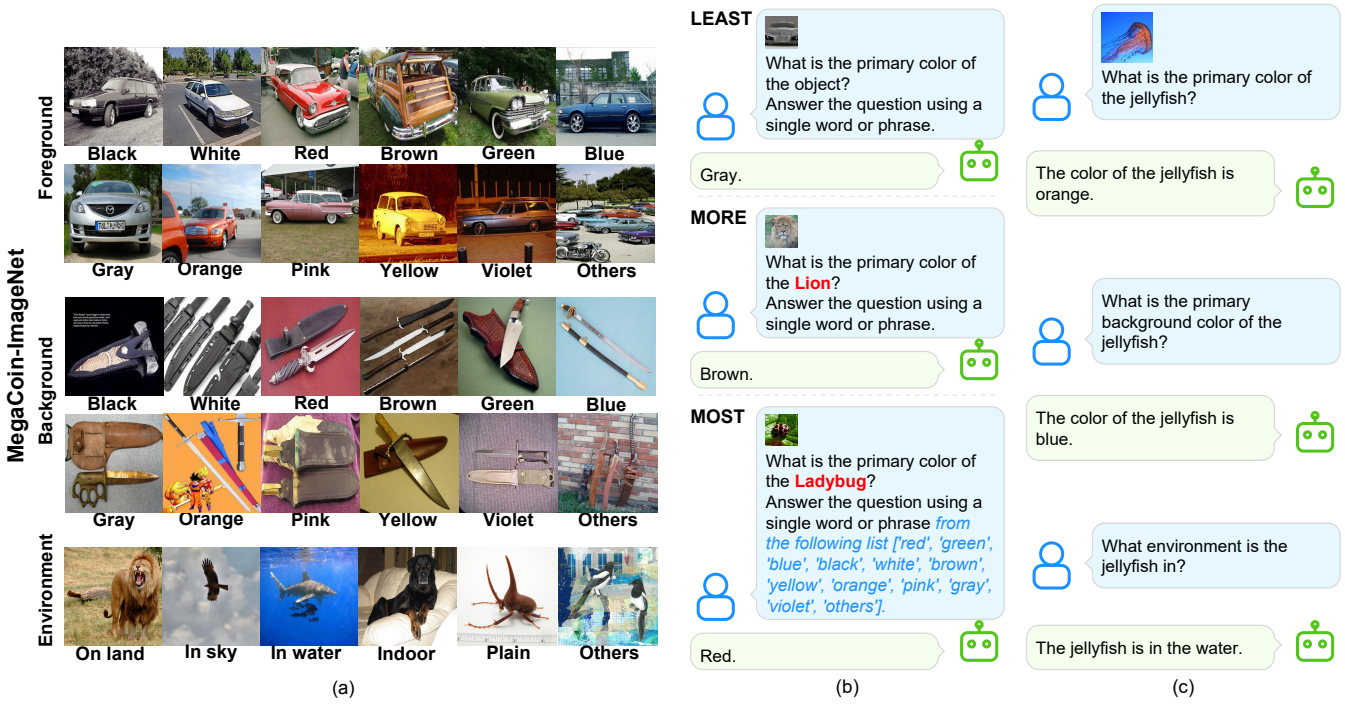

Figure 1: Overview of MEGACOIN. (a) Examples of our human-annotated MEGACOIN, consisting of three distinct attributes, foreground/background color, physical environment. (b & c) We use MEGACOIN as an instruction fine-tuning data (MEGACOIN-Instruct) and a benchmark (MEGACOIN-Bench). (b) Examples of 3-tier MEGACOIN-Bench evaluation for a single image. (c) Example of MEGACOIN-Instruct SFT pairs for a single image.

emphasis on contextual understanding and color perception.

- *Domain Generalization (DG)*: MEGACOIN can be used to explore the effectiveness of DG algorithms in different setups. This utility provides valuable insights from comparing which DG algorithms are useful, a critical capability for real-world applications.

Our experimental results show rooms for improvement for VLMs in contextual understanding and shed light on the performance of different domain generalization algorithms, revealing MMD(Li et al. 2018), CORAL(Sun and Saenko 2016) and ERM(Vapnik 1991) are on par with each other as measured by different metrics in the linear probing setting. These findings not only demonstrate the utility of MEGACOIN for domain generalization research but also provide practical insights for improving the robustness of VLMs across diverse visual contexts.

In summary, our contributions in this work include:

- MEGACOIN-Bench, a new QA benchmark and a novel Tiered-Multiple Choice QA (Tiered-MQA) scheme for evaluating VLMs' visual perception and contextual understanding capabilities.
- MEGACOIN-Instruct, a novel dataset that enhances visual alignment for VLMs with contextual medium-grained annotations and that can train a 13B or 8B model like LLaVA or Bunny to surpass GPT-4o.
- New benchmark for DG algorithms, showcasing MEGACOIN's potential to facilitate advancements in this critical area of machine learning research.

Through MEGACOIN, we aim to bridge the gap between existing datasets and the needs of advanced VLMs, providing a resource that enables more contextually-aware, robust, and generalizable VLMs.

Related Works

Multimodal SFT datasets for VLMs SFT datasets play a crucial role in the development of VLMs. Current approaches to VLM SFT typically involve exposing the model to a diverse set of task-specific instructions and corresponding image-text pairs (Liu et al. 2023c, 2024). Some popular SFT datasets for VLMs include functions like captioning (Lin et al. 2014; Laboratory 2024), visual reasoning (Liu et al. 2023a; Hudson and Manning 2019; Johnson et al. 2017), chart (Liu et al. 2023b), science (Lu et al. 2022; Kembhavi et al. 2016), OCR (Mishra et al. 2019; Singh et al. 2019), etc. This process allows the model to learn to follow natural language instructions and generate appropriate responses based on visual inputs. However, a significant limitation of many existing datasets is their reliance on synthetic data using ChatGPT (Liu et al. 2024, 2023a). Although synthetic data offer advantages in terms of scalability, they are often costly to construct and lack definitive ground truth. MEGACOIN stands out by offering a dataset of real images with human-annotated attributes. These annotations allow for efficient construction of instruction pairs for SFT while keeping future extensions to produce a guided synthetic version of MEGACOIN.

VLM QA datasets Visual Question Answering (VQA) datasets serve as important benchmarks for evaluating the capabilities of VLMs. Several popular test beds have emerged, often spanning multiple domains, such as MME(Fu et al. 2024), MMBench (Liu et al. 2025), MM-Vet(Yu et al. 2023), VQA(Antol et al. 2015), GQA(Hudson and Manning 2019), or more focused capability, such as MMC (Liu et al. 2023b), MathVista (Lu et al. 2024), etc.

While these benchmarks offer valuable insights into VLM performance across various domains, MEGACOIN takes a different approach by focusing on a fundamental visual capability: color perception and contextual understanding. By providing a dedicated benchmark for color-related queries and environmental context, MEGACOIN-Bench aims to serve researchers and practitioners a choice to evaluate basic visual comprehension skills, which has downstream safety implications (Furness et al. 2003; Chiu et al. 2022).

Domain generalization dataset Domain generalization (DG) is an important area of study in machine learning. Common image benchmarks include ColorMNIST (Arjovsky et al. 2019), Spawrious (Lynch et al. 2023), MetaShift (Liang and Zou 2022), NICO++ (Zhang et al. 2023), Waterbirds (Sagawa et al. 2019), CelebA (Liu et al. 2015), WILDS (Beery, Cole, and Gjoka 2020). We summarize and compare some distinctions between these datasets and MEGACOIN in Tab. 4.

MEGACOIN provides various desiderata of a DG dataset altogether, which is more complete than prior efforts: (1) *multiple* explicitly annotated *Spurious Correlations* (SC), (2) *Real images* that capture natural variations in visual domains, (3) *challenging tasks*: non-binary classification, intra-class heterogeneity and different difficulty levels, (4) can study *Attribute Generalization* (AG) and *Attribute Imbalance* (AI). We refer readers to (Qiao and Low 2024; Yang et al. 2023; Lynch et al. 2023) for detailed definitions.

MEGACOIN Dataset

Dataset Construction

MEGACOIN is a comprehensive dataset that builds upon three widely-used image collections: CIFAR10 (C10), Tiny-ImageNet (TI), and ImageNet (INet). To create MEGACOIN, we utilized all images from C10, the entire TI, and the validation set of INet. This combination provides a diverse range of images across various scales and resolutions, ensuring that MEGACOIN covers a broad spectrum of visual scenarios. C10 represents a good baseline for small-scale object recognition tasks and TI offers a step up in complexity from C10. As for INet, it provides high-resolution, real-world examples across 1,000 diverse categories.

To label the colors of foreground and background, we use the eleven basic colors in English, which are {*White, Black, Grey, Yellow, Red, Blue, Green, Brown, Pink, Orange, Purple/Violet*}. And for the physical environments, we put the images into {*In the Sky, In the Water, Indoor, On Land, Others, Plain*} groups.

For each image, the annotations involve three attributes:

- Foreground Color (FGD): Annotators identify the primary object of the image by the image's class label and select the dominant color from the predefined color set.
- Background Color (BGD): Annotators identify the major color in the background of the primary object.
- Physical Environment (ENV): Annotators choose from the six physical environments that the object is situated.

We show examples of MEGACOIN annotations in Fig. 1(a).

Annotation Process

We collaborate with an industry data curation company, DATUMO, to annotate the foreground, background, and physical environment for each image.

We enforced a bespoke quality control process to ensure label quality — Each image is presented to labelers alongside three predefined attribute sets for annotation. Before submission, labelers are prompted to review and confirm their selections. To maintain high standards, any incorrect selection according to company guidelines triggers an automated notification to the labeler. After repeated errors, the labeler's access is temporarily paused until an administrator reviews and re-enables their account.

MEGACOIN-Bench

Recent advancements in VLMs have unlocked the ability to perform zero-shot tasks that integrate both vision and language, such as Chart (Liu et al. 2023b), etc. However, we want to take a step back and explore their *fundamental abilities of recognizing colors and physical environments*, which is proven to be deficient in computer vision models and hard to mitigate (Chiu et al. 2022). Thus, an evaluation benchmark is essential to assess the performance of various models on these tasks and guide future research and development. Our dataset, MEGACOIN-Bench, constructed from the evaluation split of source images, is designed to fill this gap, offering the following key features: Large-scale benchmark to evaluate VLMs' ability to recognize and interpret visual elements — color, environment, and object — both individually and in combination, complementing prior benchmarks having a small number of images in the related categories (Liu et al. 2025; Fu et al. 2024). MEGACOIN leverages the new annotations and the original labels, to provides a *novel* evaluation methodology for our tasks of interest, which we call Tiered-Multiple Choice QA, a structured format going beyond a single multiple-choice QA pair.

Tiered-Multiple Choice QA (Tiered-MQA) For each image, we can provide VLMs three levels of information, to test how well VLMs rely on additional information to recognize FGD, BGD and ENV. Examples of the tiers are illustrated in Fig. 1(b).

- *Least*: We directly ask the VLM to answer *"What is the primary {FGD} of the object?"* By giving the least amount of information, we test the model's ability to identify the object and then perceive the colors or ENV.
- *More*: We provide main object's class label for the VLM, e.g. *"What is the primary {FGD} of the {Class}?"* Failure to do perform this task would mean that, in addition

to incorrect color perception, the VLM may not be able to recognize what target object is.

- *Most*: We provide both the target object's class label and options of attributes to VLM, such as *"What is the primary {FGD} of the {Class}. Please choose from {Red, Green, ..., Black}."* By giving a constrained set of choice and target object, we expect it would be easier for VLMs to comprehend and output correctly.

The resulting MEGACOIN-Bench consists of 210k images to test each of the three basic perception task, and each annotation can be used for three tiers of evaluations.

We first test MEGACOIN-Bench on the state-of-the-art closed-source VLM, GPT-4o, and discover that it may have low capability in our basic vision tasks (Tab. 2), leading us to leverage the training set of source images to construct MEGACOIN-Instruct for SFT.

MEGACOIN-Instruct

We leverage the training set from the source images of MEGACOIN to develop an SFT dataset for VLMs, referred to as MEGACOIN-Instruct. More specifically, we use the three attributes we newly annotate along with the original object label to formulate instruction-following pairs as the SFT dataset.

Our goal is to investigate how MEGACOIN-Instruct can enhance model alignment with multimodal tasks, particularly in recognizing and reasoning about colors, objects, and the physical environments in images.

The specific instruction pairs follow structured templates, as exemplified in Fig. 1. For instance, we use instructions such as *"What is the primary foreground color of the {Object}? The primary foreground color of the {Object} is {FGD}."* And similar templates for identifying background colors and physical environment.

This approach enables efficient creations large scale instruction pairs for the models to learn precise associations and improve its understanding of visual concepts within diverse contexts.

Medium Granularity MEGACOIN employs a medium-grained annotation scheme, striking a balance between detailed description and broad applicability. This choice is justified by several factors:

- VLM Compatibility: This level of detail is suited for SFT of vision-language models, providing informed details without overwhelming the model with overly specific information.
- Efficiency: This approach allows for a larger number of annotated images within the given budget, increasing the dataset's overall utility.
- Flexibility: medium-grained annotations can be easily aggregated for coarser-grained tasks or used as a basis for more fine-grained analysis if needed (e.g. Sec.).

The resulting SFT dataset consists of 450k instruction-image pairs, and we summarize MEGACOIN's statistics in Tab. 1. This rich annotation scheme provides a solid foundation for basic vision-language tasks, particularly those involving color perception and contextual understanding.

Statistics	Num
MEGACOIN-Instruct	450k
– background	150k
– foreground	150k
– environment	150k
MEGACOIN-Bench	210k
– background	70k
– foreground	70k
– environment	70k

Table 1: MEGACOIN Statistics.

Experiment

Baselines

We leverage LLaVA-v1.5-13B (Liu et al. 2024) and Bunny-Llama3-v1.1-8B (He et al. 2024) as our open-source backbones, and GPT-4o (Achiam et al. 2023) for closed model evaluation. LLaVA-1.5 excels in visual reasoning and image captioning by integrating a LLaMA-based language model with vision encoders, making it effective for tasks requiring deep understanding of visual data. Bunny-1.1, designed for efficient multimodal learning, uses high-quality training data to deliver strong visual understanding while minimizing computational costs, making it a versatile choice for VLM tasks. GPT-4o is one of the best multi-modal LLMs that lead performances across various tasks.

Experiment Setups

To fine-tune the open-source models within resource constraint, we employ the Low-Rank Adaptation (LoRA) (Hu et al. 2021) technique. We combine the original fine-tuning dataset of each backbone model with our MEGACOIN-Instruct. In addition, based on MEGACOIN's source image, we can explore three combinations of MEGACOIN-Instruct: CIFAR10 (C10), TinyImageNet (TI), and the combination of CIFAR10 and TinyImageNet (C10+TI). For each backbone model, we evaluate four variants based on the SFT data: the original backbone models as baselines, and the three using the aforementioned data combinations. This approach enables us to assess the impact of our dataset on model performance systematically. All models were fine-tuned using 8 NVIDIA A40 GPUs.

Results

MEGACOIN-Bench Results We first observe in Tab. 2 that GPT4o, one of the cutting-edge closed-source model in many aspects for its massive and surperior training data and method, only scores passing grade on MEGACOIN-Bench, re-affirming the importance of our motivation.

Comparison with GPT4o We find that fine-tuning on MEGACOIN-Instruct significantly closed the gap of these models across nearly all tasks on MEGACOIN-Bench, *some even surpass GPT-4o*. LLaVA-1.5 finetuned with MEGACOIN-Instruct, in particular, consistently outperforms GPT-4o, even in the more challenging splits like ImageNet, showcasing the effectiveness of MEGACOIN-Instruct in enhancing model performance. The fine-tuned Bunny-1.1 also outperformed GPT-4o in most tasks, although we observed a

Model		Tier	C10FGD	C10BGD	C10ENV	TIFGD	TIBGD	TIENV	INetFGD	INetBGD	INetENV
Closed	**GPT-4o**	LEAST	61.6 (+0.0)	58.29 (+0.0)	30.19 (+0.0)	49.32 (+0.0)	53.45 (+0.0)	26.08 (+0.0)	52.87 (+0.0)	53.75 (+0.0)	29.85 (+0.0)
		MOST	65.37 (+0.0)	62.44 (+0.0)	76.2 (+0.0)	48.11 (+0.0)	56.38 (+0.0)	68.09 (+0.0)	53.48 (+0.0)	57.47 (+0.0)	76.85 (+0.0)
Open-Sourced	**LLaVA-1.5-13B**	LEAST	59.74 (-1.86)	55.28 (-3.01)	31.7 (+1.51)	46.54 (-2.78)	47.72 (-5.73)	25.94 (-0.14)	47.62 (-5.25)	48.26 (-5.49)	28.69 (-1.16)
		MOST	63.16 (-2.21)	55.43 (-7.01)	64.99 (-11.21)	48.05 (-0.06)	47.99 (-8.39)	62.9 (-5.19)	52.64 (-0.84)	47.11 (-10.36)	72.14 (-4.71)
	+C10	LEAST	65.65 (+4.05)	70.43 (+12.14)	64.14 (+33.95)	52.99 (+3.67)	63.24 (+9.79)	50.24 (+24.16)	50.56 (-2.31)	59.57 (+5.82)	55.3 (+25.45)
		MOST	69.67 (+4.3)	69.69 (+7.25)	81.0 (+4.8)	53.32 (+5.21)	61.59 (+5.21)	72.55 (+4.46)	53.28 (-0.2)	58.62 (+1.15)	79.8 (+2.95)
	+TI	LEAST	66.01 (+4.41)	62.02 (+3.73)	68.59 (+38.4)	65.11 (+15.79)	61.71 (+8.26)	71.89 (+45.81)	62.79 (+9.92)	59.41 (+5.66)	74.43 (+44.58)
		MOST	66.0 (+0.63)	62.52 (+0.08)	81.19 (+4.99)	64.41 (+16.3)	59.18 (+2.8)	76.17 (+8.08)	62.07 (+8.59)	58.21 (+0.74)	82.04 (+5.19)
	+C10&TI	LEAST	70.53 (+8.93)	69.1 (+10.81)	57.02 (+26.83)	66.15 (+16.83)	62.47 (+9.02)	62.41 (+36.33)	62.96 (+10.09)	60.86 (+7.11)	64.39 (+34.54)
		MOST	69.82 (+4.45)	68.05 (+5.61)	81.46 (+5.26)	64.74 (+16.63)	62.88 (+6.5)	74.73 (+6.64)	62.23 (+8.75)	61.69 (+4.22)	81.3 (+4.45)
	Bunny-1.1-8B	LEAST	64.37 (+2.77)	58.82 (+0.53)	27.36 (-2.83)	51.98 (+2.66)	52.3 (-1.15)	27.43 (+1.35)	53.64 (+0.77)	52.87 (-0.88)	30.39 (+0.54)
		MOST	68.89 (+3.52)	60.38 (-2.06)	62.69 (-13.51)	52.94 (+4.83)	54.12 (-2.26)	58.15 (-9.94)	57.35 (+3.87)	51.75 (-5.72)	74.77 (-2.08)
	+C10	LEAST	68.86 (+7.26)	66.73 (+8.44)	25.98 (-4.21)	56.01 (+6.69)	59.47 (+6.02)	26.78 (+0.7)	55.09 (+2.22)	55.15 (+1.4)	29.33 (-0.52)
		MOST	69.98 (+4.61)	67.55 (+5.11)	66.14 (-10.06)	55.31 (+7.2)	58.81 (+2.43)	59.33 (-8.76)	58.04 (+4.56)	54.18 (-3.29)	73.71 (-3.14)
	+TI	LEAST	69.3 (+7.7)	65.75 (+7.46)	26.37 (-3.82)	59.83 (+10.51)	60.42 (+6.97)	29.26 (+3.18)	56.67 (+3.8)	55.76 (+2.01)	31.58 (+1.73)
		MOST	69.51 (+4.14)	66.63 (+4.19)	62.23 (-13.97)	57.69 (+9.58)	62.4 (+6.02)	57.56 (-10.53)	58.09 (+4.61)	54.77 (-2.7)	72.67 (-4.18)
	+C10&TI	LEAST	69.81 (+8.21)	67.14 (+8.85)	27.52 (-2.67)	60.13 (+10.81)	60.43 (+6.98)	30.02 (+3.94)	56.94 (+4.07)	55.4 (+1.65)	33.06 (+3.21)
		MOST	70.13 (+4.76)	69.46 (+7.02)	68.68 (-7.52)	58.27 (+10.16)	63.12 (+6.74)	61.51 (-6.58)	58.42 (+4.94)	55.9 (-1.57)	74.92 (-1.93)

Table 2: MEGACOIN-Bench Results (%). Small-scale MEGACOIN-Instruct-finetuned opensource VLMs can outperform GPT-4o. (.) shows the accuracy gain/gap to GPT-4o.

notable exception with the C10ENV task, where Bunny-1.1's accuracy was 2.67% to 10.06% lower than GPT-4o. Bunny-1.1 also exhibited slightly lower performance on the ImageNet dataset, which was not included in its training set, suggesting potential limitations in generalization to new domains. We want to reiterate that LLaVA-1.5 we use is a 13B model, and Bunny-1.1 being a 8B model, which is massively smaller than closed-source models and without full-finetuning.

The results show that our SFT dataset and benchmark play a vital role in enhancing model performance in the visual perception tasks. Our structured prompts and varied domains in our benchmark allowed LLaVA-1.5 to consistently outperform GPT-4o, showcasing the effectiveness of targeted fine-tuning. The occasional performance drop for C10ENV and Bunny-1.1 on unseen tasks in ImageNet highlights areas where broader generalization may be limited. Our dataset and benchmark effectively push models beyond the capabilities seen in zero-shot settings, providing a meaningful boost in domain-specific performance that zero-shot models like GPT-4o could not achieve.

Open-source VLMs On LLaVA-1.5 and Bunny-1.1, we observed improvement when models were fine-tuned and tested within the same domain (Tab. 2). For instance, LLaVA-1.5 demonstrates notable gains on the LEAST instruction within TIFGD, with accuracy rising from 46.54% to 65.11% after fine-tuning with TI. This improvement illustrates that our dataset plays a positive role in data augmentation on the color and environment recognition tasks.

OOD performances We also notice out-of-distribution (OOD) domain adaptation capabilities using MEGACOIN-Instruct, with performance gains even on previously unseen data. LLaVA-1.5, for instance, shows an increase on TIENV under the MOST prompt from 62.9% to 72.55% after fine-tuning on the C10 split of MEGACOIN-Instruct. Similarly, LLaVA-1.5's accuracy on INetBGD with MOST prompts increased from 47.11% to 61.69% when augmented with C10+TI, despite the INet dataset being both unseen and of higher resolution than C10+TI.

These findings highlight how MEGACOIN-Bench's Tiered-MQA design enables uncovering models' ability to improve specific tasks and MEGACOIN-Instruct can benefit domain adaptation.

Qualitative Results In Fig. 2, we show qualitative examples, where using MEGACOIN-Instruct helps correct wrong perception, to showcase how our MEGACOIN-Instruct improves VLM performance on MEGACOIN-Bench. (1) *FGD misperception*: in an image of a "Great White Shark" (row 1, 6th image), where the VLM may initially answer "white" based on the species name, which is a contextual bias. After MEGACOIN-Instruct, it correctly answers "gray," reflecting the actual color. Similar biases occur with images of a "Green Lizard" and "Red Fox," where the model may rely on species names instead of their true colors. (2) *BGD misinterpretation:* in a traffic light image (row 2, 5th image), the VLM misinterpreted the background as "red," which may be due to it focusing on the light's color instead of the surroundings. After training with MEGACOIN-instruct, the VLM accurately identifies the background as "blue," resolving object-background distinctions. (3) *ENV mis-judgement:* an image of a seagull flying above the sea (row 3, 1st image) have the VLM answer "in the water." We suspect that it is because the background confuses the model, but after training with MEGACOIN-Instruct, it correctly identifies the environment as "in the sky."

These qualitative examples reveal the usefulness of MEGACOIN-Instruct in enhancing the VLM ability in basic vision tasks and how MEGACOIN-Bench can facilitate such analyses of VLMs.

Common Benchmark Results Tab. 3 shows the SFT results of our fine-tuned models on common benchmarks such as MMBench, MME, VQAv2, and GQA. In addition to the four variants mentioned in Sec. , we also explore fine-tuning the VLMs with combined original SFT data and the entire MEGACOIN (dubbed 220k) because they are not evaluated on ImageNet. We suspected fine-tuning LLaVA-1.5 and Bunny-1.1 VLMs with MEGACOIN would lead to an overall performance drop as MEGACOIN focuses on more ba-

Figure 2: Failure cases on MEGACOIN-Bench before/after training with MEGACOIN-Instruct. After fine-tuning with MEGACOIN-Instruct, we are able to have the VLMs recognize the correct colors and environments.

	MMBench	MME-P	MME-C	VQAv2	GQA
LLaVA-1.5-13B	67.70	1530.43	298.57	80	63.3
+C10	**68.21**	**1530.91**	276.78	**80.07**	**63.4**
+TI	67.35	**1561.84**	294.28	**80.18**	**63.4**
+220k	66.41	**1574.82**	279.64	**80.16**	**63.49**
Bunny-1.1-8B	76.63	1647.81	305.36	82.51	64.25
+C10	**77.06**	1639.08	**344.29**	82.35	**64.46**
+TI	**76.63**	**1651.93**	**332.14**	**82.52**	64.18
+220k	**78.44**	1630.19	**352.14**	**82.65**	**64.47**

Table 3: Performance of VLMs on common benchmarks.

sic visual capabilities. Surprisingly, we observe Bunny and LLaVA show improvements on more than half of the cases; and on VQAv2 and GQA, fine-tuned models is either on par with or slightly better than the baselines. These results suggest that fine-tuning with our dataset would lead to performance gains in the majority cases, despite a few moderate performance drops on specific model-data combinations. It further corroborates the value of MEGACOIN.

Domain Generalization

Our medium-grained contextual labels in MEGACOIN can serve as *spurious* attributes, suitable for benchmarking DG algorithms along with a few other desiderata mentioned in Sec. . We evaluate the domain generalization algorithms in the *linear probing* setting of VLM in this section. We also provide the training from scratch Resnet18 results, which is closer to the original setup of (Gulrajani and Lopez-Paz 2020), in the Appendix for completeness.

Setup We adopt the implementation of DomainBed (Gulrajani and Lopez-Paz 2020) using the CLIP representation (Radford et al. 2021) and do 10 runs each with different random seeds of seven different DG algorithms, which are ERM, GroupDRO (Sagawa et al. 2019), CORAL(Sun and Saenko 2016), MMD(Li et al. 2018), EQRM(Eastwood et al. 2022), SelfReg(Kim et al. 2021) and VREx(Krueger et al. 2021) based upon their performances in previous studies (Lynch et al. 2023; Gulrajani and Lopez-Paz 2020) and recency (e.g. Eastwood et al.; Kim et al.; Krueger et al. are published after (Gulrajani and Lopez-Paz 2020)).

Merging similar colors Some subgroups have way fewer instances while being similar in hue, such as violet and blue, so we also explore a *merge* version of MEGACOINby merging pink to red, orange to yellow, and violet to blue to reduce the complexities of our dataset.

Result

MMD is the best on generalization, followed by CORAL and ERM. The generalization results across all datasets (Tab. 5) indicate that MMD achieves the highest average accuracy, at 89.06% with the lowest standard deviation (0.19), indicating its robust generalization across different dataset domains. ERM and CORAL follow closely with similar average accuracy (both at 89.04%) but slightly higher variance compared to MMD, suggesting slight sensitivity. GroupDRO and SelfReg maintain competitive performance levels, though consistently lower than MMD and CORAL, while VREx displays both the lowest accuracy (87.64%) and the highest variability (1.26), indicating less stability in its performance.

CORAL is the best with subpopulation shift. Subpopulation shift is a branch in DG and uses worst-group accuracy as the gold-standard metric. Tab. 6 presents the WGA that are evaluated on the same set of subgroups as training. We surprisingly observe that the *subgroup robust* method

Dataset	# Attr.	# Classes	# Train set	# Val. set	# Test set	Max group	Min group	SC	AI	AG
Waterbirds	2	2	4795	1199	5794	3498 (73.0%)	56 (1.2%)	x	x	
CelebA	2	2	162770	19867	19962	71629 (44.0%)	1387 (0.9%)	x		
MetaShift	2	2	2276	349	874	789 (34.7%)	196 (8.6%)	x		
Spawrious (O2O-Easy setup)	6	4	12672	n/a	12672	3073 (24.25%)	3802 (3%)	x	x	
MEGACOIN-C10-FGD	12	10	50000	10000	n/a	15734 (26.2%)	232 (0.4%)	x	x	x
MEGACOIN-TI-FGD	12	200	100000	10000	n/a	25113 (22.8%)	1036 (0.9%)	x	x	x
MEGACOIN-C10-BGD	12	10	50000	10000	n/a	15599 (26.0%)	329 (0.5%)	x	x	x
MEGACOIN-TI-BGD	12	200	100000	10000	n/a	33923 (30.8%)	638 (0.6%)	x	x	x
MEGACOIN-C10-ENV	6	10	50000	10000	n/a	43528 (72.5%)	1373 (2.3%)	x	x	x
MEGACOIN-TI-ENV	6	200	100000	10000	n/a	62311 (56.6%)	681 (0.6%)	x	x	x

Table 4: Common image domain generalization datasets. MEGACOIN imposes more challenge in terms of more attributes and more classes. Also, MEGACOIN supports all SC, AI, AG (see Sec.) evaluation scenarios.

Method	C10FGD	C10FGDmerge	C10BGD	C10BGDmerge	C10ENV	TIFGD	TIFGDmerge	TIBGD	TIBGDmerge	TIENV	Average
ERM	97.68 (0.08)	97.62 (0.13)	97.68 (0.15)	97.72 (0.17)	97.58 (0.12)	80.18 (0.19)	80.63 (0.39)	80.73 (0.34)	80.73 (0.29)	79.89 (0.25)	89.04 (0.21)
GroupDRO	97.61 (0.14)	97.60 (0.08)	97.69 (0.16)	97.75 (0.09)	97.51 (0.12)	79.42 (0.45)	79.92 (0.31)	79.53 (0.41)	80.03 (0.60)	78.54 (0.83)	88.56 (0.32)
CORAL	97.63 (0.10)	97.64 (0.10)	97.74 (0.10)	97.69 (0.11)	97.62 (0.07)	80.37 (0.24)	80.51 (0.32)	80.59 (0.38)	80.78 (0.29)	79.79 (0.37)	89.04 (0.21)
EQRM	97.62 (0.10)	97.69 (0.10)	97.66 (0.16)	97.70 (0.11)	97.54 (0.13)	80.11 (0.31)	80.35 (0.32)	80.31 (0.32)	80.56 (0.23)	79.59 (0.30)	88.91 (0.21)
MMD	97.60 (0.07)	97.64 (0.12)	97.83 (0.07)	97.78 (0.10)	97.57 (0.12)	80.35 (0.21)	80.52 (0.42)	80.54 (0.19)	80.90 (0.31)	79.82 (0.33)	**89.06** (0.19)
SelfReg	97.50 (0.09)	97.51 (0.13)	97.60 (0.08)	97.58 (0.13)	97.51 (0.09)	79.70 (0.29)	79.60 (0.25)	79.78 (0.25)	80.03 (0.32)	79.03 (0.38)	88.58 (0.20)
VREx	97.44 (0.29)	97.52 (0.20)	97.65 (0.15)	97.58 (0.16)	97.50 (0.14)	77.60 (2.29)	77.85 (2.66)	77.82 (2.00)	77.34 (3.10)	78.05 (1.62)	87.64 (1.26)
Average	97.58 (0.12)	97.60 (0.12)	97.69 (0.12)	97.69 (0.12)	97.55 (0.11)	79.68 (0.57)	79.91 (0.67)	79.98 (0.56)	80.05 (0.73)	79.24 (0.58)	88.69 (0.37)

Table 5: Domain generalization algorithms results (%, std) on MEGACOIN in linear probing. We observe that MMD is the overall best, immediately followed by ERM and CORAL.

Trained×Evaluated	ERM	GroupDRO	CORAL	EQRM	MMD	SelfReg	VREx	Avg.
C10FGD×FGD	95.00	95.24	96.19	95.00	95.00	93.57	95.00	95.00
C10BGD×BGD	94.68	93.94	94.68	94.47	94.79	94.57	93.94	94.44
C10ENV×ENV	96.03	95.94	96.23	95.88	96.05	96.01	95.98	96.02
TIFGD×FGD	75.05	75.62	76.48	76.38	74.95	75.81	72.29	75.23
TIBGD×BGD	75.40	74.92	75.81	75.08	75.81	75.56	71.85	74.92
TIENV×ENV	76.97	75.11	76.65	76.73	77.53	76.16	74.99	76.31
Average	85.21	85.13	**86.01**	85.59	85.69	85.28	84.01	85.32

Table 6: Worst group accuracy (%) when using the same training and evaluation subgroups.

like GroupDRO was not among the top performers, meaning there is room for improvement this type of method in the linear probing setup. The CORAL still outperforms the others by at least an average 0.32%, followed by MMD. And VREx still cannot perform at the same level, trailing the second worst by an average of 1.12%.

ERM excels at attribute generalization. Tab. 7 shows the WGA of algorithms trained with a certain set of attribute present, but evaluated using another set of attribute, meaning the subgroups used for evaluation are missing during training. We observe the vanilla ERM is the strongest, followed by CORAL and EQRM, contrary to MMD leading the way in the previous evaluations.

Conclusion

In this work, we constructed MEGACOIN, a large-scale human-annotated dataset containing 660k meduim-grained annotations that can be purposed into a 450k visual instruction tuning dataset (MEGACOIN-Instruct) and a 210k evaluation set (MEGACOIN-Bench), covering basic visual perception tasks. With MEGACOIN-Bench and our Tiered-MQA design, we revealed that current VLMs have room to improve on basic visual understanding, and we investigated

	Trained × Evaluated	ERM	GroupDRO	CORAL	EQRM	MMD	SelfReg	VREx	Avg.
FGD	C10FGD×BGD	93.19	94.36	94.47	94.36	94.47	93.72	94.04	94.09
	C10FGD×ENV	96.27	95.46	95.95	95.94	95.81	96.07	95.41	95.84
	TIFGD×BGD	74.52	73.31	74.19	73.63	73.79	73.79	69.76	73.28
	TIFGD×ENV	77.34	76.43	77.87	77.43	77.41	77.23	74.47	76.88
	Average	85.33	84.89	**85.62**	85.34	85.37	85.20	83.42	85.02
BGD	C10BGD×FGD	96.37	96.44	95.95	95.00	96.76	93.81	96.45	95.83
	C10BGD×ENV	96.17	96.28	96.37	96.04	96.42	95.98	95.88	96.16
	TIBGD×FGD	76.95	75.90	77.52	76.76	75.14	76.48	71.90	75.81
	TIBGD×ENV	78.02	76.26	77.53	77.70	77.72	77.21	75.26	77.10
	Average	**86.88**	86.22	86.84	86.38	86.51	85.87	84.87	86.22
ENV	C10ENV×FGD	95.95	96.32	95.71	95.71	96.43	94.52	94.52	95.59
	C10ENV×BGD	94.04	93.62	94.26	94.04	93.40	94.63	94.36	94.05
	TIENV×FGD	76.06	73.24	76.12	75.91	75.97	75.56	74.87	75.39
	TIENV×BGD	76.85	74.03	75.56	75.89	75.56	75.00	73.79	75.24
	Average	**85.73**	84.30	85.41	85.39	85.34	84.93	84.39	85.07

Table 7: Worst group performance (%) in attribute generalization (AG) setting. Contrary to the observations from generalization performance, CORAL and ERM excel in the AG setting, but MMD remain competitive.

MEGACOIN-Instruct's effectiveness in improving VLM perception capabilities. In addition, due to the medium-grained labels allowing for subgrouping, we investigated MEGACOIN's other usage in terms of benchmarking domain generalization methods and found CORAL, ERM, and MMD excel in different evaluation metrics. Our work helps to enhance the alignment from the upstream basic perception task and serve as a benchmark for domain generalization algorithms. Future work on VLM alignment can explore replacing the template instruction pairs with more powerful VLMs to generate guided instruction pairs to enhance the quality of MEGACOIN. Another promising line of work can integrate MEGACOIN with noisy data (Hendrycks and Dietterich 2019) for both VLM alignment and benchmarking algorithms focusing on out-of-distribution generalization, the dual-purpose of our work.

Acknowledgments

This work was done when Ming-Chang was at USC. We thank Datumo for partnering with us and labeled the datasets. And we thank the anonymous reviewers for their constructive feedback. This work is also supported by USC CARC.

References

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

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Beery, S.; Cole, E.; and Gjoka, A. 2020. The iWildCam 2020 Competition Dataset. *arXiv preprint arXiv:2004.10340*.

Chiu, M.-C.; Chen, P.-Y.; and Ma, X. 2023. Better may not be fairer: A study on subgroup discrepancy in image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4956–4966.

Chiu, M.-C.; Wang, Y.; Kim, D. E. G.; Chen, P.-Y.; and Ma, X. 2022. On human visual contrast sensitivity and machine vision robustness: A comparative study. *arXiv preprint arXiv:2212.08650*.

Eastwood, C.; Robey, A.; Singh, S.; Von Kügelgen, J.; Hassani, H.; Pappas, G. J.; and Schölkopf, B. 2022. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35: 17340–17358.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

Furness, S.; Connor, J.; Robinson, E.; Norton, R.; Ameratunga, S.; and Jackson, R. 2003. Car colour and risk of car crash injury: population based case control study. *Bmj*, 327(7429): 1455–1456.

Gulrajani, I.; and Lopez-Paz, D. 2020. In Search of Lost Domain Generalization. arXiv:2007.01434.

He, M.; Liu, Y.; Wu, B.; Yuan, J.; Wang, Y.; Huang, T.; and Zhao, B. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations*.

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

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2901–2910.

Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; and Farhadi, A. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 235–251. Springer.

Kim, D.; Yoo, Y.; Park, S.; Kim, J.; and Lee, J. 2021. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9619–9628.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, 5815–5826. PMLR.

Laboratory, S. A. 2024. Sharegpt-4o.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5400–5409.

Liang, W.; and Zou, J. 2022. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023a. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Liu, F.; Wang, X.; Yao, W.; Chen, J.; Song, K.; Cho, S.; Yacoob, Y.; and Yu, D. 2023b. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023c. Visual Instruction Tuning.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 216–233. Springer.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. arXiv:2310.02255.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

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

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, 947–952. IEEE.

Qiao, R.; and Low, B. K. H. 2024. Understanding domain generalization: A noise robustness perspective. *arXiv preprint arXiv:2401.14846*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, 443–450. Springer.

Vapnik, V. 1991. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.

Yang, Y.; Zhang, H.; Katabi, D.; and Ghassemi, M. 2023. Change is Hard: A Closer Look at Subpopulation Shift. arXiv:2302.12254.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhang, X.; He, Y.; Xu, R.; Yu, H.; Shen, Z.; and Cui, P. 2023. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16036–16047.