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

Ming-Chang Chiu
USC
mingchac@usc.edu

Shicheng Wen*
UC Davis
scwen@ucdavis.edu

Pin-Yu Chen
IBM Research
pin-yu.chen@ibm.com

Xuezhe Ma
USC
xuezhema@usc.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[1], *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* MEGA-COIN *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.*

1. 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 [5, 6] 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.

---

*Work done while at USC.

[1]Dataset released at https://github.com/charismaticchiu/MegaCOIN

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.

- *Evaluation Benchmark*: The annotated test set of MEGA-COIN 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 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[20], CORAL[37] and ERM[38] 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 (Sec. 4), 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 (Sec. 5), 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 (Sec. 7), 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.

2. 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 [25, 26]. Some popular SFT datasets for VLMs include functions like captioning [19, 22], visual reasoning [14, 15, 23], chart [24], science [16, 29], OCR [32, 36], 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 [23, 26]. 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

Benchmark	Size	Images	Source	Answer
VQA	>1M	General	Annotated	Open
GQA	>1M	General	Synthesized	Open
MME	1.5k	General	Annotated	Y/N
MMBench	3k	General	Repurposed	MQA
MM-Vet	0.2k	General	Repurposed	MQA
MathVista	1.4k	Math	Synthesized	MQA
MMC	2k	Chart/Plot	Internet, Annotated	Open/MQA
MEGACOIN(ours)	210k	Contextual	Annotated	Tiered-MQA

Table 1. VLM QA Benchmark Comparison.

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 (Tab. 1), often spanning multiple domains, such as MME[8], MMBench [27], MM-Vet[40], VQA[2], GQA[14], or more focused capability, such as MMC [24], MathVista [30], etc.

While these benchmarks offer valuable insights into VLM performance across various domains, MEGA-COIN 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, MEGA-COIN-Bench aims to serve researchers and practitioners a choice to evaluate basic visual comprehension skills, which has downstream safety implications [5, 9].

Domain generalization dataset Domain generalization (DG) is an important area of study in machine learning. Common image benchmarks include ColorMNIST [3], Spawrious [31], MetaShift [21], NICO++ [41], Waterbirds [35], CelebA [28], WILDS [4]. We summarize and compare some distinctions between these datasets and MEGA-COIN in Tab. 7.

MEGACOIN provides some desiderata of a DG dataset: (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 [31, 33, 39] for detailed definitions.

3. MEGACOIN Dataset

3.1. Dataset Construction

MEGACOIN is a comprehensive dataset that builds upon three widely-used image collections: CIFAR10 (C10), TinyImageNet (TI), and ImageNet (INet). To create MEGA-COIN, 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).

3.2. 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.

4. MEGACOIN-Bench

Recent advancements in VLMs have unlocked the ability to perform zero-shot tasks that integrate both vision and language, such as Chart [24], etc. However, we want to take a step back and explore their fundamental abilities of recognizing colors and physical environments. 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 [8, 27]. MEGACOIN leverages the new annotations and the original labels, to provides a novel evaluation

Dataset	split	black	blue	brown	gray	green	orange	others	pink	red	violet	white	yellow
C10FGD	train	7414	2658	12910	11085	2495	896	190	212	2946	204	7171	1819
	test	1504	518	2824	1889	447	176	42	60	547	55	1638	300
TIFGD	train	17144	3960	22842	21286	5715	3258	5176	2175	5685	931	4739	7089
	test	1638	387	2271	1993	578	315	536	241	605	105	729	602
C10BGD	train	4830	7833	6995	13129	10592	454	532	347	642	269	3408	969
	test	723	1589	1460	2470	2314	78	227	94	93	60	759	133
TIBGD	train	9172	10575	15754	30844	17064	909	3395	1037	1267	594	5864	3525
	test	894	1069	1500	3079	1754	83	227	124	124	44	784	318
INetFGD	val	9434	2018	12371	8551	2811	1477	1974	863	2175	337	4875	3114
INetBGD	val	4032	4557	10052	12179	9488	519	1813	484	632	210	4277	1757

Table 2. Statistics of MEGACOIN—Foreground (FGD) and Background (BGD) Color.

Dataset	split	in the sky	in the water	indoor	on land	others	plain
C10ENV	train	2962	4931	2261	36642	1006	2198
	test	642	816	702	6886	367	587
TIENV	train	615	6783	19310	56789	9304	7199
	test	66	648	2315	5522	591	858
INetENV	val	495	3609	14116	25663	2775	3342

Table 3. Statistics of MEGACOIN—Physical Environment (ENV).

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. 5), leading us to leverage the training set of source images to construct MEGACOIN-Instruct for SFT.

5. 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. 7).

The resulting SFT dataset consists of 450k instruction-image pairs, and we summarize MEGACOIN's statistics in

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

Table 4. MEGACOIN Statistics.

Tab. 4. This rich annotation scheme provides a solid foundation for basic vision-language tasks, particularly those involving color perception and contextual understanding.

6. Experiment

6.1. Baselines

We leverage LLaVA-v1.5-13B [26] and Bunny-Llama3-v1.1-8B [11] as our open-source backbones, and GPT-4o [1] 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.

6.2. Experiment Setups

To fine-tune the open-source models within resource constraint, we employ the Low-Rank Adaptation (LoRA) [13] 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.

6.3. Results

6.3.1 MEGACOIN-Bench Results

Comparison with GPT4o GPT-4o remains one of the best closed-source model in many aspects for its massive and surperior training data and method. However, we find that fine-tuning on MEGACOIN-Instruct significantly improved the accuracy 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 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 performance improvement when models were fine-tuned and tested within the same domain (Tab. 5). 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.

Additionally, we notice out-of-distribution 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 MORE prompts increased from 44.52% to 62.06% when augmented with C10+TI, despite the INet dataset being both unseen and of higher resolution than C10+TI.

Another interesting observation is that while a notable performance improvement is evident from the LEAST to the MOST prompt especially on ENV, the MORE prompt does not consistently benefit performance. In some cases, it

	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)
		MORE	60.04 (+0.0)	58.24 (+0.0)	30.74 (+0.0)	42.79 (+0.0)	49.89 (+0.0)	26.71 (+0.0)	46.08 (+0.0)	50.96 (+0.0)	31.37 (+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)
		MORE	63.64 (+3.6)	53.52 (-4.72)	31.99 (+1.25)	46.32 (+3.53)	43.09 (-6.8)	25.33 (-1.38)	51.6 (+5.52)	44.52 (-6.44)	27.96 (-3.41)
		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)
		MORE	68.34 (+8.3)	69.38 (+11.14)	80.14 (+49.4)	55.2 (+12.41)	62.11 (+12.22)	54.48 (+27.77)	54.89 (+8.81)	58.74 (+7.78)	58.86 (+27.49)
		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)
		MORE	66.41 (+6.37)	64.09 (+5.85)	75.32 (+44.58)	65.31 (+22.52)	61.82 (+11.93)	74.08 (+47.37)	63.43 (+17.35)	59.87 (+8.91)	79.57 (+48.2)
		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)
		MORE	70.05 (+10.01)	68.25 (+10.01)	63.87 (+33.13)	65.99 (+23.2)	63.94 (+14.05)	68.11 (+41.4)	63.58 (+17.5)	62.06 (+11.1)	71.7 (+40.33)
		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)
		MORE	65.27 (+5.23)	53.73 (-4.51)	26.79 (-3.95)	49.46 (+6.67)	45.24 (-4.65)	25.85 (-0.86)	53.32 (+7.24)	45.21 (-5.75)	28.79 (-2.58)
		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)
		MORE	69.86 (+9.82)	64.54 (+6.3)	25.85 (-4.89)	53.55 (+10.76)	52.38 (+2.49)	25.96 (-0.75)	54.44 (+8.36)	48.2 (-2.76)	27.59 (-3.78)
		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)
		MORE	69.65 (+9.61)	62.65 (+4.41)	26.39 (-4.35)	58.83 (+16.04)	56.51 (+6.62)	27.83 (+1.12)	54.98 (+8.9)	48.37 (-2.59)	30.62 (-0.75)
		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)
		MORE	70.2 (+10.16)	63.0 (+4.76)	28.05 (-2.69)	58.7 (+15.91)	54.76 (+4.87)	27.99 (+1.28)	55.14 (+9.06)	48.56 (-2.4)	31.48 (+0.11)
		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 5. MEGACOIN-Bench Results (%). Small-scale MEGACOIN-Instruct-finetuned opensource VLMs can outperform GPT-4o. (.) shows the accuracy gain/gap to GPT-4o.

even leads to performance drops, even on GPT-4o, indicating that the model may struggle to identify the target object in the image. This reveals the design of our benchmark introduces specific challenges to the models.

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

	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**
+C10&TI	66.41	**1538.36**	276.79	**80.22**	**63.49**
+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
+C10&TI	76.03	1646.26	**337.86**	**82.65**	64.01
+220k	**78.44**	1630.19	**352.14**	**82.65**	**64.47**

Table 6. Performance of VLMs on common benchmarks.

(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.

6.3.2 Common Benchmark Results

Tab. 6 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. 6.2, we also explore fine-tuning the VLMs with combined original SFT data and the entire MEGACOIN (dubbed
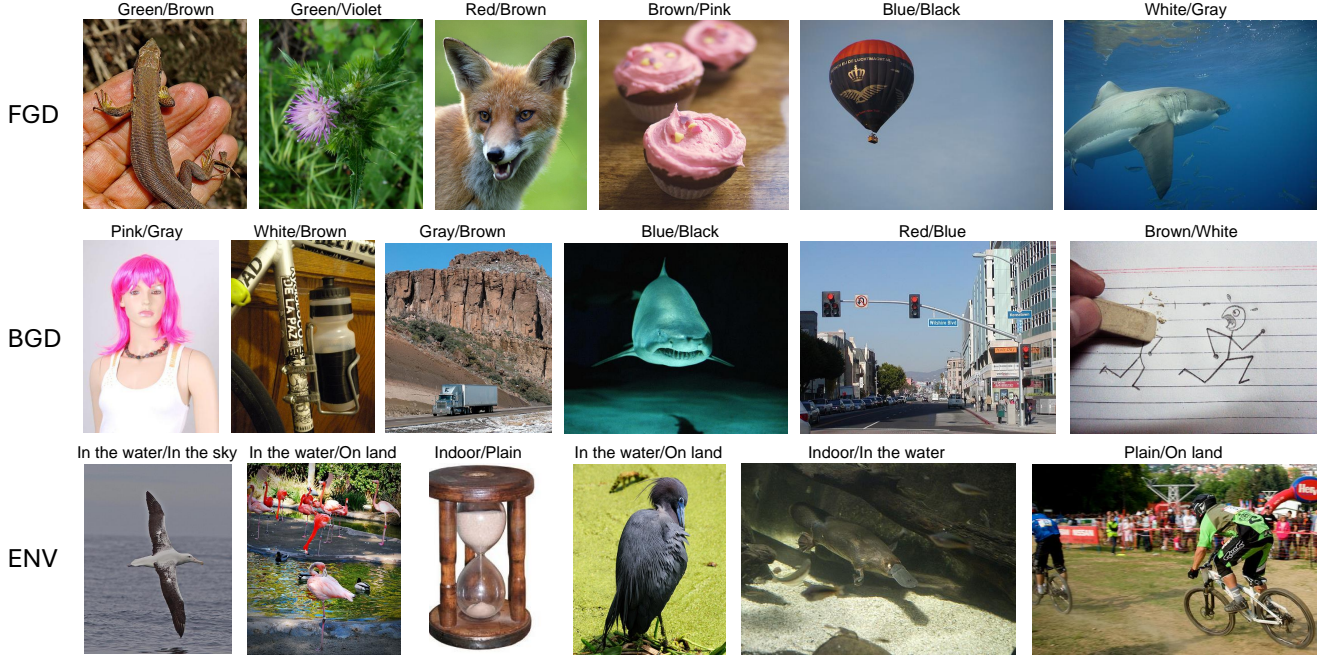
Figure 2. Failure cases on MEGACOIN-Bench before/after training with MEGACOIN-Instruct. After fine-tuning with MEGACOIN-Instruct, we are able to have the VLMs recognize the correct colors and environments.

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 basic 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.

7. 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. 2. 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 in the Appendix for completeness.

Setup We adopt the implementation of DomainBed [10] using the CLIP representation [34] and do 10 runs each with different random seeds of seven different DG algorithms, which are ERM, GroupDRO [35], CORAL[37], MMD[20], EQRM[7], SelfReg[17] and VREx[18] based upon their performances in previous studies [10, 31] and recency.

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.

7.1. Result

MMD is the best on generalization, followed by CORAL and ERM. The generalization results across all datasets (Tab. 8) 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. Group-DRO 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. 9 presents the WGA that are evaluated on the same set of subgroups as training. We surprisingly observe that the *subgroup robust* method 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%.

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-C10-FGD-*merge*	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-TI-FGD-*merge*	12	200	100000	10000	n/a	25113 (22.8%)	5383 (4.9%)	x	x	x
MEGACOIN-C10-BGD	12	10	50000	10000	n/a	15599 (26.0%)	329 (0.5%)	x	x	x
MEGACOIN-C10-BGD-*merge*	12	10	50000	10000	n/a	15599 (26.0%)	759 (1.3%)	x	x	x
MEGACOIN-TI-BGD	12	200	100000	10000	n/a	33923 (30.8%)	638 (0.6%)	x	x	x
MEGACOIN-TI-BGD-*merge*	12	200	100000	10000	n/a	33923 (30.8%)	2552 (2.3%)	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 7. 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. 2) 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 8. 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	Average
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 9. Worst group accuracy (%) when using the same training and evaluation subgroups.

	Trained × Evaluated	ERM	GroupDRO	CORAL	EQRM	MMD	SelfReg	VREx	Average
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	95.83	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 10. 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.

ERM excels at attribute generalization. Tab. 10 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, fol-

lowed by CORAL and EQRM, contrary to MMD leading the way in the previous evaluations.

8. Conclusion

In this work, we constructed MEGACOIN, a large-scale human-annotated dataset containing 660k medium-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 current VLMs have room to improve on basic visual understanding, and we investigated 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. We hope that our work can help 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 [12] for both VLM alignment and benchmarking algorithms focusing on out-of-distribution generalization.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 3

[4] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020. 3

[5] Ming-Chang Chiu, Yingfei Wang, Derrick Eui Gyu Kim, Pin-Yu Chen, and Xuezhe Ma. On human visual contrast sensitivity and machine vision robustness: A comparative study. *arXiv preprint arXiv:2212.08650*, 2022. 1, 3

[6] Ming-Chang Chiu, Pin-Yu Chen, and Xuezhe Ma. Better may not be fairer: A study on subgroup discrepancy in image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4956–4966, 2023. 1

[7] Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022. 7

[8] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 3

[9] S Furness, J Connor, Elizabeth Robinson, R Norton, S Ameratunga, and Rodger Jackson. Car colour and risk of car crash injury: population based case control study. *Bmj*, 327 (7429):1455–1456, 2003. 3

[10] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization, 2020. 7, 1

[11] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024. 5

[12] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 8

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[14] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. 2, 3

[15] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[16] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 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*, pages 235–251. Springer, 2016. 2

[17] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021. 7

[18] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021. 7

[19] Shanghai AI Laboratory. Sharegpt-4o, 2024. 2

[20] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 2, 7

[21] Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*, 2022. 3

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[23] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2

[24] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023. 2, 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 5

[27] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He,

Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 3

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 3

[29] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2

[30] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. 3

[31] Aengus Lynch, Gbètondji JS Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023. 3, 7

[32] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 2

[33] Rui Qiao and Bryan Kian Hsiang Low. Understanding domain generalization: A noise robustness perspective. *arXiv preprint arXiv:2401.14846*, 2024. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 7

[35] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 3, 7

[36] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2

[37] Baochen Sun and Kate Saenko. 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*, pages 443–450. Springer, 2016. 2, 7

[38] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991. 2

[39] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift, 2023. 3

[40] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 3

[41] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16036–16047, 2023. 3

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

Supplementary Material

9. Domain Generalization Results on Resnet18

We provide domain generalization results on MEGA-COIN using Resnet18 in Tab. 13, 11 and 12. Different from Sec. 7 testing methods in a linear probing setting, we benchmark using Resnet18.

Experimental Details. Similar to Sec. 7, we follow the implementation of DomainBed [10]. On CIFAR10, we train Resnet18 from scratch, and on Tiny-Imagenet, we use Resnet18 with weights pre-trained on ImageNet. [2]

Experiment Configuration For our experiments, we adopted specific hyperparameter settings optimized for the datasets used. Below is a summary of the configurations:

- **Learning Rate:** 1×10^{-3}
- **Batch Size:** 128
- **Weight Decay:** 1×10^{-4}
- **Learning Rate Schedulers:**
  - **CIFAR-10**: Cosine Annealing LR Scheduler with a step size of 20,000, and minimum learning rate of 0
  - **TinyImageNet**: Step LR Scheduler with a step size of 20,000, and gamma of 0.1
- **Training Steps:** 60,000 steps in total
- **Validation Split:** 5% for holdout fraction

GroupDRO is the best on generalization. GroupDRO is the best overall (Tab. 13) and shows noticeable superiority over other algorithms.

ERM is the best with subpopulation shift, followed by GroupDRO. Tab. 11 presents the WGA when evaluated on the same set of subgroups as training. We observe that ERM and GroupDRO are on par with each other. *Subgroup robust* methods like GroupDRO is a top performer, unlike in the linear probing setting.

GroupDRO excels again at attribute generalization. Tab. 12 shows that GroupDRO is consistently better than others in the attribute generalization setting.

---

[2]To ensure fair comparison, we adopt the DomainBed implementations, and we verified that on Resnet18, regular training recipe can reach 95% of accuracy.

Trained×Evaluated	ERM	GroupDRO	ARM	CORAL	EQRM	SelfReg	VREx	Average
C10FGD×FGD	85.00	89.33	84.67	85.71	85.71	84.76	88.05	86.18
C10BGD×BGD	80.00	81.33	80.43	79.67	84.67	82.80	82.26	81.59
C10ENV×ENV	81.20	79.66	80.03	79.40	81.97	81.62	81.66	80.79
TIFGD×FGD	59.00	58.17	57.26	54.80	56.78	53.36	54.42	56.26
TIBGD×BGD	63.06	62.41	62.90	55.66	59.84	55.18	54.30	59.05
TIENV×ENV	62.16	59.39	62.62	56.62	60.68	55.82	57.09	59.20
Average	**71.74**	<u>71.72</u>	71.32	68.64	71.61	68.92	69.63	70.51

Table 11. Worst group accuracy (%) when using the same training and evaluation subgroups. ERM and GroupDRO are on par with each other.

Trained × Evaluated		ERM	GroupDRO	ARM	CORAL	EQRM	SelfReg	VREx	Average
FGD	C10FGD×BGD	81.29	82.67	79.33	79.67	82.67	80.00	81.18	80.97
	C10FGD×ENV	82.96	85.70	81.94	82.88	85.19	82.99	83.40	83.58
	TIFGD×BGD	58.71	59.68	58.87	54.22	55.66	50.36	53.23	55.82
	TIFGD×ENV	60.84	61.44	60.60	55.16	57.70	54.82	55.45	58.00
	Average	67.98	**72.37**	70.31	70.95	70.19	67.04	68.32	69.59
BGD	C10BGD×FGD	85.33	86.67	85.71	87.47	86.67	87.33	85.71	86.41
	C10BGD×ENV	82.59	86.21	82.51	83.76	85.07	83.79	85.04	84.14
	TIBGD×FGD	59.00	60.93	59.18	53.13	57.84	53.36	54.22	56.81
	TIBGD×ENV	62.21	62.99	61.71	54.89	59.37	55.96	56.03	59.02
	Average	72.20	**74.20**	72.28	69.81	72.24	70.11	70.25	71.60
ENV	C10ENV×FGD	82.67	80.95	81.00	77.14	82.18	82.38	85.42	81.68
	C10ENV×BGD	77.00	80.00	75.00	75.33	79.67	79.67	80.38	78.15
	TIENV×FGD	59.51	60.37	59.81	53.61	59.48	51.54	55.19	57.07
	TIENV×BGD	60.16	59.52	62.89	55.48	59.35	54.83	53.49	57.96
	Average	69.84	**70.21**	69.68	65.39	70.17	67.11	68.62	68.72

Table 12. Worst group performance (%) in attribute generalization (AG) setting. GroupDRO is the most effective in this setup.

10. Evaluation Cost of GPT-4o

The evaluation of MEGACOIN-Bench on GPT-4o costs $560.

11. Dataset Release

We will release MEGACOIN annotations upon the acceptance of this paper to facilitate research in VLM SFT and domain generalization.

Method	C10FGD	C10FGDmerge	C10BGD	C10BGDmerge	C10ENV	TIFGD	TIFGDmerge	TIBGD	TIBGDmerge	TIENV	Average
ERM	89.39(0.27)	89.94(0.26)	89.48(0.22)	90.31(0.19)	87.67(0.29)	66.38(0.32)	67.04(0.31)	67.45(0.32)	67.64(0.43)	66.98(0.41)	78.23(0.30)
GroupDRO	91.54(0.29)	91.65(0.13)	91.24(0.27)	91.31(0.30)	88.90(0.27)	67.13(0.32)	67.10(0.45)	68.18(0.37)	68.21(0.66)	67.42(0.28)	**79.27**(0.33)
ARM	88.83(0.39)	89.23(0.15)	89.06(0.24)	89.81(0.30)	87.01(0.24)	66.23(0.48)	66.53(0.38)	67.29(0.74)	67.54(0.51)	67.09(0.36)	77.86(0.38)
CORAL	89.95(0.29)	90.56(0.24)	89.60(0.24)	90.64(0.08)	85.46(0.47)	60.56(4.99)	62.86(4.13)	59.51(6.41)	62.86(4.89)	59.43(6.22)	75.14(2.80)
EQRM	90.39(0.56)	90.70(0.33)	90.11(0.56)	90.78(0.32)	88.34(0.42)	63.09(2.19)	64.62(2.11)	64.59(2.01)	65.43(2.25)	66.08(2.27)	77.41(1.30)
SelfReg	89.86(0.25)	90.26(0.20)	89.96(0.13)	90.49(0.14)	88.19(0.10)	59.61(4.05)	61.30(2.73)	60.56(3.36)	62.72(2.24)	58.38(4.25)	75.13(1.75)
VREx	90.38(0.75)	90.99(0.66)	90.57(1.07)	90.83(0.51)	89.25(0.84)	61.22(3.53)	61.27(3.84)	61.64(3.60)	62.35(3.47)	62.45(2.62)	76.10(2.09)
Average	90.05(0.40)	90.48(0.28)	90.00(0.39)	90.60(0.26)	87.83(0.38)	63.46(2.27)	64.39(1.99)	64.17(2.40)	65.25(2.06)	63.98(2.34)	77.02(1.28)

Table 13. Domain generalization algorithms results (%, std) on MEGACOIN (Resnet18). We observe that GroupDRO is the overall best.