Toward a Holistic Evaluation of Robustness in CLIP Models

Weijie Tu, Weijian Deng, and Tom Gedeon, *Senior Member, IEEE*

***Abstract*—Contrastive Language-Image Pre-training (CLIP) models have shown significant potential, particularly in zero-shot classification across diverse distribution shifts. Building on existing evaluations of overall classification robustness, this work aims to provide a more comprehensive assessment of CLIP by introducing several new perspectives. First, we investigate their robustness to variations in specific visual factors. Second, we assess two critical safety objectives—confidence uncertainty and out-of-distribution detection—beyond mere classification accuracy. Third, we evaluate the finesse with which CLIP models bridge the image and text modalities. Fourth, we extend our examination to 3D awareness in CLIP models, moving beyond traditional 2D image understanding. Finally, we explore the interaction between vision and language encoders within modern large multimodal models (LMMs) that utilize CLIP as the visual backbone, focusing on how this interaction impacts classification robustness. In each aspect, we consider the impact of six factors on CLIP models: model architecture, training distribution, training set size, fine-tuning, contrastive loss, and test-time prompts. Our study uncovers several previously unknown insights into CLIP. For instance, the architecture of the visual encoder in CLIP plays a significant role in their robustness against 3D corruption. CLIP models tend to exhibit a bias towards shape when making predictions. Moreover, this bias tends to diminish after fine-tuning on ImageNet. Vision-language models like LLaVA, leveraging the CLIP vision encoder, could exhibit benefits in classification performance for challenging categories over CLIP alone. Our findings are poised to offer valuable guidance for enhancing the robustness and reliability of CLIP models.**

***Index Terms*—Contrastive language-image pre-training (CLIP), robustness, evaluation.**

I. Introduction

LEVERAGING contrastive training to cohesively align images and text within a singular embedding domain, the CLIP model excels in delivering versatile zero-shot generalizations. This inherent proficiency enables CLIP to handle diverse tasks without the need for task-specific fine-tuning [1], [2]. Remarkably, CLIP models exhibit outstanding zero-shot classification capabilities, even without explicit training on the target dataset. Moreover, they demonstrate commendable robustness against challenging natural distributional shifts [3], [4], [5], [6], [7]. Gaining a deeper understanding of such behaviors in CLIP models is crucial for steering the future image-text foundational models. Contemporary research has delved into multiple facets of CLIP models. This encompasses areas such as dataset formulation [8], reproducibility in scaling laws [9], strategies for fine-tuning [10], adversarial classification robustness [11] and nuances of the training distribution [12], [13].

Motivated by previous work, we conduct an in-depth analysis of CLIP models, expanding our perspective beyond overall classification robustness. Our analysis includes several key dimensions: (1) robustness to visual factors, where we assess whether CLIP models can maintain performance when encountering variations such as pose, size, color, lighting, and occlusions; (2) out-of-distribution (OOD) detection, evaluating the models' ability to identify instances with labels not present in the training distribution; (3) predictive uncertainty, examining whether CLIP models provide calibrated predictions that accurately reflect uncertainty under different testing conditions; (4) zero-shot retrieval, assessing the models' capability to associate novel textual queries with relevant visual content; (5) 3D awareness, evaluating how well CLIP models handle 3D corruptions and maintain multi-view consistency; and (6) interaction between the vision and language encoders, investigating how these components influence classification robustness. Within each of these dimensions, we analyze the impact of several crucial factors on CLIP's behavior, including variations in training distribution, model architectures, dataset sizes, contrastive loss, fine-tuning, test-time prompts, and dataset curation. This comprehensive analysis provides a thorough assessment of both the strengths and limitations of CLIP models across these critical areas.

To this end, we evaluate 84 zero-shot CLIP models with varying visual encoder architectures, training sources, and dataset sizes, as well as 44 ImageNet fine-tuned CLIP models. To establish a baseline, we compare these models against 127 ImageNet models without language-image pre-training. We examine 10 visual factors variations present in the ImageNet validation set [14], including object pose, lighting, and background, to assess models' visual factors-level robustness. As for OOD detection, we employ ImageNet as an in-distribution (ID) set following [15] and test on 5 types of OOD scenarios. Then, to investigate the predictive uncertainty, we use a set of canonical ImageNet distributions, such as texture, style, and perturbation shifts. We evaluate the effectiveness of data curation methods on the aforementioned datasets. Furthermore, we measure the

Received 30 September 2024; revised 25 April 2025; accepted 1 June 2025. Date of publication 16 June 2025; date of current version 6 August 2025. Recommended for acceptance by T. Liu. *(Corresponding author: Weijian Deng.)*

Weijie Tu and Weijian Deng are with the School of Computing, Australian National University, Canberra, ACT 0200, Australia (e-mail: weijie.tu@anu.edu.au; weijian.deng@anu.edu.au).

Tom Gedeon is with the Australian National University, University of Óbuda, 1034 Budapest, Hungary, and also with Curtin University, Bentley, WA 6102, Australia (e-mail: tom.gedeon@anu.edu.au).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2025.3580234, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3580234

TABLE I
SUMMARY OF EVALUATION DIMENSIONS AND KEY FINDINGS

	Evaluation Dimension	What We Evaluated	Key Findings
I. Task-Level	Visual Factor Robustness (Sec. IV)	Performance under visual variations (*e.g.*, shape, texture, and size)	CLIPs outperform ImageNet models on six of ten factors; Training distribution impacts visual factor robustness of CLIP.
	OOD Detection (Sec. V)	Novelty detection ability (*e.g.*, NINCO)	Zero-shot CLIP is competitive with other models; Training distribution impacts detection accuracy.
	Calibration (Sec. VI)	Alignment between prediction confidence and correctness	Fine-tuning raises calibration error; LP-FT and WiSE-FT recover with temperature scaling, while FLYP remains overconfident. Both data distribution and quantity play key roles.
	Retrieval (Sec. VII)	Image–text matching accuracy	CLIP's Zero-shot retrieval aligns with classification accuracy; data distribution and augmentation shape retrieval quality.
	3D Awareness (Sec. VIII)	Robustness to 3D corruptions and correspondence matching	CNN-based CLIPs are more stable than ViT-based CLIP.
II. Model-Level	Vision–Language Interaction (Sec. IX)	CLIP vs. LLaVA on hard category splits	LLaVA outperforms CLIP on ambiguous sets; stronger LLMs (Vicuna $>$ Mistral) amplify gains.
	Training Paradigm Impact (Sec. X-C)	CLIP, BLIP, SigLIP, ViTamin	No paradigm dominates. ViTamin is more robust to 3D corruptions; SigLIP improves robustness but weak on calibration.
	Prompt Sensitivity (Sec. X-A)	Prompt set size and quality	LLM prompts improve accuracy but may not benefit OOD detection or calibration, and no method consistently dominates.
	Fine-Tuning Impact (Sec. X-B)	Various of fine-tuning techniques	LP-FT and WiSE-FT are well-balanced. FLYP boosts accuracy but hurts calibration. PromptSRC preserves both.
III. Data-level	Dataset Curation Effect (Sec. X-D)	Filtering and data diversity	Data curation boosts classification, OOD, retrieval, and 3D robustness for ViT-CLIP, but not calibration.

Our evaluation covers three levels of analysis: task-level, model-level, and data-level, each addressing distinct aspects of model behavior and robustness.

3D awareness of CLIP by geometric, semantic correspondence estimation as in [16] and robustness against 3D-related corruptions, such as near focus and motion blur [17]. Lastly, to explore the interplay between the visual and text encoders of CLIP, we compare CLIP models with LLaVA [18] in terms of classification performance on the challenging diffusion model-generated ImageNet-D [19].

This article extends our previous conference paper [20], with the following major additions: **(1)** The experiment scale has been expanded by including 25 recent zero-shot CLIP models trained on different subsets of DATACOMP [21], allowing us to broaden the findings to the medium-to-low accuracy regime of CLIP models. **(2)** An in-depth analysis is provided to uncover the impact of fine-tuning objectives on the shape-bias of CLIP models (Section IV-B). **(3)** The zero-shot retrieval capability of CLIP models is explored, highlighting the significance of training distribution as a key factor affecting performance trends (Section VII). **(4)** A comprehensive study of fine-tuning methods, including parameter-efficient, standard, and contrastive fine-tuning, is presented (Section X-B). **(5)** A new OOD benchmark, NINOC, is added in our evaluation, which is ID-free and aggregates OOD classes from multiple existing datasets (Section V). **(6)** The 3D-awareness of CLIP models is evaluated by testing their performance on 3D correspondence estimation and robustness against 3D corruptions (Section VIII). **(7)** The interaction between visual and language encoders is investigated from a classification perspective (Section IX). **(8)** We extend the evaluation of dataset curation techniques to robustness-related tasks, including out-of-distribution (OOD) detection, calibration, visual factor-level robustness, and 3D corruption (Section X-D).

We summarize our evaluation dimensions, covering the task, model, and data levels, along with key findings in Table I.

II. RELATED WORK

Robustness: Machine learning models should generalize from training distribution to novel testing environments [22], [23], [24], [25], [26], [27]. One line of work has developed a theoretical framework to investigate model robustness [28]. Ben-David et al. [28] were the first to propose a generalization bound based on the VC dimension, which quantifies the difference in classifier error between source and target distributions using a divergence measure. Mansour et al. [29] later expanded this analysis to accommodate more general loss functions, offering improved generalization bounds through Rademacher complexity. To investigate such capability of deep models to various forms of test distributions, a commonly used approach is to introduce artificial transformations onto images, such as style transfer [30], corruptions and perturbations [31], [32]. Moreover, many real-world datasets are introduced to assess model robustness under different natural distributional shifts [3], [4], [5], [6], [7], [33]. For instance, Idrissi et al. [14] proposes ImageNet-X by relabelling the ImageNet validation set to provide detailed labels for naturally occurring factors such as pose, background, and lighting. [19] introduces 3DCC to study the robustness of networks to 3D corruptions.

CLIP Analysis: Existing studies have explored various aspects of CLIP models, including dataset formulation [8], reproducibility in scaling laws [9], adversarial classification robustness [11], fine-tuning strategies [10], nuances of the training

distribution [12], visual prompt [34], typographic attacks [35] and techniques for dataset curation [21].

Our comprehensive evaluation of CLIP goes beyond overall classification robustness to include assessments of visual-factor robustness and 3D corruption robustness. We also explore additional perspectives that are crucial for real-world applications, such as out-of-distribution (OOD) detection, which aims to filter out inputs that are irrelevant to the task at hand. Furthermore, we examine prediction uncertainty to determine whether the model can classify images with calibrated prediction probabilities that align with the empirical frequency of correctness [36], [37]. Additionally, we incorporate zero-shot retrieval tasks [9] and 3D geometry correspondence matching to investigate the potential of CLIP features.

III. Experimental Setup

A. Models of Interest

Contrastive language-image pre-training models: we use **84 zero-shot CLIP models (CLIP)** and **44 ImageNet fine-tuned CLIP models (CLIP-FT)**. They have different visual encoders, including slightly modified ResNet [38], ConvNeXt [39], ViT [40] and EVA [41]. There are various training sources, including LAION [42], WIT [1] and Conceptual Captions [43], and multiple sizes of training datasets from 3 million to 2 billion. Note that in this extended paper, we include 25 recent zero-shot CLIP models. They are trained on subsets of CommonPool [21], ranging from 14 million, 140 million to 1 billion. CommonPool draws its data from the same source as LAION, which is Common Crawl. These models allow us to validate and expand our findings in a medium-to-low accuracy regime. We also assess the performance of very recent CLIP models which are trained on filtered high-quality pre-training datasets using dataset curation techniques [44], [45]. To compare the performance with LLaVA [18], we also include SigLIP [46].

For the CLIP-FT models, the vision encoder of CLIP is fine-tuned on ImageNet-1 K. We consider different fine-tuning algorithms, including directly fine-tuned on ImageNet-1K [47], first fine-tuned on ImageNet-12 K, a subset of ImageNet-22 K before fine-tuning on ImageNet-1 K, and also fine-tuned by parameter-efficient fine-tuning methods [48], [49]. We use the default prompt template provided by [1] for zero-shot CLIP models unless specified.

Models compared: we use 127 ImageNet models with various architectures, including Convolutional Neural Networks (e.g., ResNet [38] and ConvNeXt [39]), Vision Transformers (e.g., ViT [40] and Swin [50]) and all-MLP architectures [51], [52] (e.g., MLP-Mixer [52]). Following [53], we divide them into three categories: **(i) Standard Models.** This group consists of models supervised on the ImageNet training set. **(ii) Contrastive learning models.** This category contains 8 models pre-trained by contrastive learning. There are 6 training algorithms investigated, including InsDis [54], MoCo [55], SimCLR [56]; **(iii) Pre-trained on more data.** This group contains models pre-trained on a significantly larger dataset (e.g., ImageNet-21 K) than the ImageNet training set. All the above models, including CLIP, are publicly available on TIMM [57], OpenCLIP [58].

Modern vision language models: This paper considers LLaVA [18], which combines a frozen CLIP vision encoder and a large language model (e.g., Vicuna) for general-purpose visual and language understanding. In our study, we consider six LLaVA models: the visual encoders used are CLIP-L/14-336 and SigLIP, paired with three large language models: Mistral-instruct-V2 [59], Llama-Chat [60], and Vicuna-V2-7B [60], resulting in a total of six LLaVA models. These models are available on HuggingFace, as provided by [61].

B. Test Sets and Metrics

I. Robustness: We first pinpoint model failure patterns by testing on ImageNet-X [14], which is a relabeling of ImageNet validation by 16 naturally occurring factors. This work mainly considers 10 factors labelled with a sufficient number of test samples: *Pose*, *Background*, *Pattern*, *Color*, *Smaller*, *Shape*, *Partial View*, *Subcategory*, *Texture* and *Larger*. The metric is accuracy, and high is better. In addition, we include cue-conflict stimuli and Stylized-ImageNet [30] to measure the model bias towards the shape or texture.

II. OOD detection: We use a large-scale OOD detection benchmark which is built up on ImageNet: in-distribution (ID) ImageNet *v.s.* {iNaturalist [62], SUN [63], PLACES [64], TEXTURE [65], and ImageNet-O [7] (OOD). Metrics are the area under the receiver operating characteristic curve (AUROC) and the higher is better; false positive rate (FPR@95) when the true positive rate is at 95% and a lower score is better. To evaluate OOD detection across diverse conditions, we employ the NINCO dataset [66], which is ID-contamination-free and comprises OOD classes from various existing OOD datasets. We report mean AUROC and FPR@95.

III. Calibration: We study ID and OOD datasets, where ImageNet validation is ID dataset and OOD datasets are: ImageNet-V2 [3], ImageNet-Rendition [5], ImageNet-Adversarial [7], ImageNet-Sketch [4], ObjectNet [6] and ImageNet-Vid-Robust [67]. Metrics are estimated calibration error (ECE) [68] and negative log-likelihood (NLL). A lower ECE or NLL indicates better calibration.

IV. Retrieval: We evaluate zero-shot retrieval performance on Flick30K [69] and MSCOCO [70] following the evaluation setup and splits from [71]. As in [1], we compute the cosine similarity between image and text embeddings as the image-text scores. When evaluating image retrieval, we rank the top-K images for each text caption, and vice versa for text retrieval. Recall@K is the metric with $K = 5$.

V. 3D Awareness: Two tasks are explored for this property: correspondence estimation and robustness against 3D corruptions. We use ScanNet [72], NAVI [73] and SPair-71K [74] as the evaluation datasets for correspondence estimation. The metric is recall. For robustness against 3D corruptions, we use 3DCC [19], which applies 3D-related corruptions against ImageNet-validation with 5 severity levels. The performance is measured by accuracy.

VI. Comparison to LLaVA: We compare the performance of CLIP and LLaVA on ImageNet-D [19], which consists of three splits, *Background*, *Texture* and *Material*. CLIP is evaluated

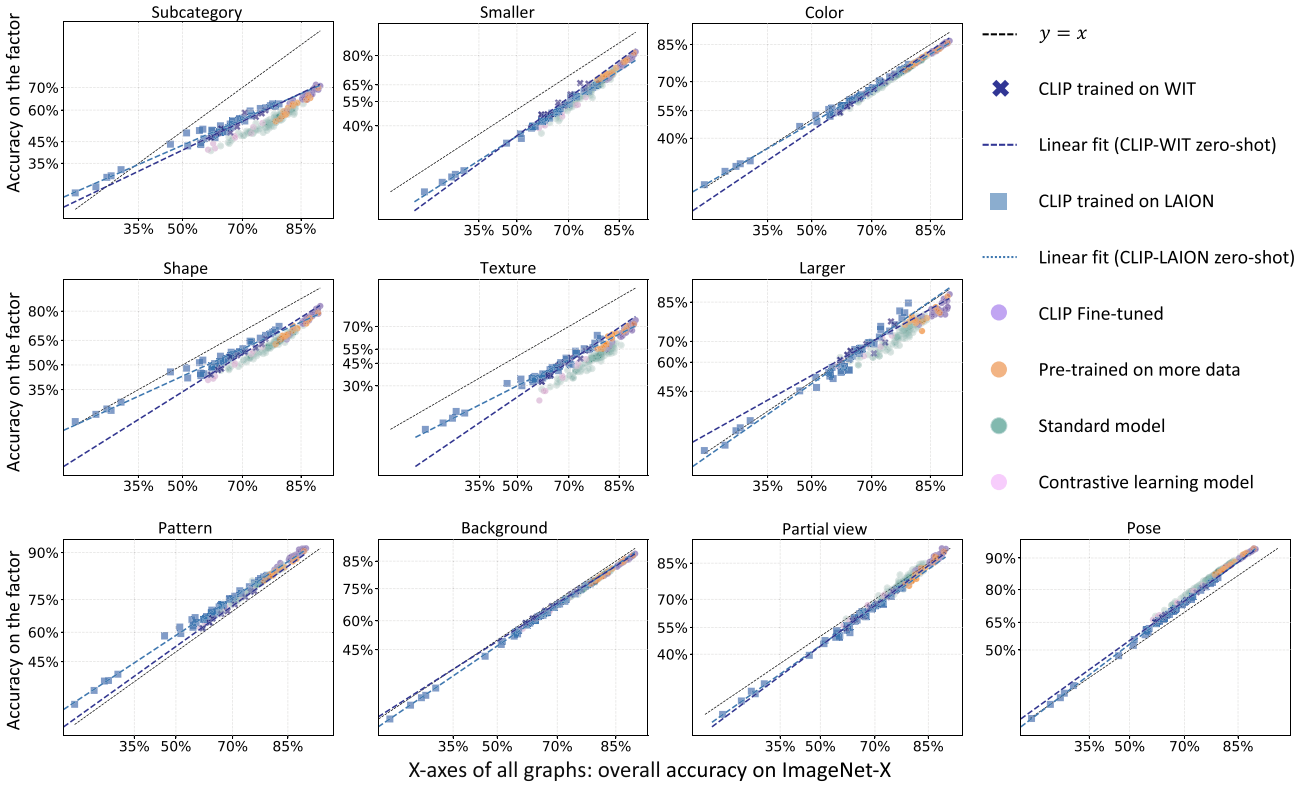


Fig. 1. The models' performance on the subset of ImageNet-X annotated with a given visual factor (y-axis) to their overall accuracy on the whole ImageNet-X (x-axis). Each point represents a model. The x-axis and y-axis are probit transformed following [53]. The black dashed line represents the ideal robust models whose performance on each visual factor is the same as the overall performance. The blue straight lines are fit with robust linear regression [75]. We include models supervised on ImageNet-1 K, pre-trained on more data, contrastive learning models, CLIP models trained on two data distributions, and their fine-tuned counterparts.

using standard zero-shot image classification protocol, while LLaVA is assessed by standard visual question answering protocol. They are both required to classify images from four classes and use accuracy as the measurement.

C. Analytical Methodology

Key Factors: to understand the underlying factors that influence the performance of CLIP models, we delve into six primary aspects: 1) training distribution, evaluating the effect of data source; 2) model architecture, looking into the potential effects of different structural choices on model performance; 3) dataset quantity, probing the interplay between the amount of data available for training and the model's efficiency; 4) contrastive loss, understanding its specific role in training dynamics 5) fine-tuning, 6) test-time prompt, assessing the impact of prompts during the evaluation on model outputs.

We follow the analytical methodology of seminal work [53], along with subsequent studies such as [8], [12], [76], to study the influential factor. Within the performance trends observed across all models, any factor causing a deviation from these trends is influential. Notably, in our research, we mainly emphasize and discuss such influential factors within each facet of our investigation. In Table I, we organize our evaluation into task-level, model-level, and data-level dimensions, highlighting the main insights observed in each.

IV. Visual Factor-Level Robustness

Our research builds upon previous findings on the robustness of CLIP models and focuses on the potential failure types of the model. Instead of solely measuring overall accuracy across distributions, this section investigates the behavior of CLIP models when faced with varying visual factors such as *Pose*, *Background*, and *Object Scale*.

A. CLIP Models Generally Exhibit Better Factor-Level Robustness Than Other Models

Factor-level effective robustness: In our study, we introduce the concept of visual factor-level effective robustness based on effective robustness [53]. It measures a model's ability to achieve higher accuracy on the subset annotated by a specific visual factor compared to what is expected based on its overall accuracy on ImageNet-X. Fig. 1 displays the accuracy on the subset annotated by a specific visual factor relative to the overall accuracy on ImageNet-X.

(1) CLIP models are generally more robust than other ImageNet models on six out of ten visual factors: Fig. 1 highlights several insights into the factor-level robustness of CLIP models. First, we find that CLIP models are more robust than other models on six out of ten visual factors, including *Subcategory*, *Smaller*, *Color*, *Shape*, *Texture*, and *Larger*. Specifically, CLIP models exhibit higher factor-level effective robustness than other

models on each of these factors. Second, we observe that CLIP models are less robust than other models on *Pose* and *Partial View*. Third, CLIP models show a similar trend to other models on the *Background* factor.

(2) Training distributions lead to different trends in CLIP models: The choice of training distribution impacts the factor-level robustness of CLIP models. Specifically, we find that training on different datasets (i.e., LAION and WIT) forms distinct trends on each visual factor for CLIP, and there is no single training source that always leads to higher factor-level robustness than another. For instance, we observe that CLIP models trained on LAION demonstrate higher robustness on *Shape* factor than those trained on WIT, while this reverses for *Background* and *Pose* factors. The results show a mixed observation on *Large* factor. Furthermore, we further point out that CLIP models trained on different subsets of LAION (LAINON-80 M, LAION-400 M, and LAION-2B) follow the same trend. The above observations highlight the importance of the choice of training source in determining not only the overall accuracy but also the factor-level behaviors of CLIP models. This suggests that factor-level robustness should be considered when choosing the training source.

(3) CLIP fine-tuned models perform slightly better than models pre-trained with more data: We compare CLIP fine-tuned models (CLIP-FT) with other models pre-trained on more data and find that CLIP-FT shows improvement in overall accuracy and robustness on visual factors of *Subcategory*, *Shape*, and *Pattern*. However, no additional robustness gain is observed on other factors. Moreover, CLIP-FT models outperform zero-shot CLIP on variations such as *Pattern* and *Partial View* but perform lower on factors like *Texture* and *Larger*. We speculate that standard fine-tuning introduces spurious correlations [77]. This may lead to a bias for CLIP towards specific visual properties, thereby compromising factor-level robustness on some factors. It would be intriguing to explore fine-tuning techniques to maintain or improve the visual factor-level robustness of CLIP.

(4) Discussion on consistent trends across visual factors: All models exhibit consistent trends across visual factors, despite differences in architecture and training data. Specifically, all models lie below the line $y = x$ under *Smaller*, *Shape*, and *Texture* conditions, which involve changes to object geometry, scale, and surface patterns. While such variations do occur in natural datasets, they are neither explicitly annotated nor emphasized, and thus may be underrepresented in the models' learned feature space. As a result, models tend to rely on statistically dominant but fragile cues—such as canonical shapes, common textures, or typical object sizes—rather than learning representations that are robust to these factors. This behavior is consistent with the concept of shortcut learning [78], where models exploit superficial but predictive patterns that fail under distribution shift. In contrast, performance on Background and Partial View remains stable, likely due to the abundance of such variations in pretraining data, which encourages models to downweight context and develop object-centric representations. The consistency across models suggests these are not architecture-specific artifacts but shared limitations shaped by training data and objectives.

(5) Pre-training analysis of factor-level coverage for training data selection: Given a candidate training set, we aim to estimate its coverage across key visual factors to anticipate potential robustness gaps before model training. ImageNet-X provides annotations across 16 visual factors, which supports pre-training analysis. Specifically, we compute a prototypical feature for each factor by averaging features from its annotated images using a fixed pretrained backbone (e.g., ResNet-50). For each image in the candidate dataset, we extract its feature and assign it to the nearest factor prototype. By counting how many training data are associated with each factor, we estimate the dataset's factor-level coverage. This lightweight and scalable analysis enables factor-aware data selection before training models.

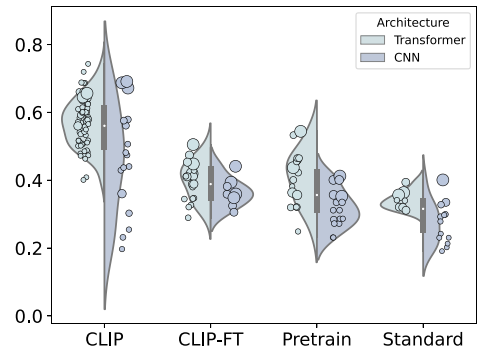


Fig. 2. Shape bias analysis of CLIP, CLIP fine-tuned (CLIP-FT), models pre-trained on more data (Pretrain), and standard models. Large points mean larger models within the group. We observe that CLIP models are more shape-biased.

B. *Texture Bias* V.s. *Shape Bias*

CLIP exhibits a shape bias: We conducted experiments using the cue-conflict stimuli dataset [30] to assess the presence of shape bias in the model's predictions. Shape bias, in this context, refers to the proportion of correct predictions that are based on the object's shape rather than texture or other features. Fig. 2 visualizes the shape bias exhibited by different models, grouped by training methods (zero-shot, CLIP fine-tuning, additional data pre-training, and standard training) and architecture (transformer versus CNN). Our results show that, among the four training methods, CLIP models exhibit a stronger shape bias compared to the other groups. While previous research has indicated that transformers show a greater shape bias than CNNs [80], [81], we found that CLIP models with CNN-based vision encoders also exhibit a significant shape bias. This suggests that CLIP can align more closely with human visual perception, which is widely acknowledged to be shape-driven [30], [82], [83]. In the following, we provide a more detailed analysis of the shape bias observed in CLIP models and explore the implications of these findings.

(1) Model size does not solely explain the shape bias of CLIP: We further observe that larger CLIP models do not necessarily have higher shape bias than smaller-size ones. For example, two models both trained on LAION-80 M, CLIP-ViT/L-14 have 0.54 shape bias, which is 0.09 lower than CLIP-ViT/B-32. This implies that the shape bias of CLIP models cannot be attributed solely to model size. Based on the above, we speculate that the

TABLE II
SHAPE BIAS OF VARIOUS FINE-TUNED CLIP MODELS

Backbone	FT methods	Shape bias
ViT-B/32	Zero shot	0.575
	Fine-tune on 1K	0.401
	Contrastive FT	0.561
	CoOp	0.549
	Tip-Adapter	0.579
ViT-B/16	Zero shot	0.473
	Fine-tune on 1K	0.345
	Contrastive FT	0.448
	CoOp	0.472
	Tip-Adapter	0.487

We include CLIP models fine-tuned using different methods: cross-entropy, contrastive loss [79], and parameter-efficient techniques such as CoOp [48] and Tip-Adapter [49].

TABLE III
THE INFLUENCE OF INPUT RESOLUTION ON SHAPE BIAS WHEN FINE-TUNING CLIP

Source	Backbone	Shape bias	IN-Val	SIN
LAION	ViT/H-14 (336/224)	0.42 /**0.51**	**0.89** /0.88	0.28 /**0.32**
	ViT/L-14 (336/224)	0.41 /**0.47**	**0.88** /**0.88**	0.27 /**0.31**
	ViT/B-16 (384/224)	0.35 /**0.43**	**0.87** /0.86	0.23 /**0.25**
	ViT/B-32 (384/224)	0.33 /**0.45**	**0.85** /0.83	0.21 /**0.22**
	ConvNeXt-B (384/224)	0.31 /**0.38**	**0.87** /0.86	0.17 /**0.21**
WIT	ViT/L-14 (336/224)	0.39 /**0.45**	**0.88** /**0.88**	0.24 /**0.30**
	ViT/B-16 (384/224)	0.35 /**0.41**	**0.87** /0.86	0.22 /**0.23**

We also report accuracy on ImageNet-Val(idation) and Stylized ImageNet (SIN). The higher value in a model pair is in bold. With the same backbone architecture, the CLIP model fine-tuned with a larger input resolution is more accurate on IN-Val but less shape-biased and less accurate on SIN.

shape bias of CLIP may be attributed to its objective, which involves training the model to associate text and image pairs.

(2) Larger input image resolution during fine-tuning of CLIP results in a stronger bias towards texture: In Table III, we observe that an input resolution during fine-tuning impacts shape bias: increasing input resolution during fine-tuning leads to better accuracy on ImageNet validation but also results in more texture-biased models with lower accuracy on Stylized-ImageNet. Across seven pairs of experiments and two training sources, we observe this pattern consistently. Given that input resolution is a crucial model dimension [84], [85], [86], it would be insightful to study its effects on shape bias beyond classification accuracy when devising scaling strategies.

(3) CLIP models tend to texture bias after fine-tuning: Our study reveals that shape bias in CLIP weakens after fine-tuning on ImageNet. Moreover, the fine-tuned CLIP models exhibit a shape bias comparable to models that are pre-trained on larger datasets. This finding is consistent when using a transformer and CNN as the visual encoder. Moreover, these results illustrate that fine-tuning discards the shape-biased property of zero-shot CLIP, which may affect their overall effective robustness [30], [87].

(4) Fine-tuning with contrastive loss maintains shape bias: By default, the CLIP-FT models are trained with standard supervised cross-entropy loss. To decouple the effect of fine-tuning methods and data source, we use zero-shot CLIP with ViT-B/32 and ViT-B/16, and fine-tune them on ImageNet training set by standard cross-entropy, contrastive loss [79], and parameter-efficient fine-tuning methods (CoOp [48] and Tip-Adapter [49]). The shape bias extents are shown in Table II: contrastive fine-tuning on ImageNet maintains the shape bias of CLIP models. This indicates that ImageNet training data might not be the primary cause of the shape-bias decrease. We believe that the alignment mechanism between visual and textual representations may play a fundamental role in shaping this bias. This is supported by our observation that fine-tuning strategies which preserve image–text embedding association tend to retain or strengthen shape bias.

V. OUT-OF-DISTRIBUTION DETECTION

Zero-shot CLIP allows for a flexible definition of in-distribution (ID) classes without re-training the model. Namely, they can conduct zero-shot OOD detection [15]. The current findings suggest that zero-shot CLIP models are competitive with other state-of-the-art models [15], [88]. Based on this finding, we conduct an extensive analysis to determine whether the purported benefits persist across various training sources, subsets, and network architectures. In the experiments, for zero-shot CLIP models, we utilize maximum concept matching [15] to detect OOD data. For models that are trained or fine-tuned on ImageNet-1 K, we employ maximum softmax score [89] for OOD detection.

(1) For CLIP models from the same source, their ID accuracy correlates with OOD detection performance: Our study includes CLIP models trained on two sources (WIT and LAION). Given the same training source, our study, conducted across five challenging OOD scenarios, reveals a strong correlation between the ID accuracy of zero-shot CLIP models and their OOD detection performance (measured by AUROC and FPR@95). This suggests that the zero-shot classification accuracy of CLIP on ID data can serve as a reliable indicator of their OOD detection performance. In contrast, such a trend is not as strong for both standard models and more data-pre-trained models. Furthermore, CLIP-FT models fine-tuned on ImageNet-1 K do not exhibit such a clear correlation.

(2) Training source impacts the trend of CLIP: Upon closer examination of the training distribution, we have observed that the correlation trend between ID accuracy and OOD detection performance is largely dependent on the training source. As illustrated in Fig. 3, our research shows two distinct trends between CLIP models trained on WIT and those trained on LAION. Moreover, with the same ID accuracy, CLIP models trained on WIT exhibit superior OOD detection performance compared to their counterparts trained on LAION on three OOD scenarios. This further indicates the importance of training sources for CLIP.

(3) Fine-tuning procedure significantly influences the OOD detection ability of CLIP: While fine-tuning generally improves CLIP's classification performance, this enhancement does not necessarily translate to better OOD detection accuracy. Some fine-tuned CLIP (CLIP-FT) models perform worse in OOD detection compared to their zero-shot counterparts. Our analysis

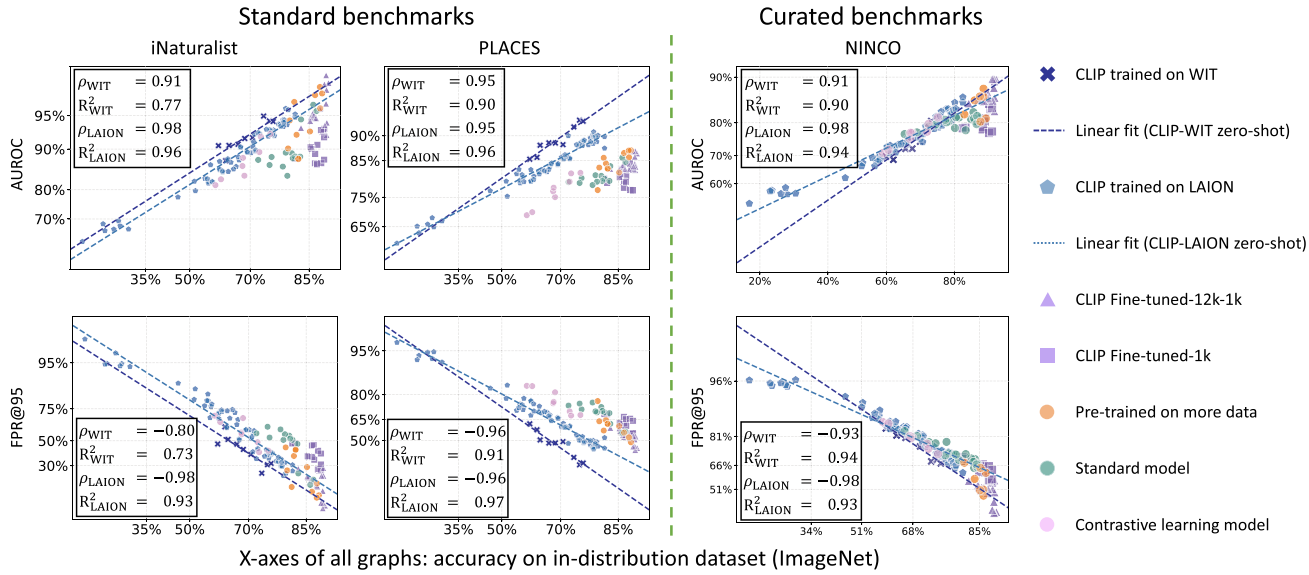


Fig. 3. OOD sample identification capability of models ***vs.* ID dataset classification accuracy.** The OOD detection ability is measured by AUROC (↑) and FPR@95 (↓). Each point represents a model. We plot the results on iNaturalist, PLACES and NINCO. The blue straight lines are fit with robust linear regression [75]. We report spearman's rank correlation and R^2 to quantify the correlation strength between ID accuracy and OOD detection performance for zero-shot CLIP trained on WIT and LAION. The x-axis and y-axis are probit transformed following [53]. We observe that training distribution has a greater impact than training dataset quantity on the OOD detection performance of CLIP. Moreover, after additionally fine-tuning on ImageNet-12 K, CLIP models are generally better at detecting OOD samples than those only fine-tuned on ImageNet-1 K.

distinguishes between two groups of CLIP-FT models based on their fine-tuning procedures: one group is fine-tuned solely on ImageNet-1 K, while the other undergoes additional fine-tuning on ImageNet-12 K. We observe that this additional fine-tuning step has a notable impact on OOD detection performance. As shown in Fig. 3, despite not yielding significant gains in classification accuracy, CLIP-FT models fine-tuned on ImageNet-12 K consistently achieve better OOD detection across all tested scenarios. These findings suggest that the fine-tuning dataset plays a critical role in enhancing OOD detection. Future work should further explore alternative fine-tuning strategies that prioritize OOD detection performance. Additionally, investigating the effects of fine-tuning on datasets beyond ImageNet-1 K/21 K presents an intriguing direction for improving the robustness of CLIP models.

(4) Evaluation on NINCO [66]: To explore the OOD detection across diverse and challenging conditions, we use a new benchmark NINCO for study. It consists of filtered samples from various existing OOD benchmarks. Fig. 3 illustrates the OOD detection performance on NINCO versus ID classification accuracy on the ImageNet validation set. The observations are consistent with those on five standard benchmarks: 1) for CLIP models from the same source, their ID accuracy correlates with OOD detection; 2) training source influences trends of CLIP; 3) additional fine-tuning on ImageNet-12 K helps OOD detection ability of CLIP. ImageNet-21 K offers broader semantic coverage than ImageNet-1 K, which may help bridge the gap between pretraining data (e.g., LAION) and downstream tasks. As an intermediate fine-tuning stage, it could help preserve model generalization, which may explain the improved OOD robustness observed compared to direct fine-tuning on ImageNet-1 K.

VI. Confidence Calibration

To better understand the well-calibrated phenomenon of zero-shot CLIP models reported by [90], this section systematically analyzes the calibration behavior of CLIP models under various training conditions. Specifically, we examine the calibration performance of CLIP models trained on different training distributions, varied training set sizes, and different architectures. Furthermore, we also investigate the calibration performance of CLIP models after fine-tuning.

A. Zero-Shot CLIP Models are Not Consistently More Calibrated Than Other Models

(1) Training data distribution and quantity significantly affect CLIP's calibration: Fig. 4 illustrates the calibration of CLIP models concerning classification accuracy under distribution shifts. We find that models trained on different distributions or dataset sizes do not always group consistently. For example, CLIP models trained on WIT and LAION tend to form distinct clusters. Additionally, within subsets of the LAION dataset, models with similar classification accuracy can display varying levels of calibration. While CLIP models are often praised for superior calibration compared to other models [90], our analysis shows this is not always the case. Notably, CLIP models trained on the LAION-80 M dataset exhibit significantly lower calibration performance compared to standard models. The superior calibration reported by [90] is primarily based on CLIP models trained on WIT. However, when we expand the analysis to models trained on the broader LAION dataset and its subsets, we observe more variability.

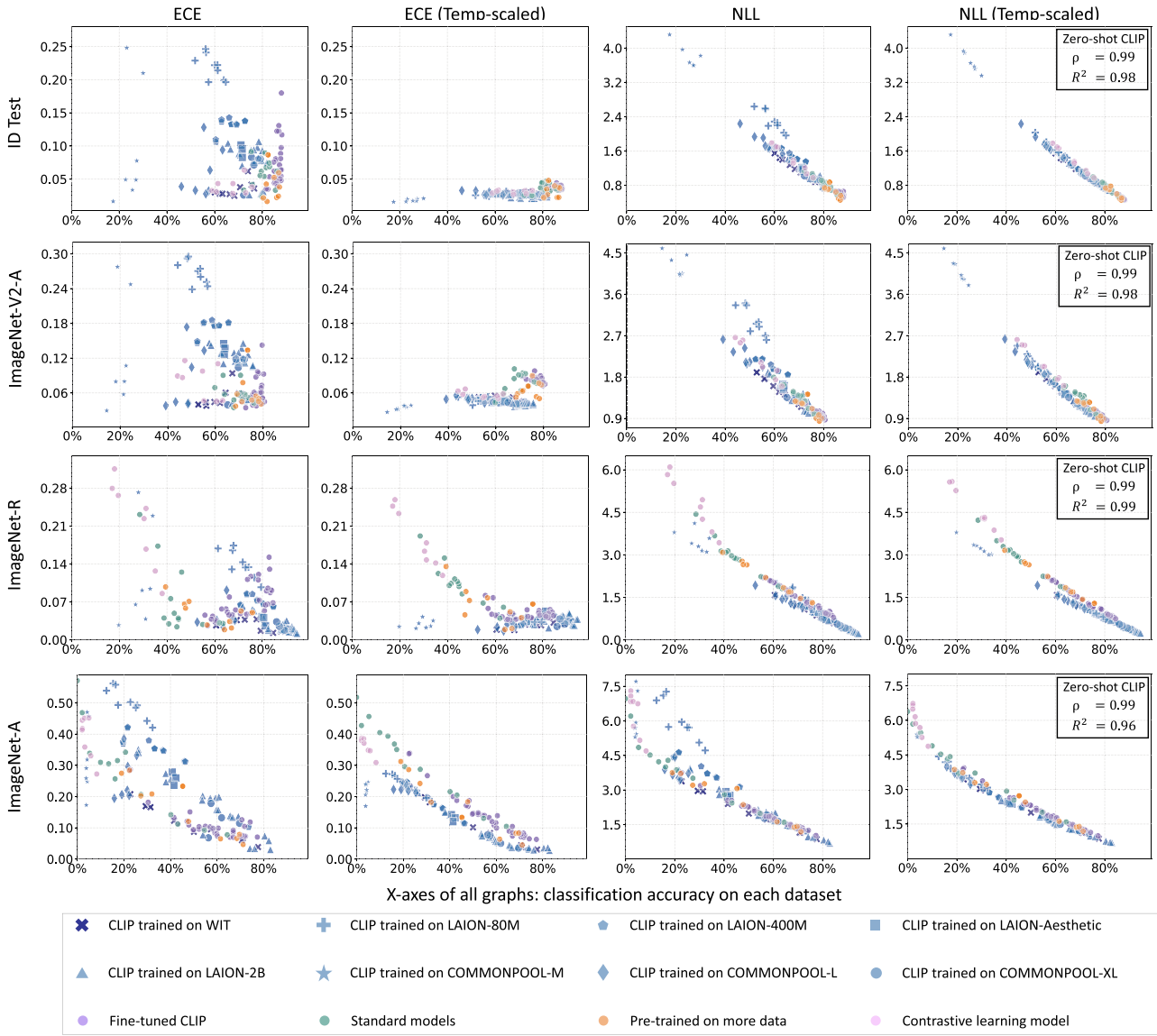


Fig. 4. Model calibration performance with respect to classification accuracy. We report results on in-distribution test set, ImageNet-V2-A, ImageNet-R, and ImageNet-A. Two metrics are considered: ECE (↓) and NLL (↓), we also include calibration performance after calibration with temperature scaling. Each point represents a model. We use colors to represent model groups. For zero-shot CLIP, we additionally use shapes to indicate training distribution and quantity. CLIP models can have higher ECE than standard models. Also, the training distribution and quantity are the key factors influencing the calibration performance of CLIP models. Moreover, temperature scaling reveals a consistent trend in CLIP models. After using temperature scaling for both CLIP and other models, CLIP models follow a distinct trend from others and show better calibration performance.

(2) CLIP fine-tuned models show a trade-off between calibration and classification: As shown in Fig. 4, fine-tuning CLIP models consistently results in higher classification accuracy but increased calibration error across all test sets. Furthermore, we did not observe that further fine-tuning CLIP on ImageNet-12 K benefits calibration performance, which contrasts with its positive impact on OOD detection. Interestingly, other model groups, including those pre-trained on larger datasets, do not show an obvious trade-off between calibration and classification. Additionally, we observe that few fine-tuned CLIP models achieve better calibration than their zero-shot counterparts, even before applying calibration techniques.

B. Temperature Scaling Highlights Well-Calibrated Properties of Zero-Shot CLIP Models

Post-hoc calibration methods, such as temperature scaling [36], are often employed to correct overconfidence or underconfidence in model predictions. Following the protocol in [91], we split the ImageNet validation set into two halves: one for temperature scaling (ID calibration) and the other for testing. We report results on both in-distribution (ID) and out-of-distribution (OOD) test sets.

(1) Classification accuracy of CLIP models correlates with calibration performance after temperature scaling: In Fig. 4, we examine the effects of temperature scaling on both CLIP and non-CLIP models, grouped based on the amount and source

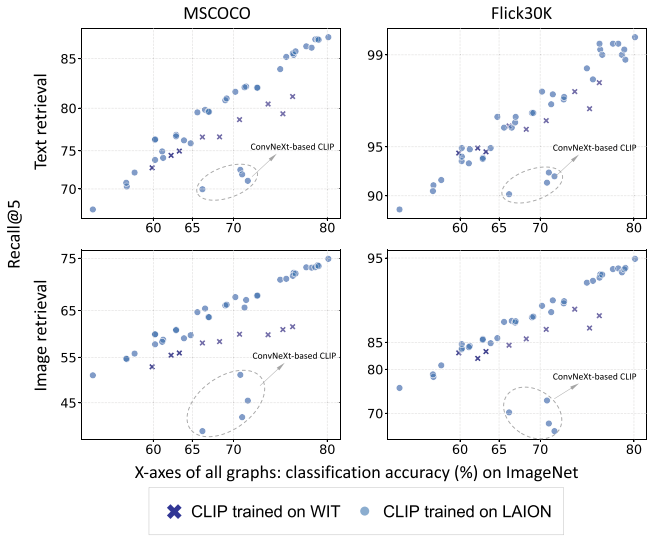


Fig. 5. Image/text zero-shot retrieval v.s classification accuracy on MSCOCO and Flick30 K measured by Recall@5. Classification accuracy is predictive of zero-shot retrieval capability. Moreover, four ConvNeXt-based CLIP models trained with a limited range of random resize crop exhibit much lower retrieval performance.

of their training data. After applying temperature scaling and evaluating with the negative log-likelihood (NLL) metric, we observe that models with higher classification accuracy generally show better calibration. Importantly, when temperature scaling is applied to both CLIP and other models, zero-shot CLIP models consistently outperform other models, including fine-tuned versions, in calibration.

This pattern persists across various testing conditions, including ID and OOD sets, with zero-shot CLIP models demonstrating superior calibration compared to other models. This trend holds across both NLL and ECE metrics.

(2) ID calibration of CLIP models transfers to OOD test sets: While prior studies [92] report in-distribution (ID) calibration often fails to generalize under distribution shifts, our findings reveal a promising result for CLIP models. After calibrating CLIP models on the ID set, they exhibit improved calibration on OOD test sets. For example, on ImageNet-A, CLIP models exhibit lower calibration error after temperature scaling, a trend not seen in other models. This suggests that CLIP models are relatively easier to calibrate across diverse distributions, indicating their potential for robust and reliable applications in real-world settings.

VII. Zero-Shot Retrieval

Since CLIP models are trained using contrastive loss to associate text and image pairs, we evaluate their zero-shot retrieval capability on the Flickr30K [69] and MSCOCO [70] datasets in this section.

We have three major observations on the two datasets. **First**, CLIP's zero-shot retrieval capability correlates with its image classification performance. Fig. 5 illustrates image and text zero-shot retrieval (gauged by Recall@5) against their accuracy on ImageNet. We observe that classification ability is predictive of their retrieval capability. **Second**, training distribution deviates from the retrieval performance trend. Specifically, CLIP models trained on WIT slightly deviate from the trend formed by CLIP models trained on LAION, and the training quantity does not affect the trend. **Last**, we observe four specific ConvNeXt-based CLIP models significantly depart from the trend of LAION. We notice that they are trained with a limited random resize crop range $(0.9, 1.0)$. This limited augmentation likely reduces training view diversity, resulting in less variation in object scale and context. In image-text retrieval, where the model must extract consistent global representations that align well with corresponding textual descriptions, this lack of variation can hinder the learning of robust embeddings, ultimately affecting retrieval performance. While this work does not consider such training augmentations, it would be interesting to explore their impact on retrieval.

VIII. 3D Awareness

CLIP models are trained using contrastive loss to associate text and image pairs in feature space, but this training does not explicitly incorporate 3D understanding, such as recognizing geometric concepts like multi-view consistency and depth. Despite being trained on 2D data, recent studies suggest that models like CLIP can still be effective in 3D-related tasks [16], [93], [94]. Building on this insight, this section evaluates the behaviors of CLIP models in 3D-specific scenarios, particularly examining their ability to capture 3D geometry and their robustness to 3D distortions.

A. Correspondence Matching

Geometric Correspondence: Given two views of the same object or scene, the objective is to identify pixels in both views that correspond to the same location in 3D space. We evaluate this using recall on the ScanNet [72] dataset for object-centric correspondence and NAVI [73] for scene-centric correspondence. Correspondence recall measures the percentage of correct correspondences that fall within a defined threshold distance. Following the protocol in [16], we categorize performance based on the magnitude of transformation between view pairs.

Semantic Correspondence: This task generalizes geometric correspondence by requiring matching of semantically similar parts across different instances of the same object class. For example, mapping the left paw of two different dogs. We use the SPair-71K [74] dataset, with performance measured by recall. Similar to geometric correspondence, we group results by the degree of view variation. Fig. 6 groups CLIP models based on their visual encoder architectures (CNN-based and ViT-based). For comparison, we also include standard supervised models such as ConvNeXt and ViT-L/16 (DeiT III) [95], which are trained on ImageNet-22 K, alongside DINO-V2 [96].

Observations: First, ViT-based CLIP models exhibit weaker performance across three datasets (ScanNet, NAVI, and Spari-71 K), falling behind the supervised model (ViT-L-16), which also uses a transformer-based architecture. In contrast, CNN-based CLIPs consistently achieve higher recall scores than their

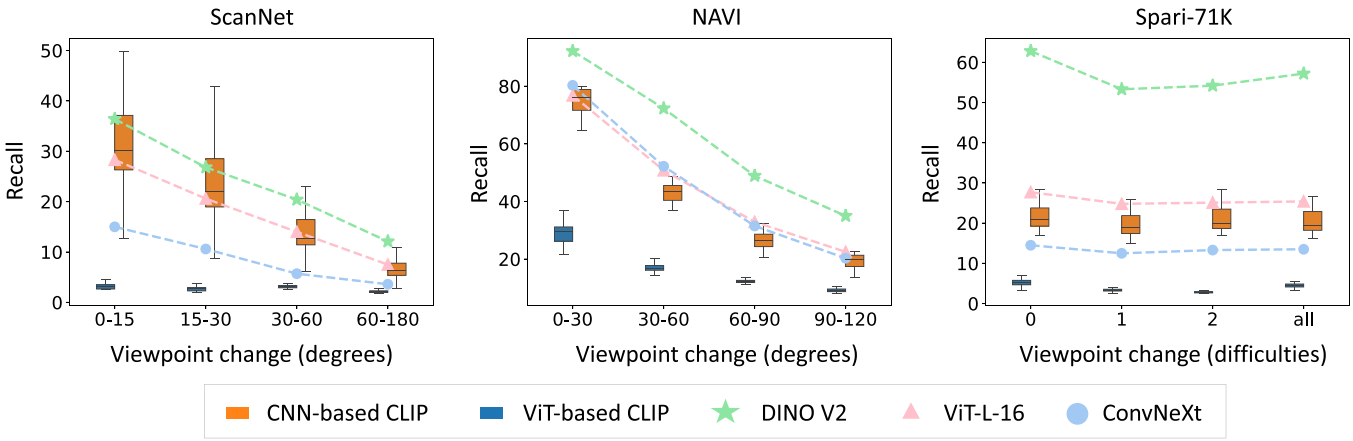


Fig. 6. Correspondence matching performance (Recall ↑) with respect to their viewpoint change. We report results on geometric correspondence matching (ScanNet, NAVI) and semantic correspondence matching (Spair-71K). CLIP models are grouped by the architecture of the visual encoder into CNN-based and ViT-based. We observe that CNN-based CLIP models consistently outperform ViT-based CLIP models, particularly in scenarios with larger viewpoint variations, and achieve competitive results compared to supervised models like ConvNeXt and ViT-L-16.

ViT-based counterparts, particularly as viewpoint changes become more extreme. Additionally, CNN-based CLIP models show competitive performance when compared to supervised CNN model ConvNeXt. This suggests the combined effect of the visual encoder architecture and training objective, which plays a crucial role in influencing CLIP's ability to manage correspondence matching. Second, our study extends the observation of [16], showing CNN-based CLIP models not only perform competitively with ViT-L/16 on NAVI but also match DINO-V2 on ScanNet. Note that, DINO-V2 emerges as the top performer across all three datasets. These findings suggest that CNN-based CLIPs generally exhibit stronger correspondence matching than ViT-based CLIPs, especially in scenarios involving significant viewpoint variations.

B. Robustness Against 3D Corruptions

We further evaluate the ability of CLIP models to handle 3D-related corruptions using the 3D Common Corruptions (3DCC) benchmark [17], which applies corruptions based on 3D transformations. Unlike the common corruptions in [31], these transformations consider the underlying geometry of the scene, producing distortions that are more reflective of real-world conditions. Sample images of corruptions are shown in the last row in Fig. 7. For example, the *fog* gets denser further away from the camera. In this study, we analyze six types of 3D-related corruptions, each with five severity levels, and examine only CLIP models pre-trained on LAION to maintain consistency in training dataset distributions. Based on correspondence matching, we categorize the CLIP models into CNN-based and ViT-based groups.

CNN-based CLIP models demonstrate stronger robustness to 3D-related corruptions as corruption intensity increases: Fig. 7 shows the performance of ViT-based and CNN-based CLIP models across various 3D-related corruptions (*Fog*, *Near Focus*, *Z-motion Blur*, *Flash*, *XY-motion-blur* and *Flash*) at different severity levels (Level 1, Level 3, and Level 5). For each row, the slope of the CNN-based models is consistently steeper than that of the ViT-based models, indicating that CNN-based models experience less degradation in performance as the clean ImageNet validation accuracy increases. This suggests that CNN-based models are more robust in maintaining accuracy under 3D distortions.

Furthermore, as the corruption intensity increases (moving from Level 1 to Level 5), the gap between the slopes, represented by $tan(\Delta_S)$, widens. This increase highlights that the advantage of CNN-based models becomes more pronounced under higher severity of corruptions, particularly for challenging distortions like *Fog* and *Z-motion Blur*. The growing slope difference indicates that CNN-based models are increasingly more capable of handling severe 3D corruptions compared to ViT-based models. These results reinforce the importance of visual encoder architecture in achieving robustness across varying corruption intensities, with CNN-based models consistently outperforming ViT-based models, especially as the corruption severity escalates. When considered alongside the results from the correspondence matching, these findings underscore the pivotal role that visual encoder architecture plays in enhancing robustness to 3D corruptions. Lastly, ViT-based CLIP models struggle with 3D geometric understanding, whereas DINO models perform better. This has implications for downstream multimodal models like LLaVA [18], which typically rely on CLIP backbones. Combining DINO with CNN-based CLIP features could improve spatial reasoning, as suggested in recent study [97].

IX. Visual and Language Encoder Interaction: A Classification Perspective

Modern large multimodal models (LMMs), such as LLaVA [18], typically use a frozen pre-trained visual encoder from CLIP as their visual backbone, with instruction fine-tuning applied to the linear projector and the language model components. This raises an important question: how does the interaction between a shared visual encoder and distinct language models affect the classification performance of LLaVA compared to CLIP-like models?

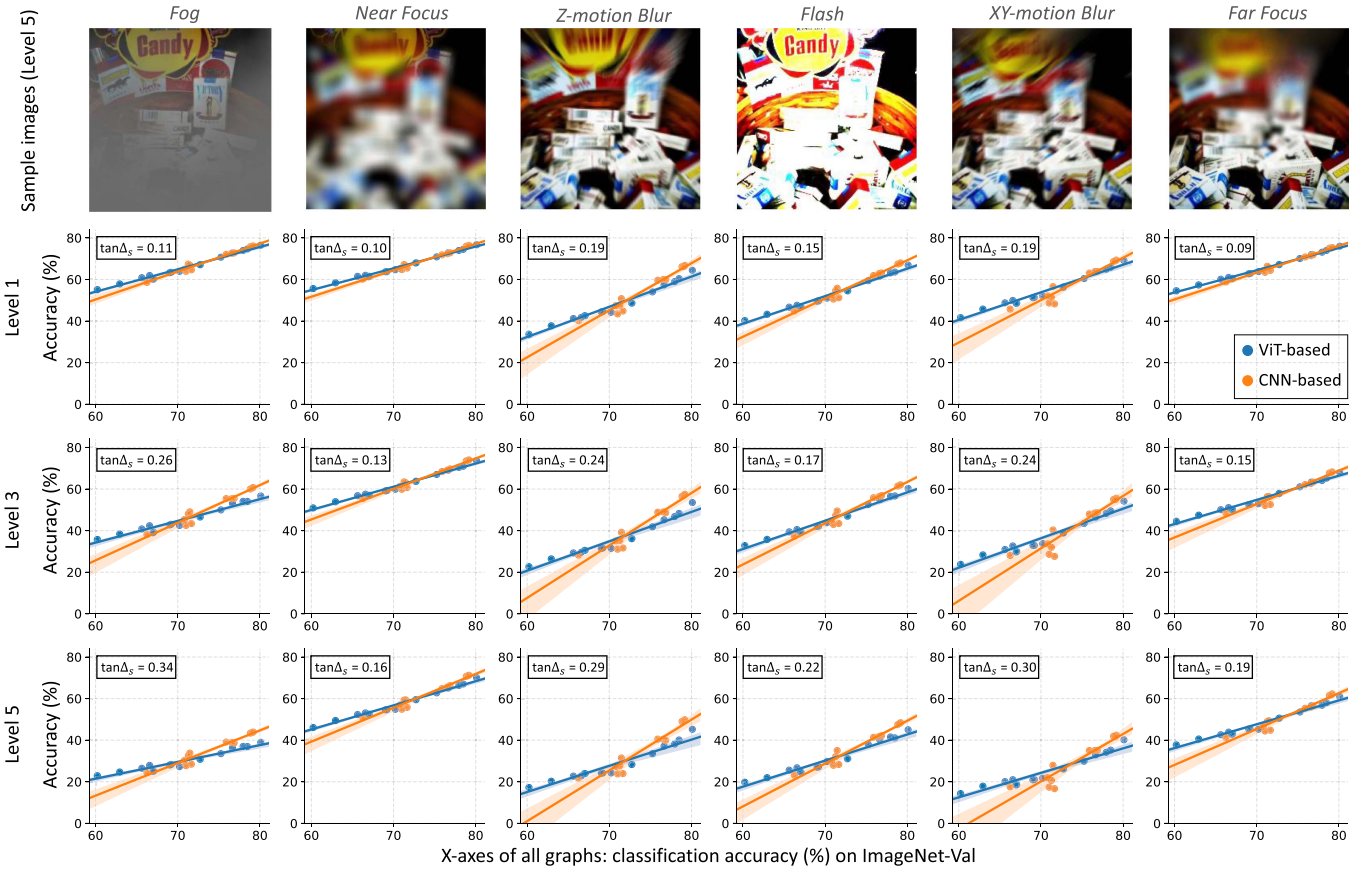


Fig. 7. Robustness comparison of ViT-based and CNN-based CLIP models under varying 3D-related corruptions. The x-axis represents accuracy on ImageNet-Val, while the y-axis represents accuracy on the corrupted dataset. We show the accuracy of ViT-based and CNN-based CLIP models across six types of 3D-related corruptions: *Fog*, *Near Focus*, *Z-motion Blur*, *Flash*, *XY-motion Blur*, and *Far Focus*, evaluated at three severity levels (Level 1, Level 3, and Level 5). Each column shows that CNN-based models consistently exhibit steeper slopes, indicating greater resilience with less performance degradation as ImageNet-Val accuracy improves. As corruption intensity increases, the gap between the slopes, represented by $tan(\Delta_S)$, widens, particularly under severe conditions like *Fog* and *Z-motion Blur*. This widening gap highlights the superior robustness of CNN-based models compared to their ViT-based counterparts, especially at higher corruption levels. This reinforces the significant impact of visual encoder architecture on CLIP's ability to handle 3D-related corruption. Sample images of Level 5 severity for each corruption are provided on the top for reference.

Driven by this, we compare the classification accuracy of CLIP and LLaVA to investigate how the interaction between the shared visual encoder and their distinct language models influences overall performance. In this section, "LLaVA" and "CLIP" refer to their training paradigms rather than specific model implementations. We also include SigLIP [46] as another representative of CLIP-like models.

Our evaluation is conducted on three splits of the ImageNet-D dataset [19]: *Background*, *Texture*, and *Material*. This dataset, generated by a text-to-image diffusion model, poses significant classification challenges. We adopt a VQA-style approach for LLaVA's classification, providing it with a category list per image and prompting it to select the correct category. The list includes the ground truth (GT) category and three "failure" categories—incorrect categories ranked with the highest confidence by a pretrained category selection model—ensuring a unique category list for each image. We evaluate the role of the category selection model using ResNet-50, CLIP-ViT-L/14-336, and SigLIP-SO-14.

To explore the interaction between the language and CLIP vision encoders, we consider six LLaVA models, combining

TABLE IV
COMPARED CLIP AND LLAVA MODELS ON IMAGENET-D

No.	Category List	Visual encoder	Type	LLM	Background	Material	Texture
1	ResNet-50	CLIP ViT-L/14-336 (WIT)	CLIP	-	0.90	0.92	0.95
			LLaVA	Mistral-Instruct-V2	0.82	0.81	0.80
			LLaVA	Llama2-Chat	0.91	0.87	0.86
			LLaVA	Vicuna-V2-7B	0.92	0.89	0.91
		SigLIP-SO-14	CLIP	-	0.97	0.96	0.99
			LLaVA	Mistral-Instruct-V2	0.84	0.73	0.79
			LLaVA	Llama2-Chat	0.93	0.91	0.93
			LLaVA	Vicuna-V2-7B	0.92	0.90	0.94
2	CLIP SigLIP-SO-14 (Webli)	ViT-L/14-336 (WIT)	CLIP	-	0.23	0.24	0.21
			LLaVA	Mistral-Instruct-V2	0.41	0.35	0.28
			LLaVA	Llama2-Chat	0.52	0.35	0.34
			LLaVA	Vicuna-V2-7B	0.57	0.48	0.42
		SigLIP-SO-14	CLIP	-	0.65	0.61	0.61
			LLaVA	Mistral-Instruct-V2	0.44	0.33	0.35
			LLaVA	Llama2-Chat	0.60	0.47	0.46
			LLaVA	Vicuna-V2-7B	0.59	0.48	0.43
3	CLIP ViT-L/14-336 (WIT)	ViT-L/14-336 (WIT)	CLIP	-	0.14	0.14	0.13
			LLaVA	Mistral-Instruct-V2	0.37	0.35	0.25
			LLaVA	Llama2-Chat	0.49	0.32	0.30
			LLaVA	Vicuna-V2-7B	0.57	0.45	0.42
		SigLIP-SO-14	CLIP	-	0.69	0.67	0.65
			LLaVA	Mistral-Instruct-V2	0.46	0.34	0.36
			LLaVA	Llama2-Chat	0.62	0.48	0.48
			LLaVA	Vicuna-V2-7B	0.59	0.52	0.44

We include two visual backbones: CLIP-L/14-336 and SigLIP-SO-L and two language models for LLaVA: Mistral-Instruct-V2, Llama2-Chat, and Vicuna-V2-7B.

two types of visual encoders—CLIP-ViT-L/14-336 and SigLIP-SO-14—and three language encoders: Mistral-Instruct-V2 [59], Llama2-Chat [60], and Vicuna-V2-7B [60]. For a fair comparison, CLIP is given the same category list using the default prompt template by [1] (e.g., "a photo of [category]"). LLaVA's prompt format is:

```
What is the main object in this image?
Choose from the following list:
  A.[Ground truth class]
  B.[Failure class 1]
  C.[Failure class 2]
  D.[Failure class 3]
  Please answer the question using the
choice from the list.
```

Observations: We report the results on ImageNet-D in Table IV and summarize the observations as follows.

First, extending the findings of [19], which evaluate CLIP (ViT/14) as a category selection model, we find that the interactions between vision and language components in selection-based networks vary significantly with task difficulty. When the category list is easy for CLIP, LLaVA models using the same visual encoder do not yield consistent improvements and sometimes exhibit slight performance drops. In contrast, when the category list is challenging for CLIP, LLaVA models using the same visual encoder show substantial gains. For example, in row 1, the most confused categories of ResNet-50 are easy for CLIP, and LLaVA brings no improvement. Similarly, in row 2, when SigLIP-SO-14 performs well, LLaVA shows a performance drop. However, in the same row, when the category list becomes difficult for SigLIP-SO-14, LLaVA improves accuracy by over 20% across three splits. A similar pattern is observed in row 3 for CLIP (ViT-L/14-336) and its LLaVA counterpart. Since LLaVA and CLIP share the same visual encoder, we speculate that the observed gains arise when CLIP's visual-text alignment is weak—likely due to ambiguous categories or limited pre-training coverage. In such cases, LLaVA's language model may help disambiguate visual features by leveraging external knowledge acquired during multimodal instruction tuning. Conversely, when CLIP already handles the token comparison effectively, the language model may over-interpret the visual input, occasionally leading to reduced performance.

Second, the choice of language model (LLM) within LLaVA significantly impacts classification performance. Vicuna-V2-7B consistently outperforms Mistral-Instruct-V2, while the choice of visual encoder also plays a critical role. LLaVA models built on SigLIP-SO-14 outperform those using ViT-L/14-336, echoing recent findings in the literature.

The above suggests that LLaVA's performance gains are not solely the result of architectural complexity or additional training data, but rather the interaction between the two. The model's effectiveness depends on how the language model, visual encoder, and learned projection layer work together in response to varying input complexity. This indicates the importance of designing vision-language models with careful attention to how components are integrated and how they respond under different levels of visual-text alignment difficulty.

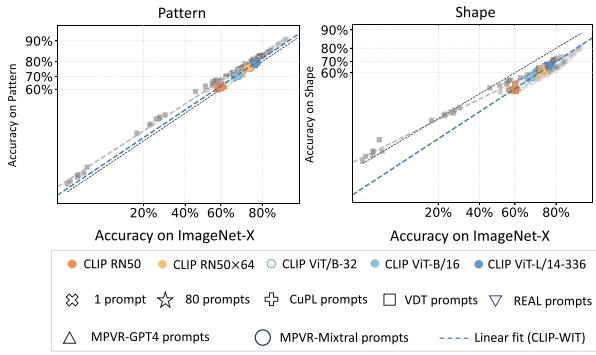


Fig. 8. Influence of test-time prompts on CLIP's visual-factor robustness. We evaluate five CLIP models trained on WIT, represented by different colors for architectures and different shapes. The dashed grey line represents robust linear regression [75] based on the original CLIP-WIT models with 80 prompts. Different prompt sets may influence classification performance but do not significantly impact visual factor robustness because models still lie on the original line.

TABLE V
INFLUENCE OF TEST-TIME PROMPTS ON CLIP'S CLASSIFICATION, OOD DETECTION AND CALIBRATION

Backbone	Pre-training dataset	Classification	OOD Detection		Calibration	
		Accuracy (↑)	AUROC (↑)	FPR (↓)	Before-temp ECE (↓)	After-temp ECE (↓)
RN50	1 Pompt	0.41	0.84	0.61	0.08	0.08
	80 Prompts	0.43	0.83	0.64	0.08	0.08
	CuPL	0.44	0.83	0.63	0.09	0.09
	VDT	0.42	0.82	0.65	0.10	0.09
	MPVR-GPT4	0.43	0.82	0.65	0.09	0.09
	MPVR-Mistral	0.43	0.82	0.66	0.08	0.09
	REAL	0.42	0.81	0.69	0.08	0.09
ViT-B/16	1 Pompt	0.57	0.86	0.55	0.05	0.06
	80 Prompts	0.59	0.85	0.57	0.05	0.06
	CuPL	0.60	0.86	0.54	0.06	0.06
	VDT	0.59	0.86	0.54	0.07	0.06
	MPVR-GPT4	0.60	0.85	0.58	0.05	0.06
	MPVR-Mistral	0.60	0.85	0.60	0.05	0.06
	REAL	0.58	0.83	0.63	0.06	0.06

We evaluate CLIP models trained on WIT with ResNet-50 and ViT-B/16 as the visual encoder. We find that prompt sets generated by large language models may improve zero-shot CLIP models' classification accuracy, but it does not enhance other OOD detection or calibration.

X. IMPACT OF TRAINING AND INFERENCE STRATEGY ON MODEL ROBUSTNESS

A. Impact of Test-Time Prompts

In the previous analyses, we used the default prompt set provided by [1]. Here, we investigate how varying test-time prompts influence CLIP's performance in out-of-distribution (OOD) detection, visual factor robustness, and predictive uncertainty. We experiment with five additional prompt sets: (1) a single prompt ("a photo of a {label}"); (2) a set generated by GPT-3 following [98]; (3) a prompt set generated by GPT-4 [99] using the chain-of-thought strategy [100] (VDT) [101]; (4) prompts generated by GPT-4 (MPVR-GPT4) or Mistral (MPVR-Mistral) with a target-task-oriented design [102]; (5) prompts generated by GPT-4 with Retrieval-Augmented Learning (REAL) [103]. These prompts are tested across five CLIP models—RN50, RN50× 64, ViT-B/16, ViT-B/32, and ViT-L/14-336—all trained on the WIT dataset.

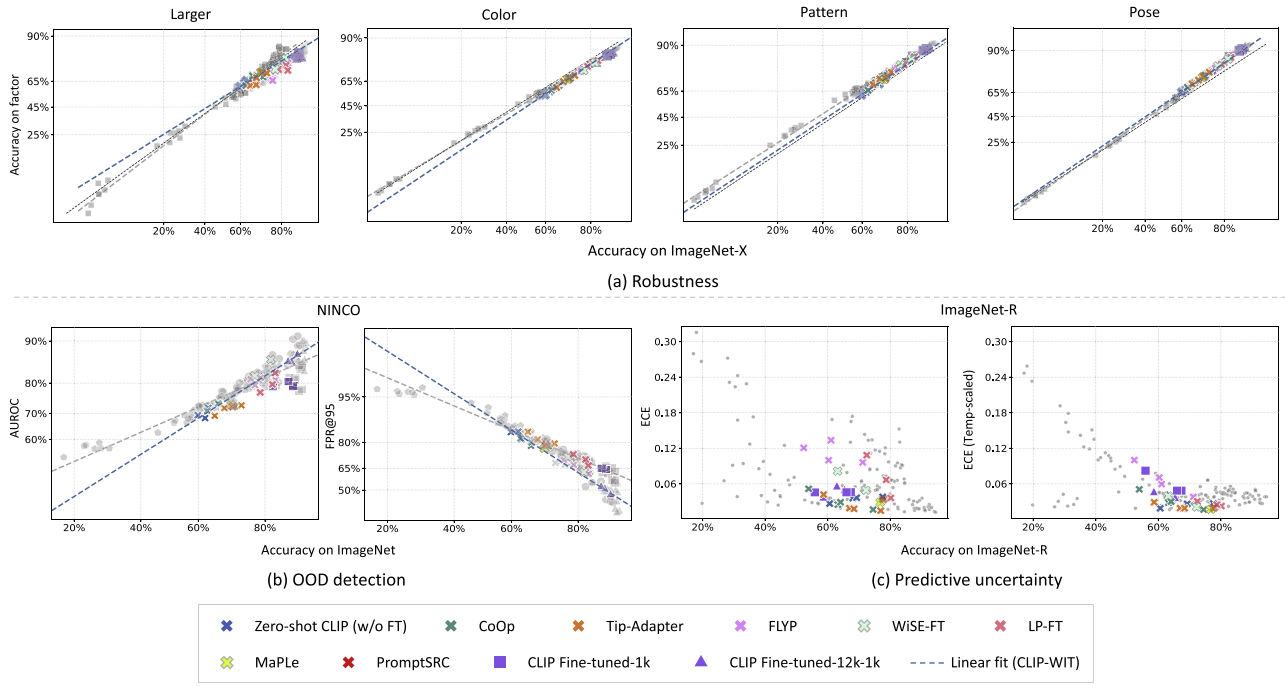


Fig. 9. Influence of fine-tuning algorithms on CLIP's robustness, OOD detection, and predictive uncertainty. We fine-tune four CLIP models trained on WIT using various algorithms. Different colors represent model architectures, and different shapes denote fine-tuning algorithms. The blue dashed line is fit with robust linear regression [75] for original CLIP-WIT models, while the grey dashed line represents zero-shot CLIP trained on LAION. Results show that contrastive fine-tuning improves overall classification accuracy but negatively impacts predictive uncertainty.

Fig. 8 and TableV summarize the effects of various prompt strategies on CLIP's classification, robustness, OOD detection, and calibration. Using fewer prompts (e.g., a single prompt) reduces classification accuracy but improves OOD detection and calibration. Factor-level robustness, such as on the Pattern task, remains largely unchanged regardless of prompt type, with models following the original CLIP-WIT trend. Prompt sets generated by large language models—including VDT, MPVR-GPT4, MPVR-Mistral, and REAL—consistently improve classification accuracy, but show no clear advantage in OOD detection or calibration. These results highlight a key challenge: how to design prompt strategies that simultaneously enhance other objectives beyond classification accuracy only.

TABLE VI
PERFORMANCE OF VARIOUS TRAINING PARADIGMS ON CLASSIFICATION, OOD DETECTION, CALIBRATION AND 3D ROBUSTNESS

Training paradigm	Classification	OOD Detection		Calibration		3D Robustness
	Accuracy (↑)	AUROC (↑)	FPR (↓)	Before-temp ECE (↓)	After-temp ECE (↓)	Accuracy (↑)
CLIP-ViT-B/16	0.59	0.85	0.57	0.06	0.06	0.34
BLIP-Base	0.51	0.76	0.70	0.07	0.08	0.27
SigLIP-ViT-B/16	0.68	0.89	0.50	0.08	0.05	0.38
ViTamin-Base	0.62	0.84	0.67	0.37	0.23	0.35
CLIP-ViT-L/14	0.72	0.88	0.47	0.06	0.05	0.46
BLIP-2	0.53	0.66	0.89	0.14	0.08	0.37
SigLIP-ViT-L/16	0.75	0.91	0.40	0.07	0.04	0.46
ViTamin-L-256px	0.79	0.89	0.49	0.21	0.17	0.53

We evaluate five vision-language training paradigms: CLIP, BLIP, BLIP-2, SigLIP and ViTamin. We find that no training paradigm is the most performant on all considered safety-related objectives.

B. *Effect of Fine-Tuning Procedures*

In addition to standard fine-tuning methods (i.e:, cross-entropy fine-tuning on ImageNet), we examine seven alternative fine-tuning strategies: contrastive fine-tuning (FLYP) as introduced by [79], two robsut fine-tuning methods–WiSE-FT [10] and LP-FT [104], and four parameter-efficient methods–CoOp [48], Tip-Adapter [49], MaPLe [105] and PromptSRC [106]. They are applied to fine-tune zero-shot CLIP models pre-trained on WIT.

In Fig. 9, we present the performance of fine-tuned CLIP models across visual factor robustness, OOD detection, and calibration. The results reveal mixed effects across different tuning methods. For visual factor robustness, CoOp and PromptSRC preserve the properties of zero-shot CLIP, aligning with prior findings that test-time prompts have limited influence on robustness. FLYP and Tip-Adapter improve robustness against the *Pattern* factor but reduce it under *Larger* visual changes. LP-FT maintains robustness across several factors, while WiSE-FT slightly weakens it on *Larger*. On OOD detection, most methods—including FLYP, LP-FT, WiSE-FT, and PromptSRC—enhances both accuracy and detection performance, with LP-FT showing strong generalization as their models lie above the zero-shot CLIP-WIT trend. For calibration, FLYP increases calibration error, while CoOp, Tip-Adapter, and PromptSRC maintain well-calibrated predictions. LP-FT and WiSE-FT increase error slightly before temperature scaling but recover calibration performance afterward, outperforming FLYP in uncertainty estimation.

These findings suggest that while fine-tuning can improve certain aspects of CLIP's performance, achieving a balance between classification accuracy, OOD detection, and predictive uncertainty remains a challenge, highlighting the need for further

TABLE VII
COMPARISON OF CLIP TRAINED WITH FILTERED PRE-TRAINING DATA ON SIX TASKS

Backbone	Pre-training dataset	Data Filtering	Classification	OOD Detection		Calibration		Visual factor robustness			Retrieval	3D robustness
						Before-temp	After-temp	Larger	Shape	Color		
			Accuracy (↑)	AUROC (↑)	FPR (↓)	ECE (↓)	ECE (↓)	Accuracy (↑)	Accuracy (↑)	Accuracy (↑)	Recall@5 (↑)	Accuracy (↑)
ViT-B/16	LAION-400M	No	0.61	0.84	0.65	0.13	0.05	0.67	0.56	0.63	0.82	0.32
	MetaCLIP-400M	Yes	0.67	0.85	0.62	0.09	0.08	0.75	0.61	0.67	0.83	0.35
	LAION-2B	No	0.64	0.85	0.64	0.13	0.05	0.69	0.60	0.67	0.84	0.34
	DFN-2B	Yes	0.70	0.88	0.52	0.12	0.07	0.80	0.66	0.73	0.85	0.40
	CommonPool-L	No	0.43	0.73	0.86	0.06	0.07	0.45	0.46	0.58	0.64	0.19
	CommonPool-L-*CLIP*	Yes	0.53	0.77	0.81	0.11	0.07	0.61	0.53	0.56	0.72	0.26
ViT-L/14	LAION-400M	No	0.68	0.86	0.59	0.17	0.06	0.75	0.64	0.70	0.85	0.38
	MetaCLIP-400M	Yes	0.76	0.89	0.50	0.09	0.06	0.74	0.67	0.74	0.85	0.45
	LAION-2B	No	0.72	0.88	0.52	0.11	0.04	0.82	0.66	0.72	0.87	0.42
	DFN-2B	Yes	0.78	0.91	0.39	0.07	0.04	0.85	0.74	0.79	0.88	0.50
	CommonPool-XL	No	0.72	0.87	0.54	0.03	0.04	0.72	0.65	0.70	0.80	0.43
	CommonPool-XL-*CLIP*	Yes	0.75	0.88	0.54	0.08	0.03	0.74	0.67	0.74	0.84	0.46
ConvNeXt-Base	LAION-2B	No	0.64	0.85	0.64	0.12	0.05	0.69	0.59	0.67	0.72	0.35
	LAION-Aesthetic	Yes	0.65	0.85	0.63	0.14	0.05	0.71	0.62	0.67	0.68	0.33

For the classification task, we report average accuracy on ImageNet validation, ImageNet-V2-A, ImageNet-S, ObjectNet, ImageNet-A, ImageNet-R and ImageNet-Vid. We report averaged AUROC and FPR on NINCO, iNaturalist, DTD, Place, SUN and ImageNet-O. We report ECE before and after calibration. The calibration set is ID-val and test set is the same as OOD generalization. For visual factor robustness, we evaluate *Larger*, *Shape* and *Color*. We use averaged recall@5 to measure text-to-image and image-to-text retrieval on MSCoCo and Flick30K. For 3D robustness, we use accuracy to metric their mean performance on six 3D-related corruptions with severity level 5. The best performance for each architecture is in green . We find that data curation technique is an effective method for enhancing model performance beyond classification.

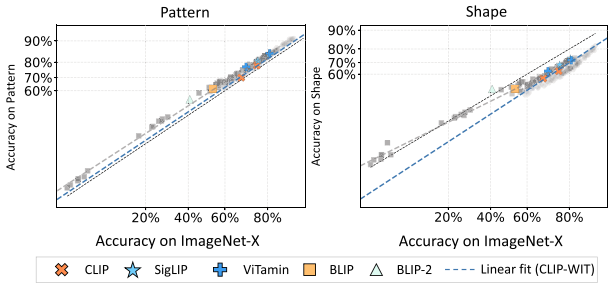


Fig. 10. Visual-factor robustness of different training paradigms from CLIP. We evaluate five vision-language training paradigms: CLIP, BLIP, BLIP-2, SigLIP and ViTamin. The dashed blue line represents robust linear regression [75] based on the original CLIP-WIT models. Different training paradigms effectively impact visual factor robustness.

research into fine-tuning strategies that can address all of these objectives.

C. Extending to Different Training Paradigms

We have expanded our analysis beyond CLIP to include four additional vision-language training paradigms: (1) BLIP [79], (2) BLIP-2 [107], (3) SigLIP [46], and (4) ViTamin [108]. We summarized the results in Fig. 10 and Table VI We observe that different training paradigms yield trade-offs across robustness, OOD detection, calibration, and 3D performance. SigLIP achieves the best overall balance, with strong OOD detection, low calibration error, and competitive 3D robustness.ViTamin-L-256px leads in classification accuracy and 3D robustness but suffers from poor calibration. In addition, BLIP and BLIP-2 consistently underperform across most metrics. These results reinforce our main claim: no single paradigm excels universally, underscoring the need for multi-dimensional evaluation beyond accuracy alone.

D. Robustness Evaluation of Dataset Curation

High-quality training sets are crucial for developing CLIP models, and as a result, recent research has increasingly emphasized dataset curation (DC) to create these datasets [21], [44], [45]. In this work, we extend the evaluation of DC techniques to robustness-related tasks, including out-of-distribution (OOD) detection, calibration, visual factor-level robustness, and 3D corruption.

To ensure a clear and fair comparison, we control the architecture of the CLIP models and categorize the methods based on their pretraining dataset sources. We consider four DC techniques: 1) CommonPool [21], which uses a trained CLIP model as a filter; 2) MetaCLIP [45], which leverages metadata for curation and balancing of raw web-sourced data; and 3) DFN-2B [44], which employs a network trained on high-quality datasets for filtering; 4) Aesthetic [109], which is filtered using perceptual hashing for deduplication and an aesthetic score threshold.

Table VII shows that DC techniques consistently improve performance in classification, OOD detection, visual factor robustness, and 3D robustness—particularly for transformer-based models. However, their impact on calibration is limited. We also evaluate CNN-based CLIP models trained on LAION-Aesthetic and observe that while dataset filtering improves classification and robustness, it shows limited benefits for retrieval, calibration, and 3D robustness. These observations suggest that the effectiveness of filtering strategies depends on both model architecture and the specific evaluation objective, emphasizing the need for multi-dimensional assessment beyond classification accuracy.

XI. CONCLUSION AND DISCUSSION

Our research contributes to the ongoing discussion regarding the robustness and capabilities of CLIP models, particularly in response to visual factor robustness, OOD detection, the reliability of uncertainty estimation, zero-shot retrieval capabilities, and 3D awareness. To achieve these insights, we performed comprehensive experiments and comparative analyses, systematically evaluating CLIP models against diverse model families. Through an in-depth exploration of critical factors—including training sources, contrastive learning objectives, network architecture, fine-tuning strategies, and test-time prompt

variations—our findings provide substantial insights into the unique advantages CLIP models offer.

Discussion on Dataset Overlap: Given that CLIP models are pretrained on large-scale web-crawled datasets such as LAION-5B, potential overlap with evaluation benchmarks is a valid concern. Prior work suggests that such overlap is unlikely to significantly affect our findings. For classification robustness, LAION-5B paper [21] and OpenAI [1] report only isolated cases where overlap impacts performance, and do not view it as a major threat to result validity. For OOD detection, Bitterwolf et al. [66] show that overlapping class semantics between pretraining (e.g., IN-21 K) and test sets (e.g., NINCO) does not substantially alter detection performance. For calibration, since our evaluation datasets are shared with classification and focus on relative model comparison, any sample-level overlap is unlikely to influence conclusions. Moreover, our emphasis on relative trends, rather than absolute scores, further mitigates this concern.

This work leaves open many interesting and promising directions and we discuss a few. **First**, we offer an analysis of LLaVA and demonstrate that its large language model can assist in classification where CLIP's text and visual tokens are misaligned. Future work could explore other modern large vision models (LVMs), such as BLIP-3 [110] and Otter [11], to deepen this analysis. Further exploration into the interaction between language models and CLIP's visual encoder could also yield valuable insights. We see our analysis as a starting point. **Second**, our study includes two academic training sources—WIT and LAION—for CLIP. Future work should investigate whether our findings generalize to other training sources, such as datasets generated by Stable Diffusion [111], to advance our understanding of multi-modal dataset design. **Lastly**, our analysis reveals a critical need for more refined fine-tuning strategies tailored to CLIP models, aimed at improving both classification accuracy and robustness.

References

[1] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[2] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[3] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5389–5400.

[4] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 10506–10518.

[5] D. Hendrycks et al., "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8320–8329.

[6] A. Barbu et al., "ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9453–9463.

[7] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15262–15271.

[8] T. Nguyen, G. Ilharco, M. Wortsman, S. Oh, and L. Schmidt, "Quality not quantity: On the interaction between dataset design and robustness of clip," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 21455–21469.

[9] M. Cherti et al., "Reproducible scaling laws for contrastive language-image learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2818–2829.

[10] M. Wortsman et al., "Robust fine-tuning of zero-shot models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7959–7971.

[11] Y. Zhao et al., "On evaluating adversarial robustness of large vision-language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 54111–54138.

[12] A. Fang et al., "Data determines distributional robustness in contrastive language image pre-training (CLIP)," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 6216–6234.

[13] Z. Shi et al., "Effective robustness against natural distribution shifts for models with different training data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 73543–73558.

[14] B. Y. Idrissi et al., "ImageNet-X: Understanding model mistakes with factor of variation annotations," in *Proc. Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=HXz7Vcm3VgM

[15] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li, "Delving into out-of-distribution detection with vision-language representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 35087–35102 .

[16] M. El Banani et al., "Probing the 3 d awareness of visual foundation models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 21795–21806.

[17] O. F. Kar, T. Yeo, A. Atanov, and A. Zamir, "3 d common corruptions and data augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18963–18974.

[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. neural Inf. Process. Syst.*, 2024, pp. 34892–34916.

[19] C. Zhang, F. Pan, J. Kim, I. S. Kweon, and C. Mao, "ImageNet-D: Benchmarking neural network robustness on diffusion synthetic object," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 21752–21762.

[20] W. Tu, W. Deng, and T. Gedeon, "A closer look at the robustness of contrastive language-image pre-training (CLIP)," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 13678–13691.

[21] S. Y. Gadre et al., "DataComp: In search of the next generation of multimodal datasets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 27092–27112.

[22] J. Djolonga et al., "On robustness and transferability of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16458–16468.

[23] P. W. Koh et al., "Wilds: A benchmark of in-the-wild distribution shifts," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5637–5664.

[24] A. Kirsch and Y. Gal, "A note on "assessing generalization of SGD via disagreement"," *Trans. Mach. Learn. Res.*, 2022. [Online]. Available: https://openreview.net/forum?id=oRP8urZ8Fx

[25] Z. Huang, C. Liu, Y. Dong, H. Su, S. Zheng, and T. Liu, "Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 19973–20003. [Online]. Available: https://openreview.net/forum?id=LwOfVWgEzS

[26] Z. Huang et al., "Harnessing out-of-distribution examples via augmenting content and style," in *Proc. Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=boNyg20-JDm

[27] Z. Huang et al., "Winning prize comes from losing tickets: Improve invariant learning by exploring variant parameters for out-of-distribution generalization," 2023, arXiv: *2310.16391*.

[28] B. Schölkopf, J. Platt, and T. Hofmann, "Analysis of representations for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 137–144.

[29] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," 2009, arXiv: *0902.3430*.

[30] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bygh9j09KX

[31] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=HJz6tiCqYm

[32] E. Mintun, A. Kirillov, and S. Xie, "On interaction between augmentations and corruptions in natural corruption robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 3571–3583.

[33] C. Baek, Y. Jiang, A. Raghunathan, and J. Z. Kolter, "Agreement-on-the-line: Predicting the performance of neural networks under distribution shift," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 19274–19289.

[34] A. Shtedritski, C. Rupprecht, and A. Vedaldi, "What does clip know about a red circle? visual prompt engineering for vlms," 2023, arXiv: *2304.06712*.

[35] H. Cheng, E. Xiao, and R. Xu, "Typographic attacks in large multimodal models can be alleviated by more informative prompts," 2014, arXiv: *2304.06712*.

[36] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[37] K. Nguyen and B. O'Connor, "Posterior calibration and exploratory analysis for natural language processing models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1587–1598.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016.

[39] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11966–11976.

[40] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[41] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," 2023, arXiv: *2303.15389*.

[42] C. Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. 36th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2022, pp. 25278–25294. [Online]. Available: https://openreview.net/forum?id=M3Y74vmsMcY

[43] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.

[44] A. Fang, A. M. Jose, A. Jain, L. Schmidt, A. Toshev, and V. Shankar, "Data filtering networks," in *Proc. Adv. Neural Inf. Process. Syst. Workshop Distrib. Shifts*, 2024. [Online]. Available: https://openreview.net/forum?id=KAk6ngZ09F

[45] H. Xu et al., "Demystifying clip data," in *Proc. Int. Conf. Learn. Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=5BCFlnfE1g

[46] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11975–11986.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[48] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

[49] R. Zhang et al., "Tip-adapter: Training-free adaption of CLIP for few-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 493–510.

[50] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[51] X. Ding, C. Xia, X. Zhang, X. Chu, J. Han, and G. Ding, "RepMLP: Re-parameterizing convolutions into fully-connected layers for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2022, pp. 578–587.

[52] I. O. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 3615–3624.

[53] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18583–18599.

[54] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[55] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[56] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[57] R. Wightman, "Pytorch image models," 2019. [Online]. Available: https://github.com/rwightman/pytorch-image-models

[58] G. Ilharco et al., "Openclip," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5143773

[59] A. Q. Jiang et al., "Mistral 7B," 2023, arXiv: *2310.06825*.

[60] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, arXiv: *2307.09288*.

[61] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, "Prismatic VLMs: Investigating the design space of visually-conditioned language models," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 23123–23144.

[62] G. Van Horn et al., "The inaturalist species classification and detection dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8769–8778.

[63] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3485–3492.

[64] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[65] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3606–3613.

[66] J. Bitterwolf, M. Müller, and M. Hein, "In or out? fixing imagenet out-of-distribution detection evaluation," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 2471–2506.

[67] V. Shankar, A. Dave, R. Roelofs, D. Ramanan, B. Recht, and L. Schmidt, "Do image classifiers generalize across time?," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9661–9669.

[68] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2901–2907.

[69] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014. [Online]. Available: https://aclanthology.org/Q14-1006

[70] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, arXiv: *1504.00325*.

[71] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[72] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3 D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5828–5839.

[73] V. Jampani et al., "Navi: Category-agnostic image collections with high-quality 3D shape and pose annotations," in *Proc. Adv. Neural Inf. Process. Syst. Dataset Benchmark Track*, 2023, pp. 76061–76084.

[74] J. Min, J. Lee, J. Ponce, and M. Cho, "Spair-71 k: A large-scale benchmark for semantic correspondence," 2019, arXiv: *1908.10543*.

[75] P. J. Huber, "Robust statistics," in *International Encyclopedia of Statistical Science*. Berlin, Germany: Springer, 2011, pp. 1248–1251.

[76] J. P. Miller et al., "Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 7721–7735.

[77] Y. Xiao, Z. Tang, P. Wei, C. Liu, and L. Lin, "Masked images are counterfactual samples for robust fine-tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 20301–20310.

[78] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.

[79] S. Goyal, A. Kumar, S. Garg, Z. Kolter, and A. Raghunathan, "Finetune like you pretrain: Improved finetuning of zero-shot vision models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19338–19347.

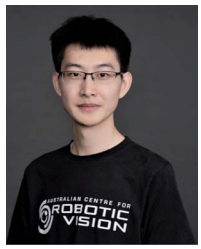
[80] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 23296–23308.

[81] C. Zhang et al., "Delving deep into the generalization of vision transformers under distribution shifts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7277–7286.

[82] K. Hermann, T. Chen, and S. Kornblith, "The origins and prevalence of texture bias in convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 19000–19015.

[83] T. Li, Z. Wen, Y. Li, and T. S. Lee, "Emergence of shape bias in convolutional neural networks through activation sparsity," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024.
[84] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[85] I. Bello et al., "Revisiting resnets: Improved training and scaling strategies," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 22614–22627.
[86] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[87] R. Geirhos et al., "Inducing a human-like shape bias leads to emergent human-level distortion robustness in cnns," *J. Vis.*, vol. 19, no. 10, pp. 209c–209c, 2019.
[88] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 7068–7081.
[89] D. Hendrycks and K. Gimpel, "A Baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Hkg4TI9xl
[90] M. Minderer et al., "Revisiting the calibration of modern neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 15682–15694.
[91] K. Gupta, A. Rahimi, T. Ajanthan, T. Mensink, C. Sminchisescu, and R. Hartley, "Calibration of neural networks using splines," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=eQe8DEWNN2W
[92] Y. Ovadia et al., "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14003–14014.
[93] J. Zhang et al., "Telling left from right: Identifying geometry-aware semantic correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 3076–3085.
[94] J. Zhang et al., "A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 45533–45547.
[95] H. Touvron, M. Cord, and H. Jégou, "Deit iii: Revenge of the ViT," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 516–533.
[96] M. Oquab et al., "Dinov2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2024. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt
[97] O. F. Kar, A. Tonioni, P. Poklukar, A. Kulshrestha, A. Zamir, and F. Tombari, "Brave: Broadening the visual encoding of vision-language models," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 113–132.
[98] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 15691–15701.
[99] J. Achiam et al., "GPT-4 technical report," 2023, arXiv: *2303.08774*.
[100] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 24824–24837.
[101] M. Maniparambil, C. Vorster, D. Molloy, N. Murphy, K. McGuinness, and N. E. O'Connor, "Enhancing CLIP with GPT-4: Harnessing visual descriptions as prompts," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 262–271.
[102] M. J. Mirza et al., "Meta-prompting for automating zero-shot visual recognition with llms," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 370–387.
[103] S. Parashar et al., "The neglected tails in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 12988–12997.
[104] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=UYneFzXSJWh
[105] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19113–19122.
[106] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, "Self-regulating prompts: Foundational model adaptation without forgetting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 15190–15200.
[107] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.
[108] J. Chen, Q. Yu, X. Shen, A. Yuille, and L.-C. Chen, "Vitamin: Designing scalable vision models in the vision-language ERA," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 12954–12966.
[109] C. Schuhmann and R. Beaumont, "Laion-aesthetics," 2022. Accessed: Apr. 25, 2025. [Online]. Available: https://laion.ai/blog/laion-aesthetics/
[110] L. Xue et al., "XGEN-MM (BLIP-3): A family of open large multimodal models," 2024, arXiv: *2408.08872*.
[111] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10684–10695.

Weijie Tu received the bachelor degree in advanced computing with first class honours from Australian National University, Canberra, Australia, in 2021. He is currently working toward the PhD degree with the Australian National University. His research focuses on trustworthy machine learning, particularly in OOD generalization. He has published papers in top-tier conferences and journals, including ICML, NeurIPS, CVPR, and TMLR, with a focus on unsupervised model evaluation and vision language models.



Weijian Deng received the PhD degree from Australian National University, Canberra, Australia, in 2023. He is currently a research fellow with the Australian National University. His research focuses on 3D content modeling, machine learning safety, and model generalization prediction. He has published extensively in top-tier conferences and journals, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, ICML, NeurIPS, CVPR, and ICCV, with a focus on unsupervised evaluation techniques and advancing AI robustness.

Tom Gedeon (Senior Member, IEEE) received the BSc and PhD degrees from the University of Western Australia, and Grad Dip Management from UNSW. He is the Human-Centric Advancements Chair in AI with Curtin University. He is an international research professor with Obuda University in Hungary. He is an Honorary professor with the Australian National University, where he was formerly Deputy Dean and Head of Computer Science. He is former president of the Asia-Pacific Neural Network Assembly, and of the Computing Research and Education Association of Australasia. He is a member of the Governing Boards of the *Asia-Pacific Neural Network Society* and the *IEEE Systems Man and Cybernetics Society*. His research interests are in responsive AI and responsible AI.