Evaluating the Robustness of Vision Foundation Models on an ImageNet Class-Subset Using Photorealistic Images Synthesized in Unreal Engine

Master Thesis

Master of Science in Applied Computer Science

Markus Leinberger

August 03, 2026

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: M.Sc Sebastian Dörrich

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

The rapid emergence of Vision Foundation Models (VFMs), such as CLIP and DINOv2, has fundamentally transformed computer vision by providing robust representations pre-trained on web-scale data. Despite their high performance on standard benchmarks, these models often exhibit significant fragility when subjected to Out-of-Distribution (OOD) shifts in physical factors like viewpoint, material, and environmental context. This thesis investigates the representational boundaries of VFMs by introducing a controlled, procedurally driven synthetic image rendering pipeline developed in Unreal Engine 5 (UE5). We aligned ShapeNet 3D models with the ImageNet-1K taxonomy to create two new evaluation benchmarks. The Multi-Factor Dataset isolates the causal effects of five nuisance factors: viewpoint, material, illumination, background, and fog. The Multi-Color Dataset examines texture bias by applying uniform color overrides. Using a triple evaluation strategy involving logits, k-nearest neighbors (k-NN), and linear probing, this study assesses how well different architectures perform at various levels of semantic understanding. The findings show a significant drop in performance when models switch from natural to synthetic data. Standard supervised models, such as ResNet-50, face drastic declines in effectiveness. However, self-supervised models like DINOv2 and DINOv3 hold up much better, although they still face a common failure issue when viewed from top-down perspectives. Furthermore, a representational analysis using equivariance metrics and Grad-CAM spatial attribution demonstrates that supervised fine-tuning (SFT) often destroys the latent invariance of foundation backbones by introducing a strong contextual bias toward background structures. Finally, by evaluating models at the ShapeNet superclass level, we show that VFMs maintain higher semantic stability than fine-grained metrics suggest. Ultimately, this research demonstrates that programmatic 3D rendering provides a necessary diagnostic lens for thoroughly evaluating advanced vision models.

The plugin code, evaluation code, and datasets are available on GitHub[1].

[1] `https://github.com/Beamfreak/VFM-Robustness-UE5`

Contents

List of Figures

List of Tables

List of Acronyms

Foundational Concepts and Models:	
VFM	Vision Foundation Model
CLIP	Contrastive Language–Image Pre-training
DINO	self-DIstillation with NO labels
ViT	Vision Transformer
VLM	Vision-Language Model
UE5	Unreal Engine 5
PUG	Photorealistic Unreal Graphics
Technical and Evaluation Metrics:	
OOD	Out-of-Distribution
ID	In-Distribution
k-NN	k-Nearest Neighbors
SFT	Supervised Fine-Tuning
AUROC	Area Under the Receiver Operating Characteristic curve
PCA	Principal Component Analysis
Grad-CAM	Gradient-weighted Class Activation Mapping
i.i.d.	Independent and identically distributed
Domain and Task-Specific Acronyms:	
AD	Anomaly Detection
ND	Novelty Detection
OSR	Open Set Recognition
LOD	Level of Detail
CNN	Convolutional Neural Network
DNN	Deep Neural Network
SPAR	Scene, Position, Attribute, and Relation
CGI	Computer-Generated Imagery
RGB	Red, Green, Blue (color spectrum)
Other Technical References:	
AI	Artificial Intelligence
FAISS	A similarity search library
iBOT	Image BERT Pre-Training

Notation

Sets and Spaces

$\mathcal{X}$	The input space (image domain)
$\mathcal{Y}$	The label space
$\mathcal{Y}_{\text{IN}}$	The set of 1,000 fine-grained ImageNet-1K class labels
$\mathcal{Y}_{\text{SN}}$	The set of coarse ShapeNet superclass labels
$\mathcal{I}_{1000}$	The set of ImageNet class indices $0, \ldots, 999$
$\mathcal{S}$	The set of ShapeNet superclasses
$\mathcal{E}$	The set of evaluable samples containing a valid ShapeNet mapping
$\mathcal{D}_{f=v}$	Subset of data where nuisance factor f is set to value v
$\varnothing$	Indicator for an unmapped class

Probability and Information Theory

$P(X, Y)$	Joint probability distribution over input space $\mathcal{X}$ and label space $\mathcal{Y}$
$P(X)$	Marginal distribution over the input space $\mathcal{X}$
$P(Y \mid X)$	Conditional probability of labels given input features
$P(\mathcal{Y}_{\text{ID}})$	Probability distribution of labels for In-Distribution data
$P(\mathcal{Y}_{\text{OOD}})$	Probability distribution of labels for Out-of-Distribution data
$\hat{p}_i$	Predicted class probability distribution for sample i

Functions and Mappings

ϕ	Taxonomic mapping function: $\mathcal{Y}_{\text{IN}} \to \mathcal{Y}_{\text{SN}} \cup \varnothing$
$\mathcal{M}$	Formal taxonomic mapping function bridging ImageNet and ShapeNet
$\Phi(\cdot)$	Feature extraction function used for representation analysis
$\mathbb{1}$	Indicator function: 1 if the condition is true, 0 otherwise
Top5$(\cdot)$	Function returning the five most likely predicted labels
TopK	Function returning the top K predicted indices or probabilities

Linear Algebra and Model Components

$\ell(x)$	The raw output logit vector in $\mathbb{R}^{1000}$ for input x
z	High-dimensional feature embedding vector in $\mathbb{R}^d$
$\tilde{z}$	L_2-normalized feature embedding vector
W	Weight matrix for a trained linear classifier or probe
b	Bias vector for a trained linear classifier or probe
τ	Temperature parameter for weighted voting in k-NN classification
$\Delta_{v_i}(a \to b)$	Latent transition vector representing an object's shift from perspective a to b
$\text{Sim}_{i,j}(a \to b)$	Cosine similarity between transition vectors of objects i and j

Evaluation Metrics and Statistics

y_i	Ground-truth ImageNet label for sample i
$\hat{y}_i$	Model's predicted label for sample i
c_i^{IN}	Binary indicator of exact ImageNet class correctness
c_i^{SN}	Binary indicator of ShapeNet superclass correctness
A^{IN}	Top-1 ImageNet-level classification accuracy
$A^{\mathrm{IN(5)}}$	Top-5 ImageNet-level classification accuracy
A^{SN}	ShapeNet-level superclass accuracy
$A_{f=v}$	Per-factor accuracy for a specific factor level
$\Delta_{f=v}$	Absolute robustness degradation (accuracy drop) from baseline v_0
$R_{f=v}$	Relative robustness degradation ratio
$\overline{\Delta}_f$	Average absolute degradation across a range of factor values
ρ	Spearman rank correlation coefficient for benchmark comparison
ϵ	A small constant used to avoid division by zero

1 Introduction

The rapid advancement of Vision Foundation Models (VFMs), such as CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024), has fundamentally transformed the landscape of computer vision, establishing a paradigm where large-scale models trained on web-scale data serve as a common basis for a wide range of downstream applications (Bommasani et al., 2021). These models are often deployed under the assumption of universal generalization. However, they frequently exhibit significant fragility when encountering distribution shifts in physical factors such as camera perspective, material properties, lighting conditions, and environmental context (Geirhos et al., 2020; Bordes et al., 2023). This lack of robustness often stems from a model's tendency to exploit spurious correlations, such as localized texture statistics or specific background cues, rather than learning the invariant geometric structure of the 3D objects they are designed to recognize (Geirhos et al., 2022; Xiao et al., 2020). The ImageNet-1K dataset (Russakovsky et al., 2015) serves as the standard training and evaluation split for many models, but lacks the diversity in viewpoints and environmental factors required to foster true geometric invariance (Alcorn et al., 2019; Geirhos et al., 2020; Barbu et al., 2019; Xiao et al., 2020).

A primary obstacle in diagnosing these failures is the lack of programmatic control provided by traditional benchmarks (Barbu et al., 2019). While natural image datasets capture real-world complexity, they do not allow the isolation of individual failure modes by manipulating single nuisance factors such as viewpoint, material, illumination, background, and fog (Geirhos et al., 2020). Synthetic images generated via diffusion models, while visually diverse, frequently lack the precise geometric and physical control required to rigorously test a model's spatial invariants (Zhang et al., 2024; Malik et al., 2024). Consequently, there is an urgent scientific need for an evaluation framework that provides causal granularity, enabling systematic variation of specific scene parameters while holding all other variables constant (Bordes et al., 2023). Unreal Engine 5 (UE5), with its advanced real-time rendering technologies such as Lumen and Nanite, offers a powerful solution by enabling the synthesis of photorealistic images with absolute control over parameters, thereby further closing the visual gap between synthetic and natural data (Bordes et al., 2023; Bommasani et al., 2021).

Evaluating modern vision systems also requires a nuanced understanding of the semantic hierarchy of the visual world (Silla and Freitas, 2011). Traditional classification metrics, which rely on exact fine-grained leaf class matches, may systematically underestimate a model's underlying semantic stability when a domain shift occurs (Dünkel et al., 2025). A central objective of this research is to determine whether models that fail to identify a specific ImageNet subclass can still maintain *semantic stability* at a broader superclass level, such as the ShapeNet taxonomy (Silla and Freitas, 2011). By investigating this gap, we seek to answer whether the observed fragility of VFMs is primarily a loss of fine-grained precision or a more fundamental failure to recognize core object geometry (Dünkel et al., 2025).

Furthermore, this study aims to isolate the causal mechanisms of model failure by subjecting these architectures to a controlled factorial design of five nuisance factors: camera viewpoint, object material, illumination, background context, and atmospheric fog (Bordes et al., 2023). Through this approach, we investigate the specific extent to which individual environmental perturbations, as opposed to an entangled domain gap, contribute to the degradation of a model's latent representations (Geirhos et al., 2020). Of particular interest is whether certain perspectives, such as top-down views, or surface variations, like non-natural materials, represent universal blind spots for current architectures regardless of their training data volume (Alcorn et al., 2019).
This thesis investigates representational boundaries through a new synthetic image rendering pipeline created in UE5. By using automated mesh-scale normalization and systematic variation of factors, this approach enables the creation of new benchmarks with detailed control over parameters. Two primary evaluation sets were created: the *Multi-Factor Dataset*, which isolates the effects of viewpoint, object material, illumination, background context, and atmospheric fog, and the *Multi-Color Dataset*, which investigates color bias by replacing object materials with uniformly colored plastic. Using a triple evaluation strategy that includes logits, k-nearest neighbors (k-NN), and linear probing, this work quantifies the robustness of a diverse suite of architectures and identifies universal failure modes.

The remainder of this work is organized into eight chapters. Chapter 2 provides the theoretical background on VFMs and their evaluation, as well as the mathematical foundations of distribution shifts, and some insights on UE5. Chapter 3 reviews the different taxonomies of VFMs, related work in synthetic benchmarking, and identifies current research gaps in representation bias. Chapter 4 details the technical setup of the UE5 rendering pipeline and the semantic alignment of ShapeNet to ImageNet. Chapter 5 establishes the evaluation methodology and the metrics used. Chapter 6 presents the empirical results, including factor breakdowns and scaling trajectories. Chapter 7 discusses the implications of our findings, focusing on texture and context bias. Chapter 8 addresses the limitations of the current framework and proposes avenues for future research. Finally, Chapter 9 concludes the thesis by synthesizing the core contributions of this research.

2 Theoretical Background

This chapter establishes the theoretical framework necessary for diagnosing the limitations in robustness of modern vision systems. We begin by examining the VFM paradigm, discussing how reliance on large-scale self-supervision impacts representational stability. Next, we delve into the concept of robustness itself and provide a diagnostic basis for the Distribution Shift taxonomy, with a specific focus on covariate shifts related to physical factors. We then investigate how the hierarchical detail level affects evaluation metrics. Finally, we explore the core technologies of

Unreal Engine 5, Lumen and Nanite, explaining how these advancements facilitate the photorealistic imagery required for our evaluation.

2.1 The Vision Foundation Model Paradigm and Evaluation Strategies

In this work, we focus on *VFMs*, a paradigm shift in artificial intelligence characterized by large-scale models trained on broad, heterogeneous data that serve as a versatile basis for diverse downstream applications (Bommasani et al., 2021). Unlike traditional computer vision models designed for single tasks, VFMs are defined by their ability to find recurring, generalizable patterns in web-scale data without relying on expensive, task-specific human annotation (Bommasani et al., 2021; Awais et al., 2025). Their capabilities are driven by two structural properties: the massive scale and diversity of pre-training data, and their adaptability to specific tasks through mechanisms such as fine-tuning, linear probing, or zero-shot prompting (Bommasani et al., 2021; Oquab et al., 2024).
To evaluate vision backbones across diverse training paradigms, researchers utilize various evaluation strategies:

Logits-Based Classification As defined by Hendrycks and Gimpel (2018), logits-based evaluation assesses model confidence using the raw outputs before applying the softmax function, with the Maximum Softmax Probability commonly used to identify errors and Out-of-Distribution (OOD) samples. This approach relies on the observation that correctly classified inputs often have higher maximum softmax probabilities compared to misclassified or OOD examples. By extracting the largest probability from the softmax output, it becomes possible to flag instances where the model is less certain. Small changes in logits can notably influence the resulting probabilities, making this method sensitive and practical for error detection.

k-NN Probing Caron et al. (2021) note that k-NN probing evaluates the quality of learned features by classifying data points according to their closest neighbors in the feature space. When the learned representations are strong, this method achieves high classification accuracy, making it a useful benchmark for representation quality. k-NN results are commonly used to assess and compare models, and the protocol is valued for its simplicity and effectiveness without requiring additional training.

Linear Probing Linear probing assesses the semantic organization of learned representations by training a simple linear classifier on top of fixed model features (Cherti et al., 2023; Bommasani et al., 2021; Siméoni et al., 2025). This approach is widely adopted because it requires minimal computation compared to full end-to-end fine-tuning but still reveals the extent of class information encoded within the penultimate representations (Cherti et al., 2023; Oquab et al., 2024). Strong results from linear probing indicate that the model's feature space is linearly

separable, suggesting that visual information has been organized into coherent clusters where semantic details are readily available for downstream tasks (Oquab et al., 2024; Radford et al., 2021).

Historically, computer vision already embraced supervised pretraining on ImageNet as a precursor to the foundation model paradigm. The shift toward foundation models arises from replacing curated supervision with large-scale self-supervised or multimodal learning, thereby lessening reliance on expensive annotation and potentially improving generalization (Bommasani et al., 2021).

2.2 Robustness and Domain Shifts

Robustness measures how reliably a model performs when inputs are noisy, incomplete, adversarial, or drawn from a distribution different from the one seen during training (Yang et al., 2024; Miyai et al., 2025). Despite impressive benchmark results, modern vision models frequently exhibit fragility in deployment, a property that traces to three root causes: overfitting to training-specific patterns that do not generalize, insufficient diversity in the training data relative to the range of scenarios encountered in production, and dataset biases that induce skewed or unstable predictions (Teney et al., 2023; Hendrycks and Gimpel, 2018; Prabhu et al., 2024; Yang et al., 2024). Addressing these failure modes requires both a principled theoretical account of what constitutes a distribution shift and a rigorous empirical framework for measuring and comparing model behavior under such shifts (Yang et al., 2024; Prabhu et al., 2024).

2.2.1 Theoretical Foundations and Distributional Shifts

The foundational paradigm of visual recognition has historically relied on the *closed-world assumption*, which posits that the test-time data distribution is independent and identically distributed (i.i.d.) relative to the training distribution, referred to as In-Distribution (ID) (Yang et al., 2024; Miyai et al., 2025). Real-world deployment, however, necessitates robustness against *OOD* samples, where the joint distribution $P(X, Y)$ over the input space $\mathcal{X}$ and label space $\mathcal{Y}$ undergoes significant shifts (Yang et al., 2024; Hendrycks and Gimpel, 2018). Robust systems must identify and adjust their predictions for unknown examples based on their current knowledge (Hendrycks and Gimpel, 2018; Miyai et al., 2025).

Distribution shifts are categorized based on which component of $P(X, Y)$ varies:

1. **Covariate Shift:** A change in the distribution $P(X)$ without changing the conditional $P(Y \mid X)$. This encompasses domain variations such as changes in lighting, viewpoint, texture, or background, where the semantic identity of the depicted object is preserved (Yang et al., 2024; Miyai et al., 2025).

2. **Semantic Shift:** A fundamental alteration in the label space $\mathcal{Y}$, where test samples belong to categories not represented in training ($P(\mathcal{Y}_{\text{ID}}) \neq P(\mathcal{Y}_{\text{OOD}})$). The detection of semantic shift is the primary focus of generalized OOD detection (Yang et al., 2024; Miyai et al., 2025).

This thesis is primarily concerned with robustness to covariate shift, specifically under controlled variations in object appearance introduced through synthetic rendering. Semantic shift and generalized OOD detection are discussed below to situate this work within the broader literature.

2.2.2 Taxonomy of Generalized OOD Detection

To unify the disparate sub-fields addressing the open-world challenge, the *Generalized OOD Detection* framework organizes five related problems along four criteria: distribution shift type, the nature of ID data (single-vs. multi-class), necessity of ID classification, and learning setting (inductive vs. transductive) (Yang et al., 2024; Miyai et al., 2025).

1. **Anomaly Detection:** Detects deviations from a predefined notion of normality. *Sensory AD* addresses covariate shift within a fixed semantic context, while *Semantic AD* targets novel categories arising from label shift (Yang et al., 2024; Miyai et al., 2025).

2. **Novelty Detection:** Identifies test samples not belonging to any training category. ND is a binary task and does not require classification among known ID classes (Yang et al., 2024; Miyai et al., 2025).

3. **Open Set Recognition (OSR):** Requires a classifier to simultaneously assign correct labels to known classes and reject unknown open-set inputs, typically without access to auxiliary OOD data during training (Yang et al., 2024; Miyai et al., 2025).

4. **OOD Detection:** Canonically similar to OSR but encompasses a broader solution space, including methods such as *Outlier Exposure*, where models are regularized using auxiliary real or synthetic OOD datasets (Yang et al., 2024).

5. **Outlier Detection:** A transductive task in which the model identifies samples differing significantly from the majority of a given, potentially contaminated, observation set (Yang et al., 2024; Miyai et al., 2025).

Miyai et al. (2025) note that Vision-Language Models (VLMs) such as CLIP have substantially blurred the boundaries between these sub-tasks, as their open-vocabulary inference capabilities partially subsume problems that were historically treated separately. Under the updated Generalized OOD Detection v2 framework, the most practically demanding challenges have converged on OOD detection and

anomaly detection as the two central open problems (Miyai et al., 2025). The combination of these challenges indicates that assessing models like CLIP should focus on evaluating their semantic stability across a wide range of open-world scenarios rather than isolated tasks. This research examines the model's ability to adapt to the practical requirements of OOD detection in real-world settings.

2.2.3 Misspecification and the ID–OOD Performance Trade-off

A widely held assumption is that ID accuracy serves as a reliable proxy for OOD generalization: models that perform better on the training distribution are expected to generalize better under shift. Teney et al. (2023) systematically challenge this view, demonstrating that ID and OOD performance are not uniformly positively correlated and can in fact be *inversely* correlated on real-world datasets.

According to Teney et al. (2023), this phenomenon is rooted in *misspecification*. The Empirical Risk Minimization objective, by rewarding any feature predictive of the training labels, incentivizes models to exploit *spurious correlations*, patterns that are predictive within the ID distribution but misleading under shift. This could be a specific background texture co-occurring with a class only in the training set. Theoretically, incorporating an additional spurious feature can decrease ID risk while increasing OOD risk, producing an inverse correlation between the two. Teney et al. (2023) propose a spectrum of ID/OOD behavioral regimes as a function of shift severity: mild shifts tend to produce positive correlation; moderate shifts produce *underspecification*, where multiple models achieve comparable ID accuracy but diverge markedly on OOD data; severe shifts yield an inverse correlation (*misspecification*); and extreme shifts result in near-zero transfer.
Evaluating models solely on ID accuracy is thus ineffective for ensuring robustness. Achieving optimal OOD performance may necessitate a strategic compromise with ID accuracy. This thesis presents and builds an evaluation framework that measures shifted performance separately from standard ID results, aiming to tackle the shortcomings of traditional benchmarking methods.

2.2.4 Benchmarking and Evaluation Protocols

The evaluation of distribution shift has progressively moved from static, fixed-size datasets towards more dynamic protocols designed to prevent community-level overfitting to benchmark idiosyncrasies (Prabhu et al., 2024).

Modern evaluation settings can be organized into three categories (Miyai et al., 2025):

1. **Near and Hard OOD:** Benchmarks in which ID and OOD classes are semantically related or derived from the same large-scale dataset (e.g., *ImageNet-10/20*, *ImageNet-X*). These represent the most challenging regime, with cur-

rent state-of-the-art models frequently failing to exceed 80% AUROC on the hardest splits (Miyai et al., 2025).

2. **Lifelong Benchmarks:** Ever-expanding test pools (e.g., *Lifelong-CIFAR10*, *Lifelong-ImageNet*) that continuously incorporate diverse domains, diffusion-generated samples, and web-queried data to maintain benchmark representativeness over time (Prabhu et al., 2024). Prabhu et al. (2024) introduce the Sort & Search framework to address the associated computational cost, drastically reducing evaluation time across many models by adaptively selecting informative sub-sample test instances.

3. **VLM-specific Benchmarks:** Tasks such as Unsolvable Problem Detection evaluate whether Vision-Language Models can correctly withhold answers when presented with mismatched or unanswerable question–image pairs (Miyai et al., 2025).

While these emerging protocols reflect a shift toward more dynamic and semantically complex OOD challenges, they frequently rely on natural images or generative models that lack the causal granularity necessary to isolate specific physical failure modes. This thesis addresses this limitation by situating our photorealistic UE5-based benchmarks within the Hard OOD regime, providing the programmatic control required to systematically diagnose the spatial and textural invariants of foundation models that traditional benchmarks often leave entangled.

2.3 Semantic Hierarchy and Evaluation Metrics

Evaluating VFMs requires a nuanced understanding of semantic relationships between heterogeneous datasets. While ImageNet-1K serves as the standard for training and evaluating VFMs, creating our own robustness dataset requires 3D models that do not always conform to the ImageNet hierarchy (Lang et al., Retrieved January, 2026; Russakovsky et al., Retrieved April, 2025, 2015).

2.3.1 ImageNet

ImageNet is constructed upon the backbone of the WordNet noun taxonomy, organizing images into a dense semantic hierarchy, grouping semantically overlapping images together (Deng et al., 2009). In this structure, object classes are represented as nodes in a semantic tree where parent nodes represent general categories and child nodes represent fine-grained subclasses (Deng et al., 2009; Bostock, 2018). For example, the general concept of a `chair` serves as a parent to specific ImageNet leaf classes such as `office chair`, `folding chair`, and `barber chair`.
ImageNet-1K serves as a popular benchmark for training and evaluating modern computer vision models (Lang et al., Retrieved January, 2026; Russakovsky et al., Retrieved April, 2025, 2015). By expanding the scope of previous popular benchmarks like PASCAL VOC from 20 object classes to 1,000 and image counts from

roughly 20,000 to over 1.4 million, ImageNet-1K provided the data density to transition from hand-tuned feature pipelines to large-scale convolutional neural networks (CNNs) that learn visual representations directly from raw image data (Deng et al., 2009; Russakovsky et al., 2015). Beyond simple classification, ImageNet-1K has standardized performance metrics for single-object localization and object detection, forcing models to learn sophisticated feature representations across a vast array of viewpoints, scales, and background clutter (Deng et al., 2009; Russakovsky et al., 2015). Nevertheless, this scope is restricted to natural images sourced via manual collection, a process that inevitably introduces class imbalance and semantic skew into the benchmark.

2.3.2 ShapeNet

ShapeNet is a large-scale dataset containing various 3D models organized in different categories. These categories are similarly organized under the WordNet taxonomy, ensuring broad compatibility with ImageNet and other multimedia data (Chang et al., 2015). However, while ImageNet-1K typically predicts fine-grained leaf classes, ShapeNet categories like `car`, `plane`, or `chair` often correspond to coarser, real-world object classes (Chang et al., 2015; Bostock, 2018). To bridge this gap, we created a one-to-many semantic mapping to align ShapeNet categories with their corresponding ImageNet subclasses.
ShapeNet represents the most comprehensive, freely available dataset containing a vast diversity of 3D objects in a large-scale, multi-category repository for this thesis (Chang et al., 2015). Alternative sources would involve significant financial costs to purchase high-quality models or require an impractical amount of time to search for and curate single models from disparate sources manually.

2.3.3 Hierarchical Evaluation Metrics

Traditional strict class accuracy can be viewed as insufficient for evaluating modern foundation models on rendered datasets for several reasons. A label-space mismatch can occur when comparing ImageNet-1K logits directly to ShapeNet ground-truth labels because ShapeNet provides broader labels than ImageNet's specific subtypes. More importantly, treating a semantically valid subcategory prediction as an error (e.g., a `sports car` prediction when the ground truth prediction is `car`) misrepresents the model's true object recognition capabilities. Thus, measuring robustness as a single aggregated value fails to capture the full capabilities, as model rankings can vary significantly with the severity of the shift.

A robust hierarchical evaluation protocol can assess whether performance degradation under a shift is simply a loss of fine-grained precision or a significant breakdown in broader semantic understanding.
Additionally, the act of predicting a semantically related object, such as an `office chair` in place of a `folding chair`, represents a different categorization error com-

pared to the prediction of an entirely unrelated item. Recent research into continuous nuisance shifts suggests that models often degrade gracefully rather than failing suddenly, implying that evaluation should identify specific failure points along a spectrum of variation rather than relying on binary success or failure (Dünkel et al., 2025). In conclusion, a hierarchical evaluation protocol can offer a model-agnostic diagnosis of VFMs, indicating a loss of fine-grained precision or a fundamental breakdown in broader semantic understanding for occurring shifts.

2.4 Unreal Engine 5 for Synthetic Data Generation

The adoption of UE5 in computer vision research represents a significant shift from previous versions, such as UE4, by offering a substantially higher degree of visual realism and more efficient workflows (Gutbier, 2025; Karis et al., 2021; Epic Games, Retrieved May, 2026b, 2025, Retrieved May, 2026c). Unlike earlier synthetic pipelines that often struggled with realism, UE5 provides photorealism out of the box through its core technologies, making it ideal for generating densely labeled datasets for robustness evaluation (Gutbier, 2025; Karis et al., 2021; Epic Games, Retrieved May, 2026b, 2025, Retrieved May, 2026c).

Core Technologies for Photorealistic Rendering The photorealistic ability to render ShapeNet meshes under diverse conditions relies on two new UE5 technologies. Lumen is a fully dynamic global illumination solution (see Figure 1). It traces thousands of microrays per pixel to deliver multi-bounce indirect lighting (Epic Games, Retrieved May, 2026b; Gutbier, 2025; Karis et al., 2021). With real-time updates in geometry or light sources, researchers can dynamically animate environmental factors, such as sun angle or light intensity, without the time-consuming process of baking lightmaps (Gutbier, 2025; Epic Games, Retrieved May, 2026b). Nanite is a virtualized micropolygon geometry system replacing fixed Level of Detail (LOD) meshes with a hierarchical cluster structure, allowing for film-quality assets comprised of millions of polygons to be rendered without noticeable fidelity loss (Karis et al., 2021; Epic Games, Retrieved May, 2026b). By keeping GPU work proportional to screen pixels, Nanite enables the seamless integration of complex 3D meshes, such as those from ShapeNet, without the need for manual optimization or traditional LOD authoring (Gutbier, 2025; Karis et al., 2021; Epic Games, Retrieved May, 2026b). Lastly, Virtual Shadow Maps complement Nanite and Lumen, providing soft, plausible shadows by intelligently streaming only the detail that is perceivable to the camera, largely removing previous constraints on draw cells and poly counts (Epic Games, Retrieved May, 2026b; Karis et al., 2021).

Programmatic Control and Generation Efficiency UE5 utilizes an actor-component paradigm, where every placeable entity is an *actor* that serves as a container for components like geometry meshes, lights, and virtual sensors (Gutbier,

Figure 1: Screenshot of "Fantasy Ruins Demo UE5 - Standard Pipeline vs Nanite/Lumen RTX 4080 Graphics Comparison" (Cycu5, 2024) taken at *15:03*. Standard Lighting on the left and UE5 Lumen on the right.

2025). This modularity is essential for experimental setups that require precise, programmatic manipulation of generative factors (Gutbier, 2025). While Blueprint visual scripting allows for the rapid prototyping of scene logic, a native C++ workflow is often required for high-throughput data generation (Epic Games, Retrieved May, 2026c; Gutbier, 2025). Gutbier (2025) demonstrated that migrating control logic to a C++ event-driven state machine can increase throughput by a factor of 16.6 compared to standard reference implementations. This efficiency allows for the generation of tens of thousands of photorealistic images with precise, automated metadata export for every frame.

3 Related Work

Building upon the theoretical foundations of distribution shifts established in the previous chapter, this section reviews the current landscape of model robustness and synthetic benchmarking. It examines how the vision community has transitioned from static, real-world test sets to high-fidelity, semantically controllable environments such as PUG-ImageNet to diagnose specific model failures. Furthermore, this chapter provides a detailed taxonomy of evaluated vision architectures, specifically examining the pre-training objectives and architectural inductive biases of models like CLIP and the DINO family. Additionally, we analyze existing literature on representation biases, specifically the tendency of deep neural networks to prioritize local textures over global geometry.

While natural adversarial benchmarks like ImageNet-A are effective at identifying broad recognition failures, there is still a significant research gap in understanding how DINOv3's hierarchical spatial features and CLIP's open-vocabulary priors respond to isolated physical factors. By contextualizing our work within the landscape of severe distribution shifts such as ObjectNet and ImageNet-A, this chapter identifies the remaining gaps in fine-grained semantic robustness evaluation that this thesis aims to address.

3.1 A Taxonomy of Evaluated Vision Architectures

For the purpose of this thesis, understanding the internal mechanics of the chosen VFM models is essential, as their robustness to domain shift is fundamentally tied to their pre-training objectives and architectural inductive biases (Geirhos et al., 2020). We have selected a heterogeneous suite of the following models: ViT, Swin, Hiera, CLIP, DINO, DINOv2, and DINOv3 to represent the current landscape of vision research for the following strategic reasons.
Firstly, to investigate how spatial processing affects robustness. For that, we compare standard global attention (ViT) against hierarchical and windowed attention architectures (Swin, Hiera) (Dosovitskiy et al., 2021; Liu et al., 2021; Ryali et al., 2023). This allows us to test if the local-to-global priors of hierarchical models provide different stability under geometric shifts than purely global transformers. Secondly, the inclusion of CLIP (language-supervised contrastive learning) and the DINO family (self-supervised instance discrimination) enables a comparison between two dominant methods of acquiring semantic knowledge (Radford et al., 2021; Oquab et al., 2024). This comparison determines whether language-aligned or purely visual self-supervision fosters a stronger *shape-bias* over *texture-bias*.

Vision Transformer (ViT) The Vision Transformer (ViT) serves as the primary baseline for architectures with minimal architectural inductive bias. ViT applies a standard Transformer encoder directly to sequences of image patches, departing from the convolutional paradigm (Dosovitskiy et al., 2021). Unlike convolutional networks, ViT processes images as sequences of non-overlapping patches and computes self-attention globally from the very first layer (Dosovitskiy et al., 2021). ViT is a critical subject of study because its lack of built-in spatial hierarchy requires the model to learn 3D geometric structure entirely from data statistics rather than having it enforced by architectural design (Dosovitskiy et al., 2021; Liu et al., 2021). Consequently, ViT underperforms comparably sized ResNets on mid-scale datasets but matches or surpasses convolutional baselines at sufficient pretraining scale (Dosovitskiy et al., 2021).

Swin Transformer and Hiera To contrast the global nature of ViT, hierarchical transformers such as Swin and Hiera were included to evaluate the impact of spatial regularization. The Swin Transformer introduces a local-to-global prior through shifted window attention and a hierarchical patch-merging pyramid (Liu et al., 2021). Similarly, Hiera achieves spatial inductive bias through a streamlined architecture optimized via Masked Autoencoder pre-training (Ryali et al., 2023). These models are essential for determining if structural constraints, specifically local windowing and hierarchical downsampling, provide better stability under geometric shifts than purely global architectures (Liu et al., 2021; Ryali et al., 2023).

Contrastive Language–Image Pre-training (CLIP) CLIP represents the multimodal, language-supervised paradigm, which replaces fixed categorical labels with

joint visual-textual representations learned from free-form textual descriptions collected at web-scale (Radford et al., 2021). This design enables open-vocabulary recognition: at inference time, downstream classification tasks are formulated as text matching problems (Radford et al., 2021). Class labels are converted into natural language prompts (e.g., a photo of a {class}), embedded via the text encoder, and compared to the image embedding using cosine similarity (Radford et al., 2021). The predicted class corresponds to the highest similarity score, enabling zero-shot transfer without task-specific fine-tuning (Radford et al., 2021). In this research, CLIP is utilized to investigate whether semantic priors derived from natural language descriptions help ground a model in an object's global form rather than its local texture (Oquab et al., 2024). By comparing CLIP against purely visual self-supervised models, we can identify whether language alignment facilitates a stronger "shape bias" that remains robust even when surface materials are synthetically overridden (Radford et al., 2021; Geirhos et al., 2022).

The DINO Family (v1, v2, and v3) The DINO (self-*di*stillation with *no* labels) series serves as the cornerstone for investigating object-centric representations. DINOv1 is notable for the emergence of unsupervised object segmentation (Caron et al., 2021). In this approach, a student network is optimized to replicate the output probability distribution of a teacher network. The teacher itself is defined by applying an exponential moving average to the student's weights over training (Caron et al., 2021).
DINOv2 builds upon the discriminative self-supervised learning methods of DINO and iBOT by combining image-level and patch-level objectives within a student-teacher framework (Oquab et al., 2024). Additionally, it builds on DINO with large-scale curated pre-training, producing frozen features that currently represent the state of the art for OOD robustness (Oquab et al., 2024). Notably, DINOv2 features perform competitively even without fine-tuning, reinforcing the foundation model paradigm in vision: pretrained representations can be used as-is across tasks (Oquab et al., 2024).
Finally, DINOv3 is included to explore the boundaries of hybrid learning objectives and scaling behavior (Siméoni et al., 2025). It extends this line of work by further scaling model capacity, training data, and optimization strategies, while refining the self-distillation framework. DINOv3 employs ViT architectures and incorporates improved initialization schemes and normalization strategies to ensure stable optimization at very large model sizes (Siméoni et al., 2025). These changes enable the model to improve scaling across model and data dimensions, increase robustness to distribution shifts, enhance patch-level consistency and structural coherence, and improve transfer to dense prediction and open-world recognition tasks (Siméoni et al., 2025). Qualitative analyses like the Principal Component Analysis (PCA) indicate that DINOv3 features encode emergent semantic part structure and foreground-background separation, despite the absence of explicit supervision (Siméoni et al., 2025). The inclusion of this latest iteration of the DINO family enables this thesis to evaluate whether the most recent advancements in large-scale self-supervision and hybrid

learning objectives yield a more robust and physically grounded representational baseline than its predecessors.

3.2 Robustness Evaluation Datasets Overview

The evaluation of neural network robustness spans various dimensions of distribution shift, utilizing different image modalities to isolate specific model vulnerabilities:

1. **Photorealistic Unreal Graphics (PUG) Family:** Introduced by Bordes et al. (2023), these datasets measure factor-specific robustness (e.g., pose, lighting, size, and texture) and compositional reasoning for vision-language models. They utilize synthetic images rendered via UE5, allowing for perfect semantic control and disentanglement of environmental factors.

2. **ImageNet-C:** Hendrycks and Dietterich (2019) benchmark corruption robustness against 75 common visual corruptions (e.g., blur, noise, weather), tracking the stability of predictions across subtle temporal sequences. It utilizes altered real images created by applying algorithmic distortions to the original ImageNet validation set.

3. **ImageNet-A:** ImageNet-A measures robustness to natural adversarial real-world images that exploit classifier blind spots without synthetic modification (Hendrycks et al., 2021b). It consists of real, unmodified natural images collected via adversarial filtration.

4. **ImageNet-R and ImageNet-Sketch:** ImageNet-R evaluates generalization to abstract visual styles (e.g., paintings, embroidery, toys), while ImageNet-Sketch specifically targets domain shifts toward superficial patterns (Hendrycks et al., 2021a; Wang et al., 2019). They utilize real natural images (renditions) and altered/natural sketches to probe a model's reliance on texture over shape.

5. **ObjectNet:** Borrowing controls from physical sciences, Barbu et al. (2019) measure robustness to random rotations, backgrounds, and viewpoints. It consists of newly photographed real-world objects captured in controlled but randomized environments to ensure models cannot exploit trivial correlations.

6. **ImageNet-V2:** Recht et al. (2019) evaluate the generalization of models to a new split created by strictly following the original ImageNet collection protocol. It uses new real images from the same source to test if models have overfit to the original validation set.

7. **ImageNet-Hard:** Taesiri et al. (2023) construct ImageNet-Hard by filtering real natural images from multiple existing robustness benchmarks. This dataset probes fundamental recognition failure and spatial/zoom biases, containing images that remain unclassifiable by state-of-the-art models even after hundreds of zoom attempts.

8. **ImageNet-D:** This benchmark targets shared perception failures by mining synthetic images that deceive multiple surrogate models (Zhang et al., 2024). It utilizes synthetic images generated by steering diffusion models with natural language to place objects in unusual contexts (e.g., a bench in a swimming pool).

9. **ImageNet-9:** Xiao et al. (2020) designed this dataset to measure model robustness against object background variations. It utilizes altered real images created by taking foreground objects from the original ImageNet and placing them onto new backgrounds.

10. **ImageNet-Mini and Tiny-ImageNet:** These are smaller-scale versions from the original ImageNet dataset (Russakovsky et al., Retrieved April, 2025).

While these popular datasets capture real-world complexity, their lack of factor-wise control necessitates a shift toward the high-fidelity synthetic environments discussed in the following section.

3.3 Synthetic Datasets for Robustness Evaluation

The advancement of synthetic datasets for robustness evaluation has moved beyond simple image augmentations toward high-fidelity, semantically controllable environments that enable precise diagnosis of model failures (Bordes et al., 2023; Idrissi et al., 2022). This transition is driven by the need to understand how specific, often entangled factors interact with neural networks.

Photorealistic Unreal Environments (PUG) Building on the standard established by Bordes et al. (2023), the family of PUG datasets represents a new standard in synthetic data, leveraging UE5 to achieve visual realism comparable to cinematic CGI. Unlike earlier synthetic benchmarks that lacked visual fidelity, PUG environments enable the precise disentanglement of environmental factors and object assets. PUG-ImageNet serves as a robustness test set for models pretrained on ImageNet. It provides granular annotations for 151 object classes across systematic variations in factors like camera yaw, lighting, and object scale. To probe the compositional reasoning of VLMs, PUG: SPAR introduces a progressive evaluation scheme to identify specific failure modes, such as a model's inability to distinguish spatial relations, such as left and right, in complex scenes. This is supported by PUG: AR4T, which provides a large training set of $200K$ images for fine-tuning VLMs on attributes and relations.
Building on this design philosophy, Gutbier (2025) introduced a modular UE5 plugin that extends these capabilities to a broader set of everyday ImageNet objects while providing more detailed metadata for each rendered scene.

Semantic Context Manipulation: Object Composure Malik et al. (2024) evaluate model resilience against complex context variations with the ObjectCompose framework by introducing an automated methodology for generating diverse object-to-background compositional change. While previous efforts like LANCE or ImageNet-E often resulted in the distortion of object semantics, ObjectCompose conditions the diffusion process on accurate object boundaries to help retain the object's original identity in the generated result. This is achieved by integrating the Segment Anything Model for mask generation and BLIP-2 for visual-textual descriptions. The resulting datasets, ImageNet-B and COCO-DC, allow researchers to quantify the role of background context through natural shifts, such as color and texture changes, and adversarial background perturbations crafted by optimizing latent embeddings. Evaluations using ObjectCompose reveal significant vulnerabilities in standard classifiers, showing an average performance drop under natural shifts and a significant drop when exposed to adversarial backgrounds.

Diffusion-Based Synthetic Datasets Recent breakthroughs in denoising diffusion probabilistic models have enabled the generation of high-fidelity synthetic data that can both improve and stress-test vision models (Azizi et al., 2023; He et al., 2023). Azizi et al. (2023) demonstrate that large-scale text-to-image models like Imagen can be fine-tuned to produce class-conditional samples that boost ImageNet classification accuracy across architectures. Similarly, He et al. (2023) show that utilizing GLIDE for generative data augmentation can improve zero-shot recognition results by an average of 4.31% across diverse datasets. Beyond augmentation, diffusion models are used to extract perception failures to create challenging benchmarks like ImageNet-D (Zhang et al., 2024). By steering Stable Diffusion with natural language, ImageNet-D generates objects in diverse representations, such as unusual materials and backgrounds, causing accuracy drops of up to 60% for standard classifiers and foundation models like CLIP (Zhang et al., 2024).

Advantages of 3D Rendering for Robustness Analysis While generative models provide high-fidelity visuals, 3D rendering engines offer unique scientific advantages for experimentation (Bordes et al., 2023; He et al., 2023). According to Bordes et al. (2023), rendering allows for the precise manipulation of individual factors (e.g., exact camera yaw or light intensity) while holding all other variables constant, enabling precise attribution of model failures. Rendering also provides easy access to granular labels and captions for rare events or specific distribution shifts that are difficult to find in scraped datasets. Researchers can test a model's invariance by changing only one factor at a time, such as changing an animal's texture while keeping its pose and background identical. Lastly, synthetic rendering avoids the privacy, bias, and copyright concerns inherent in large-scale web-scraping. While PUG established the potential for this approach, Gutbier (2025) demonstrated that these pipelines can be made more accessible by operating entirely within the game engine, eliminating the need for external proprietary infrastructure or complex Python wrappers.

3.4 Context, Texture, and Viewpoint Bias.

The systematic variation of viewpoint, material, background, light color, and fog is essential for isolating the specific failure modes of modern vision models, as standard benchmarks often reveal performance drops without pinpointing the underlying causal factors (Idrissi et al., 2022). By employing a controlled factorial design, researchers can independently vary nuisance factors to evaluate a model's true robustness against domain shifts. These are factors that do not alter the object's identity (Geirhos et al., 2020; Bordes et al., 2023; Zhang et al., 2024; Malik et al., 2024). This approach is grounded in the observation that deep neural networks (DNNs) frequently rely on spurious correlations and local cues rather than a global understanding of object geometry (Xiao et al., 2020).

Viewpoint variation Viewpoint variation is a critical parameter because vision models demonstrate very high sensitivity to object orientation, often failing to recognize familiar objects when presented in unusual poses (Alcorn et al., 2019). Sensitivity tests have shown that disturbances in rotation, specifically pitch, roll, and yaw, can lead to median failure rates exceeding 80% for high-confidence classifiers like Inception-v3 (Alcorn et al., 2019). These failures occur because standard training sets are often biased toward canonical perspectives, leaving viewpoints like the top or back underrepresented (Geirhos et al., 2020; Idrissi et al., 2022). Rendering objects from the five principal axes (front, back, left, right, and top) provides comprehensive coverage of the viewing sphere while maintaining a statistically balanced design.

Object Materials The choice to vary object materials is justified by the bias for textures in both CNNs and ViTs, where models prioritize local textural statistics over global shape cues (Idrissi et al., 2022; Geirhos et al., 2022). Because local textures are often sufficient to achieve high accuracy during training, models adopt these shortcuts and consequently fail when applied to objects with OOD surface properties (Xiao et al., 2020; Geirhos et al., 2022). Researchers can systematically investigate failure modes by rendering objects with various textures, from natural to semantically misleading. This method keeps the object's geometry constant, allowing for a clear evaluation of whether a model genuinely recognizes an object's shape or is simply reacting to its surface reflectance.

Background Variation Background variation is necessary to evaluate the extent to which models depend on environmental context for recognition. Evidence suggests that standard image classifiers do not just use but often require class-consistent backgrounds for successful classification (Xiao et al., 2020). Furthermore, selecting backgrounds in an adversarial manner can cause standard models to misclassify up to 87.5% of images (Xiao et al., 2020). By introducing diverse, photorealistic environments, such as indoor rooms, natural vegetation, or open landscapes, researchers

can quantify contextual robustness and ensure that the model's decision-making process is anchored to the foreground object rather than the surrounding scene (Malik et al., 2024).

Light Color In addition, the systematic adjustment of illumination and light color is necessary to evaluate a model's learned invariance to environmental lighting conditions. Research indicates that changes in illumination significantly impact model performance, revealing a lack of robustness to variations in shading, shadows, and reflections (Zhang et al., 2024). Studies recording the most frequent classes for objects under different lighting settings found a median intersection over union score of only 47.10%, suggesting that a model's internal class rankings are highly susceptible to lighting shifts (Alcorn et al., 2019). By varying directional and ambient light intensities, ranging from bright to dark settings, and incorporating different light sources, such as artificial home lighting versus natural light sources, researchers can simulate realistic domain shifts (Alcorn et al., 2019). Probing these factors is crucial because standard vision models often rely on specific illumination cues encountered during training, rather than the intrinsic properties of the object itself.

Fog The inclusion of atmospheric effects, specifically fog, is justified by the documented vulnerability of deep learning models to common environmental corruptions (Malik et al., 2024). Vision-based systems frequently fail when subjected to weather-related distribution shifts that were underrepresented in the training data, such as snow, frost, and fog (Malik et al., 2024). Benchmark evaluations on datasets like ImageNet-C have shown that standard architectures, such as vanilla ResNet-50, exhibit significant performance degradation when encountering fog, with corruption error rates reaching as high as 66.1 (Geirhos et al., 2022). A dedicated fog setting allows for a controlled assessment of how atmospheric attenuation and reduced visibility interfere with feature extraction and object-to-background composition (Malik et al., 2024).

3.5 Limitations of Current Robustness Benchmarks

Despite their utility, current robustness benchmarks face significant limitations that complicate the development of universally robust vision systems.

A primary limitation of traditional synthetic datasets is the simulation-to-reality gap, as early renderings often lacked the photorealism necessary for broad representation learning (Bordes et al., 2023). While modern engines like Unreal Engine address this, generative synthetic data often lacks quality control, sometimes ignoring conditioning prompts or introducing privacy concerns by replicating training data (Bordes et al., 2023; He et al., 2023). Furthermore, models trained solely on synthetic data frequently underperform compared to those trained on real data, requiring significantly more samples to achieve comparable efficiency (He et al., 2023).

PUG-ImageNet isolates specific shifts, such as camera pose or light intensity, to measure their independent effects on recognition performance (Bordes et al., 2023). However, Gutbier (2025) identified that the PUG framework relies on incomplete external Python wrappers and offers limited metadata annotation. To address this, he developed a native UE5-based pipeline that not only provides exhaustive variable control but also increases rendering performance, making large-scale synthetic OOD evaluation more computationally feasible.

Natural datasets suffer from the entanglement of factors, making it difficult to determine if a model fails due to pose, background, or lighting variations (Bordes et al., 2023). Many natural adversarial benchmarks like ImageNet-A and ObjectNet have been found to harbor severe spatial biases, where simply zooming into the center of the image de-clutters the scene and artificially boosts accuracy (Taesiri et al., 2023; Hendrycks et al., 2021b; Barbu et al., 2019).

Generative context manipulation frameworks often distort object semantics or lack the guidance necessary to preserve original object appearances during background editing (Malik et al., 2024). Additionally, most robustness datasets capture model failures but fail to explain why they occur (Idrissi et al., 2022). Diagnostic tools like ImageNet-X address this but reveal a new limitation: robustness interventions often have spill-over effects, where improving robustness to one factor (e.g., color via jittering) can inadvertently degrade robustness to another (e.g., pose) (Idrissi et al., 2022). Finally, the field lacks consensus because no single method consistently improves robustness across all distribution shifts, necessitating multi-faceted evaluation protocols (Hendrycks et al., 2021a).

4 Dataset and Rendering Setup

This section details the UE5 pipeline used to synthesize photorealistic datasets under controlled, factor-wise distribution shifts.

4.1 Experimental Overview

The experimental pipeline begins with selecting and semantically aligning ShapeNet object classes with the ImageNet hierarchy. Then, photorealistic renderings of objects under systematically varied nuisance factors are generated. Afterward, the images are classified using various VFMs to evaluate their robustness while considering hierarchical structures and to provide a detailed comparison with other state-of-the-art robustness datasets.

Figure 2 illustrates the dataset creation pipeline. First, 3 different 3D object meshes are selected from each of the 55 ShapeNet categories (Chang et al., 2015). Then we select our desired factor variations. Afterwards the objects are rendered in

UE5 under our variations in viewpoint, material, illumination, and background. The rendering pipeline builds upon prior synthetic data generation systems such as PUG (Bordes et al., 2023) and subsequent adaptations (Gutbier, 2025). Additionally, our pipeline creates segmentation masks for every created image.
The rendered images are then evaluated using the mentioned VFMs (see Section 3.1). These models produce semantic predictions corresponding to ImageNet classes, which can be organized in a hierarchical structure derived from WordNet (Deng et al., 2009; Silla and Freitas, 2011; Bostock, 2018). To account for semantic relationships between object categories, evaluation is performed at two hierarchical levels: superclass correctness, corresponding to ShapeNet category alignment, and exact ImageNet class correctness.

This controlled rendering approach enables a systematic analysis of model robustness for objects under domain shifts affecting appearance while preserving the semantic identity (Bordes et al., 2023; Zhang et al., 2024; Malik et al., 2024). Unlike natural image benchmarks containing uncontrolled variations, our synthetic rendering approach enables precise manipulation of individual nuisance factors, providing valuable insights into robustness limitations and representation biases in modern vision systems.

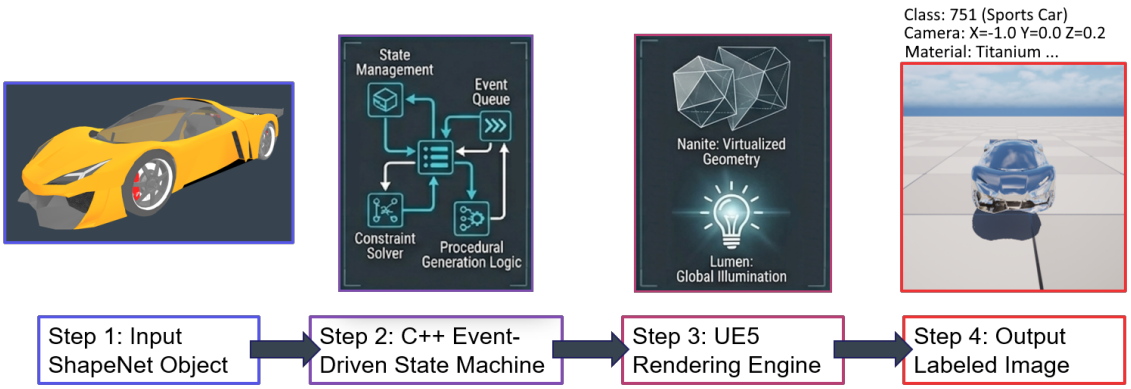


Figure 2: Dataset creation pipeline. Rendered 3D car model from Chang et al. (2015).

4.2 ShapeNet Class Subset and Semantic Alignment

To ensure semantic compatibility with ImageNet-trained models, object classes are selected from ShapeNetCore (Chang et al., 2015), which provides richly annotated 3D models organized into semantic categories. ShapeNet categories correspond closely to real-world object classes like `car`, `plane`, and `chair`.

Each ShapeNet category is mapped to one or more corresponding ImageNet classes based on the WordNet semantic hierarchy underlying ImageNet (Deng et al., 2009; Bostock, 2018). This hierarchical organization represents object classes as nodes in a

semantic tree, where parent nodes correspond to more general categories and child nodes correspond to fine-grained subclasses. Hierarchical classification has been widely studied in machine learning as a means to capture semantic relationships between classes and improve evaluation interpretability (Silla and Freitas, 2011). For example, the ShapeNet category `chair` corresponds to multiple ImageNet subclasses, including `office chair`, `folding chair`, and `barber chair`. Predictions to any of these subclasses are considered semantically correct at the superclass level, while only exact matches are considered correct at the fine-grained level.

This hierarchical mapping allows for the evaluation of whether domain shifts primarily impact fine-grained recognition or overall semantic understanding. However, there are two exceptions to this rule. The ShapeNet classes `bicycle` and `wine bottle` are not part of the ImageNet-1K class list, so they have been excluded from the metrics.

4.3 Synthetic Dataset Generation

Synthetic images are generated using UE5, a real-time rendering engine capable of producing photorealistic imagery. The rendering pipeline is adapted from prior work on photorealistic synthetic dataset generation (Gutbier, 2025).

All materials used are based on license-free stock photos or freely available online. All background imagery was obtained from the FAB Marketplace and selected according to three criteria (Epic Games, Retrieved May, 2026a). They are free, are distributed under FAB's standard license, and the provider permits use of these assets for AI-related research.

The *Dataset Renderer* plugin from Gutbier (2025) constitutes the baseline Unreal Engine rendering pipeline employed in this work. It is implemented as a runtime plugin that enables the generation of synthetic datasets through controlled scene composition and systematic variation of rendering parameters. Static meshes are loaded from a CSV-based mapping and placed within an UE5 scene. The system then iterates through predefined combinations of camera positions, light colors, materials, and fog settings. For each configuration, the plugin captures rendered images and records the corresponding metadata in a structured CSV file. In addition, the pipeline incorporates automatic texture-streaming synchronization to ensure that relevant level and object textures are fully resident before image acquisition. Optional mesh-scale normalization further contributes to the consistency and usability of the generated data. Overall, the plugin provides a reproducible and controllable framework for synthetic data generation under photorealistic rendering conditions.

Our plugin extends this baseline pipeline with several adaptations tailored to the experimental requirements of this thesis. In particular, the capture process was restructured into explicit phases, separating variations in camera, lighting, material,

and fog into distinct rendering stages. Moreover, the pipeline was augmented to distinguish between a default background and additional background-variation levels, thereby enabling the background to be a normal factor variation. The adapted version also preserves and restores the scene's original lighting and material state, thereby preventing unintended accumulation of rendering changes across capture runs.
To maintain consistent visual proportions across the different objects, the pipeline implements an adjustable Object-Frame-Percentage parameter. For the following dataset generation, the value was set to 60%. This approach offers adequate padding for transformer-based patch extraction while also maximizing foreground resolution. It ensures that all objects occupy no more than 60% of the rendered frame's height or width, allowing us to minimize the spatial biases observed in some other benchmarks, such as ImageNet-A (see Section 3.5). The parameter remains fully editable within the plugin configuration, enabling future research into the effects of varying object-to-frame ratios on model zoom bias and scale invariance.
The output directory structure was reorganized into factor-specific subfolders to facilitate downstream analysis. A new component of this capture process is the automated generation of a black-and-white (binary) segmentation mask for each rendered photorealistic image. These masks provide precise, pixel-level ground truth by isolating the foreground ShapeNet object geometry from the background environment, thereby significantly enhancing the consistency and usability of the generated data for evaluating spatial attribution and dense prediction tasks in future research. These modifications refine the baseline system into a more structured, experiment-oriented dataset-generation pipeline.

Each object is consistently rendered multiple times independently under all the different nuisance factors predefined in the plugin.
The rendering pipeline operates automatically, loading ShapeNet meshes, applying predefined materials, configuring lighting and background environments, and rendering images from predefined camera viewpoints. All rendered images are stored along with metadata specifying the rendering configuration, enabling precise evaluation of robustness with respect to individual factors.

We generated two datasets: one containing all described factor variations (Multi-Factor Dataset) and the other containing specifically object color variations (Multi-Color Dataset).

4.4 Domain Shift Factors

To evaluate robustness under realistic appearance variations, five types of domain shifts are introduced during rendering: viewpoint, material, illumination, background, and fog. These factors were selected based on prior work demonstrating their strong impact on vision model robustness (Geirhos et al., 2020; Bordes et al., 2023; Zhang et al., 2024; Malik et al., 2024).

As the thesis goal is robustness evaluation under controlled nuisance factors, we want to ensure a balanced design with equal statistical power across factors, unbiased performance estimates, and valid factor-wise comparisons. To achieve that, we select the same number of variations for each domain shift.

4.4.1 Viewpoint Variation

Objects are rendered from five viewpoints corresponding to the principal axes of the object coordinate system: front, back, left, right, and top. This configuration provides almost complete coverage of the viewing sphere while maintaining a balanced experimental design. With the object being close to the ground and not levitating in the air to avoid impacting performance, the view from under the object was omitted. Using axis-aligned viewpoints ensures interpretable and statistically balanced evaluation while isolating viewpoint sensitivity as an independent factor.

4.4.2 Material Variation

Objects are rendered using multiple physically based materials, including both naturalistic and artificial textures. Material variation alters surface reflectance and texture statistics while preserving object geometry.
To systematically investigate these failure modes, we have selected materials that vary across several categories: natural versus artificial, matte versus specular, structured versus structureless, and semantically inconsistent. The chosen materials consist of baseline options such as wood; high-frequency artificial textures like procedural noise; metallic materials including titanium; transparent and refractive materials such as glass; and semantically misleading materials, which encompass camouflage patterns. This approach enables a thorough examination of the various failure modes.

4.4.3 Background Variation

Objects are rendered in multiple photorealistic environments using high-dynamic-range maps. Background variation alters contextual information and environmental appearance while preserving object identity.

To provide an evenly distributed environment, we selected multiple different backgrounds:

1. Indoor Home Environment
    (a) Living Room (combines artificial lights and natural lighting)
    (b) Basement (darker artificially lighted wooden room)

2. Natural Vegetation Environment (rocky desert with plants)
3. Natural Vegetation Environment (forest hills)
4. Open Landscape Environment (plain desert)

4.4.4 Illumination Variation

Lighting conditions are varied using different environment maps and illumination parameters. Illumination variation alters object appearance through changes in shading, shadows, and reflection.
We have chosen five different lighting colors: red, green, blue, yellow, and dark dim, to encompass most of the RGB spectrum.

4.4.5 Fog

The fog factor is implemented as a binary environmental condition, where atmospheric occlusion is either completely disabled (serving as the control) or enabled at a fixed intensity throughout the rendered scene. By leveraging UE5's volumetric fog capabilities, we simulate atmospheric attenuation to evaluate how reduced visibility affects the model's ability to extract foreground features from the background.

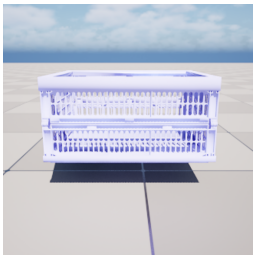
4.4.6 Dataset Statistics

The final dataset consists of a balanced factorial combination of our factor variations. Each rendering configuration is represented equally, ensuring unbiased evaluation of individual factors (see Figure 3).
All rendered images are stored together with metadata specifying object identity, rendering parameters, and semantic class mappings. This enables precise analysis of robustness with respect to individual domain shift factors and semantic hierarchy levels.

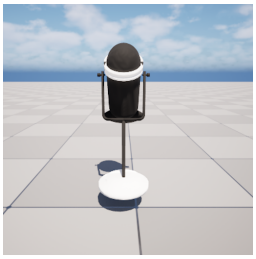
4.4.7 Multi-Color Dataset

For the second dataset, we changed the render pipeline to apply all materials for all viewpoints. We created a matte-worn plastic material for the main color spectrum (green, cyan, blue, pink, red, yellow) as well as black and grey. These materials were applied to all objects for the default background on all camera perspectives mentioned in 4.4.1.

The rendering pipeline was extended to apply all material variants across all viewpoints, enabling a dedicated analysis of model robustness with respect to object coloration. Unlike the one-factor-at-a-time design employed in the first dataset, this

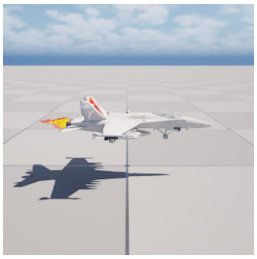


(a) Basket Base View



(b) Microphone Base View

(c) Jet Fighter Base View



(d) Jet Left Side View

(e) Jet Desert Background

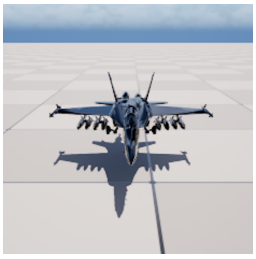
(f) Jet Red Lighting

(g) Jet Camouflage Material

(h) Jet Fog Enabled

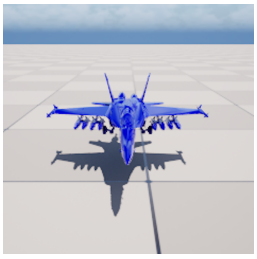
Figure 3: Qualitative examples from the Multi-Factor Dataset. Row 1 shows various contained objects and one exemplary viewpoint. Row 2 demonstrates exemplary environmental and texture variations including background changes, lighting shifts, camouflage patterns, and atmospheric fog. Zoomed in for better visibility.

dataset follows a full factorial design, rendering every object under every combination of material and camera perspective. These material instances were applied to all objects rendered against the default Unreal Engine background and the full set of camera perspectives defined in Section 4.4.1 (see Figure 4).

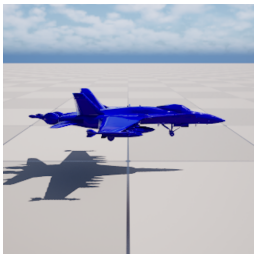


(a) Jet Black Plastic

(b) Jet Green Plastic



(c) Jet Blue Plastic



(d) Jet Blue Plastic Left Side View

Figure 4: Qualitative examples from the Multi-Color Dataset containing exemplary plastic colors and one exemplary viewpoint. Zoomed in for better visibility.

4.5 Dataset Comparisons

To systematically contextualize the difficulty of our proposed dataset and ensure the broad applicability of our findings, we integrate our evaluation into a wider ecosystem of established distribution-shift benchmarks defined in Section 3.2.
A critical technical distinction in our setup involves preprocessing protocols. The PUG family employs a specialized pipeline, scaling assets based on bounding box dimensions before rendering or, in some cases, cropping and then resizing, while the standard ImageNet preprocessing involves resizing the image's smaller dimension to 256, followed by a center-crop to 224 (Bordes et al., 2023). We utilized the standard preprocessing. By anchoring our synthesized distribution shifts against these diverse and well-documented benchmarks, we establish a baseline difficulty spectrum, mapping the severity of our parametrically controlled variations directly onto the known landscape of OOD robustness datasets.

5 Evaluation Methodology

This section details a comprehensive multi-tiered evaluation protocol designed to systematically quantify fine-grained discriminative power, broad semantic stability, and structural equivariance in pretrained vision backbones. Evaluating models solely on standard, identically distributed test sets often masks critical failure modes under domain shift or environmental perturbations. Therefore, our framework measures performance across two semantic granularities: exact ImageNet-1K classification and hierarchy-aware ShapeNet superclass identification. By subjecting a heterogeneous model suite to both novel procedurally generated datasets and established benchmarks, we establish a comparative framework to verify pipeline fidelity and empirically rank datasets by relative difficulty.

Because our model tracking includes both supervised classifiers and pure feature extractors, we adopt a triple evaluation strategy. For models equipped with pretrained ImageNet-1K classification heads, we measure robustness via zero-shot logit predictions. For self-supervised models lacking a task-specific projection head, we standardize evaluation using a k-NN protocol on the extracted representation-space features, ensuring a fair comparison of the underlying representation quality. Finally, we trained a classifier head for a few selected models to demonstrate how linear probing performs compared to the out-of-the-box alternatives.

Crucially, by subjecting this identical model suite to both our novel dataset and the established ImageNet variants, we establish a two-fold comparative framework:

1. *Verification of Reliability*: By comparing our models' top-1 accuracy on standard benchmarks (e.g., ImageNet-1K, ImageNet-A) to values reported in the existing literature, we verify the fidelity of our inference pipeline, feature extraction, and class-mapping procedures.

2. *Empirical Difficulty Ranking*: By aggregating the performance degradation of the standard model suite across all datasets, we empirically rank the datasets by relative difficulty. This ranking provides a quantitative measure of how the isolated variations in our dataset compare in severity to the complex semantic shifts present in datasets like ObjectNet or ImageNet-A. Ultimately, this allows us to rigorously demonstrate where our dataset sits on the spectrum of modern robustness challenges.

5.1 Evaluated Models

We evaluate a representative set of large-scale VFMs spanning convolutional and transformer-based architectures, including CLIP, DINO (v1, v2, v3), Swin, ViT, Hiera, and ResNet-50 as a baseline comparison. These models represent various training paradigms, including contrastive multimodal and self-supervised learning. We exclude generative or instruction-tuned models to avoid prompt-dependent variability and focus on controlled visual recognition robustness. These models are trained on broad, web-scale data and exhibit strong transfer capabilities, thereby satisfying the defining characteristics of foundation models (Awais et al., 2025).

To ensure a scientifically sound comparison between models with native classification heads and pure feature extractors, we adopt a comprehensive triple evaluation strategy:

1. **Logits-Based Evaluation**: For models equipped with pretrained ImageNet-1K heads (e.g., CLIP-B-IN1K, DINOv2-B-IN1K, ResNet-50-IN1K), we measure robustness via zero-shot direct logit predictions using their native 1000-way classification heads.

2. **Embedding-Based (k-NN) Evaluation**: For self-supervised models lacking a task-specific head (e.g., DINOv1-B, DINOv3-B), we standardize evaluation using a non-parametric k-NN protocol on extracted features. We utilize $\ell_2 - normalized$ embeddings and a distance-weighted voting scheme. For higher efficiency, the backend is implemented via FAISS. With this approach we can measure the true semantic quality of the underlying representation space.

3. **Linear Probe**: To further assess the linear separability of the learned representation spaces, we train a supervised linear classifier on top of the frozen extracted features. This approach provides a parametric counterpart to the k-NN evaluation, allowing us to evaluate the extent to which the raw semantic features can be linearly mapped to class labels without end-to-end fine-tuning.

Our goal with this multi-tiered approach is to isolate the intrinsic robustness of the backbone from any potential biases present in the classification head. By comparing frozen features using k-NN against supervised fine-tuned (SFT) logits, we can determine whether the geometric information remains in the representation space, even

when a model's final predictions seem to fail. Additionally, we use linear probing as a middle-ground diagnostic tool to evaluate the linear separability of these features, while avoiding the representational distortion that often occurs with full end-to-end fine-tuning.
Table 1 summarizes the model families, their source of supervision, and the evaluation mode used in the framework.

Model key	**Backbone / pretrained source**	**Mode**
CLIP-IN1K	CLIP ViT with ImageNet-1K classifier head	logits, linear
CLIP	CLIP ViT representation model	k-NN
DINOv2-IN1K	DINOv2 ViT with ImageNet-1K classifier head	logits, linear
DINOv2	DINOv2 ViT representation model	k-NN
DINOv1	DINOv1 ViT representation model	k-NN
DINOv3	DINOv3 ViT representation model	k-NN
Swin-IN1K	Swin with ImageNet-1K classifier head	logits, linear
Swin	Swin representation model	k-NN
ViT-IN1K	ViT with ImageNet-1K classifier head	logits
ResNet-50-IN1K	ResNet-50 with ImageNet-1K classifier head	logits
ResNet-50	ResNet-50 representation model	k-NN
Hiera-IN1K	Hiera with ImageNet-1K classifier head	logits

Table 1: Backbones evaluated in the study. Models with a native ImageNet-1K head are reported in logits mode. Representation-only models are evaluated through frozen features using nearest-neighbor classification and linear probing.

This model set is deliberately heterogeneous. Convolutional baselines such as ResNet-50 provide a classical reference point, while transformer-based models such as CLIP, DINOv2, Swin, ViT, and Hiera capture the behavior of more recent large-scale pretrained architectures. The comparison is therefore not restricted to a single training recipe or a single architectural family. Instead, the benchmark measures how robustly each pretrained representation transfers to the evaluation corpus under exact-class and superclass-level criteria.
Except for DINOv2 (518x518) and DINOv3 (256x256), all other chosen models require images with 224×224 pixel inputs. Depending on the backbone (Small, Base, Large), the parameter count ranges from $22M$ (DINOv2-S) to $50M$ (Swin-S), and $52M$ (Hiera-B) to $149M$ (CLIP-B), and $197M$ (Swin-L) to $420M$ (CLIP-L).

5.2 Label Hierarchy and Class Mapping

A primary challenge in cross-domain evaluation is the semantic granularity gap. Our evaluated models are predominantly pre-trained on the highly granular ImageNet-1K dataset ($\mathcal{Y}_{IN}$, containing 1,000 specific synsets), whereas our procedurally generated 3D assets derive from the broader taxonomic categories of ShapeNet ($\mathcal{Y}_{SN}$, representing 55 high-level object superclasses).

To structurally bridge this gap without penalizing models for selecting semantically valid subsets, we formalize a taxonomic mapping function $\mathcal{M}$. For each ShapeNet superclass target $c \in \mathcal{Y}_{SN}$, we define a set of valid ImageNet child classes $V_c \subset \mathcal{Y}_{IN}$. This mapping utilizes the WordNet hierarchy to handle subcategory aggregations. Formally, let $\mathcal{Y}_{\text{IN}}$ denote the set of 1000 ImageNet classes and let $\mathcal{Y}_{\text{SN}}$ denote the set of ShapeNet superclasses. We define a mapping

$$\phi : \mathcal{Y}_{\text{IN}} \to \mathcal{Y}_{\text{SN}} \cup \{\varnothing\},$$

where $\phi(y)$ returns the ShapeNet superclass associated with ImageNet class y, and $\varnothing$ indicates that the class is not mapped. The mapping is constructed offline from a curated correspondence file between ShapeNet categories and ImageNet class indices. During preprocessing, each sample is augmented with both the human-readable ImageNet label and the ShapeNet superclass assigned by ϕ.

This mapping is many-to-one in the sense that multiple ImageNet labels may collapse into a single ShapeNet superclass. For example, different bottle-related or bed-related ImageNet synsets are grouped together because they correspond to the same coarse semantic object category. ImageNet labels that do not belong to any ShapeNet superclass and are marked as unmapped. These samples are excluded from ShapeNet-level scoring but remain valid for ImageNet-level evaluation.
The mapping is computed once and stored in the metadata so that evaluation can be performed without any runtime semantic lookup. This keeps the metrics deterministic and allows the same label transformation to be reused across models and datasets. Table 2 shows a representative excerpt of the label hierarchy.

5.3 Hierarchy-Aware Evaluation Protocol

The evaluation protocol has two complementary goals. First, it measures whether the model predicts the exact ImageNet class label. Second, it measures whether the model preserves the broader ShapeNet superclass when exact classification fails. This dual view is important because pretrained vision models often learn semantically useful representations even when they confuse fine-grained subclasses within the same object family. This way, we can determine whether the model is degrading gracefully, retaining the coarse semantic family, or completely failing.

Exact ImageNet class correctness For each sample i, let $y_i \in \mathcal{Y}_{\text{IN}}$ be the ground-truth ImageNet label and let $\hat{y}_i$ be the model prediction. Exact correctness is defined by

$$c_i^{\text{IN}} = \mathbb{1}[\hat{y}_i = y_i].$$

The corresponding ImageNet top-1 accuracy over N samples is

$$A^{\text{IN}} = \frac{1}{N} \sum_{i=1}^{N} c_i^{\text{IN}}.$$

ShapeNet superclass	Synset	ImageNet indices	Representative ImageNet labels
airplane	n02691156	404, 405, 895	airliner; warplane, military plane; airship, dirigible
bag	n02773838	414, 636, 728, 748, 797	bag; backpack, knapsack, rucksack; plastic bag; mailbag; purse
basket	n02801938	790, 519, 588	basket; shopping basket; crate; hamper
bathtub	n02808440	435	bathtub, bathing tub, bath, tub
bed	n02818832	431, 520, 564, 831	bed; bassinet; four-poster; studio couch, day bed; crib, cot
bench	n02828884	703	bench; park bench
bicycle	n02834778	444, 671	bicycle; bike; tandem bicycle; mountain bike
birdhouse	n02843684	448	birdhouse
bookshelf	n02871439	453	bookshelf; bookcase
bottle	n02876657	440, 441, 720, 737, 898, 899, 901, 907	bottle; beer bottle; wine bottle; water bottle; pop bottle; pill bottle; whiskey jug
bowl	n02880940	659, 809	bowl; soup bowl; mixing bowl
bus	n02924116	654, 779, 874	bus; school bus; minibus; trolleybus
cabinet	n02933112	495, 553, 648	cabinet; file cabinet; medicine cabinet; china cabinet
camera	n02942699	732, 759	camera; Polaroid camera; reflex camera
can	n02946921	653	can; milk can

Table 2: Representative excerpt of the ImageNet-to-ShapeNet mapping used for hierarchy-aware evaluation.

When the model outputs ranked probabilities, we also report top-5 accuracy,

$$A^{\mathrm{IN}(5)} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[y_i \in \mathrm{Top5}(\hat{p}_i)],$$

where $\hat{p}_i$ is the predicted class distribution.

ShapeNet superclass correctness Let $\phi(y_i)$ denote the ShapeNet superclass of the ground-truth label and $\phi(\hat{y}_i)$ the superclass associated with the model prediction. The sample is considered semantically correct if both belong to the same superclass:

$$c_i^{\mathrm{SN}} = \mathbb{1}[\phi(\hat{y}_i) = \phi(y_i)].$$

This criterion measures semantic robustness rather than fine-grained label fidelity. A prediction such as `beer bottle` for a `bottle` image is incorrect at the ImageNet level but can still be correct at the ShapeNet superclass level because both labels belong to the same higher-order object family.

Only samples with a valid ShapeNet superclass mapping are included in ShapeNet-level evaluation:

$$\mathcal{E} = \{i \mid \phi(y_i) \neq \texttt{unmapped}\}.$$

The ShapeNet superclass accuracy is therefore

$$A^{\text{SN}} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} c_i^{\text{SN}}.$$

Samples whose ground truth class is unmapped remain part of ImageNet evaluation but are excluded from ShapeNet-specific metrics.

Interpretation The relationship between the two levels of correctness is informative. A large gap between A_{IN} and A_{SN} indicates that the model often confuses semantically close subclasses but still recognizes the coarse category. A small gap indicates that the model is both fine-grained and semantically stable. In contrast, low accuracy at both levels suggests that the model fails to capture the object family itself rather than just the subclass identity.

5.4 Model-Agnostic Prediction Handling

To ensure uniform evaluation across heterogeneous architectures, the framework converts all model outputs into a shared prediction record with: (i) top-K ImageNet class indices and probabilities, and (ii) a mapped ShapeNet superclass prediction for top-1.

Let $I_{1000} = 0,\ldots,999$ be ImageNet class indices, and let

$$\phi : \mathcal{I}_{1000} \to \mathcal{S} \cup \texttt{unmapped}$$

map each ImageNet index to a ShapeNet superclass in S, or to $\emptyset$ if unmapped. For each sample with ground-truth ImageNet label i^*, the predicted superclass is

$$\hat{s} = \phi(\hat{i}_1),$$

where $\hat{i_1}$ is the top-1 predicted ImageNet index. Superclass correctness is only evaluated when the ground-truth superclass is mapped.

Current mapping protocol The framework currently applies direct class-to-superclass mapping with predicted ImageNet Top-1 being mapped via ϕ to ShapeNet superclass space. ShapeNet metrics are computed only for samples whose ground truth maps to a valid superclass. Should the dataset metadata provide a set of acceptable ImageNet indices (for ambiguous or grouped labels), Top-1/Top-5 ImageNet correctness is computed against that set rather than a single class index.

Logits-based models For models with native ImageNet heads, the network outputs logits $\ell(x)R^{1000}$. Probabilities are obtained by softmax:

$$p(i \mid x) = \frac{\exp(\ell_i)}{\sum_{j \in \mathcal{I}1000} \exp(\ell_j)}.$$

Top-K predictions are

$$\hat{i}1 : K = \mathrm{TopK}\, i \in \mathcal{I}_{1000}, p(i \mid x).$$

ImageNet correctness is computed either as exact-match ($\hat{i_1} = i^*$) or, if provided, membership in a valid class set. Superclass prediction is then obtained by $\hat{s} = \phi(\hat{i_1})$, and ShapeNet correctness is $\hat{s} = \phi(i^\star)$ on evaluable samples only.

Embedding-based models (k-NN mode) For representation models without classifier heads, the backbone is used as a frozen feature extractor:

$$z = f(x) \in \mathbb{R}^d, \qquad \tilde{z} = \frac{z}{|z|}.$$

Classification is then performed by k-NN in embedding space (cosine similarity via normalized inner product), with temperature-weighted voting:

$$\hat{i} = \arg\max c \in \mathcal{I}_{1000} \sum_{j \in \mathcal{N}_k(\tilde{z})} \exp! \left(\frac{\mathrm{sim}(\tilde{z}, \tilde{z}_j)}{\tau} \right) \mathbf{1}[i_j = c].$$

In current experiments, the k-NN reference bank is fixed to ImageNet-1K features. If the query dataset is ImageNet-1K itself, leave-one-out is applied (the query sample is excluded from its own neighbor list). For other datasets, no leave-one-out filtering is enforced in the active run path.
The resulting top-K ImageNet predictions are stored in the same schema as logits mode, and superclass evaluation is performed identically via ϕ.

Embedding-based models (Linear Probe mode) For representation models evaluated via linear probing, the backbone remains a frozen feature extractor to yield embeddings:

$$z = f(x) \in \mathbb{R}^d.$$

A supervised linear classifier, parameterized by a weight matrix $W \in \mathbb{R}^{1000 \times d}$ and bias $b \in \mathbb{R}^{1000}$, is trained on the ImageNet-1K training set using these frozen features. During inference, the network outputs logits via the learned linear projection:

$$\ell(x) = Wz + b.$$

(If L_2-normalized features $\tilde{z}$ were used during probe training, the projection is instead $\ell(x) = W\tilde{z} + b$). Probabilities are then computed via softmax, yielding top-K predictions identical to the logits-based protocol:

$$p(i \mid x) = \frac{\exp(\ell_i)}{\sum_{j \in \mathcal{I}_{1000}} \exp(\ell_j)}, \qquad \hat{i}_{1:K} = \mathrm{TopK}_{i \in \mathcal{I}_{1000}}\, p(i \mid x).$$

As with the other modes, the resulting ImageNet predictions are stored in the standard schema, and superclass evaluation is performed identically via ϕ.

5.5 Metrics and Multi-Factor Analysis

The benchmark metadata contains multiple nuisance factors that vary independently or semi-independently across the rendered dataset. These include camera position, lighting, fog, material, and background level. For each factor, we compute stratified accuracy so that the effect of a single perturbation can be isolated from the rest of the corpus.

Let f denote a nuisance factor and let v denote one of its values. We define the subset

$$\mathcal{D}_{f=v} = \{i \mid \text{factor } f \text{ of sample } i \text{ equals } v\}.$$

The per-factor accuracy is then

$$A_{f=v} = \frac{1}{|\mathcal{D}_{f=v}|} \sum_{i \in \mathcal{D}_{f=v}} \mathbb{1}[\hat{y}_i = y_i].$$

Analogously, a ShapeNet-level factor accuracy can be computed by replacing exact ImageNet correctness with the superclass criterion:

$$A^{\mathrm{SN}}_{f=v} = \frac{1}{|\mathcal{D}_{f=v} \cap \mathcal{E}|} \sum_{i \in \mathcal{D}_{f=v} \cap \mathcal{E}} \mathbb{1}[\phi(\hat{y}_i) = \phi(y_i)].$$

This factor-wise evaluation is essential because aggregate accuracy alone can hide strong brittleness to one particular nuisance variable. For example, a model may achieve high average accuracy but degrade sharply under fog or lighting variation, revealing a lack of robustness that would not be visible from a single global score.

5.5.1 Metric Suite

Thus, the evaluation reports a compact set of summary metrics containing the: *ImageNet top-1 accuracy*, *ImageNet top-5 accuracy*, *ShapeNet superclass top-1 accuracy*, and *per-factor accuracy*.
In addition to accuracy-based measures, the framework can report confidence-sensitive statistics such as expected calibration error and AUROC for misclassification detection. While these are not the core metrics of the methodology, they help contextualize whether a model's confidence scores are aligned with its actual correctness. This is particularly useful when comparing logits-based models to feature-based models whose confidence calibration may differ substantially.

5.5.2 Equivariance

To move beyond aggregate classification metrics, this work employs a formal mathematical framework to quantify the structural consistency of the model's latent space through equivariance analysis. This metric evaluates whether a model's internal representation responds to physical transformations in a predictable, object-independent manner.

We hypothesize that latent representation stability does not guarantee downstream generalization. By introducing this metric, we can examine whether models can maintain stable internal features despite variations in factors, while still experiencing significant drops in classification performance due to those features crossing decision boundaries.

The analysis uses transition vectors to measure how an object's representation shifts in the latent space in response to a change in a specific factor. Following the framework established in PUG (Bordes et al., 2023), for a given object i and a transition between two camera perspectives a and b, the transition vector $\Delta v_i^{(a \to b)}$ is defined as the difference between the extracted feature embeddings:

$$\Delta v_i^{(a \to b)} = \Phi(x_i^{(b)}) - \Phi(x_i^{(a)})$$

where $\Phi(\cdot)$ denotes the feature extraction function (embedding) of the model, and $x_i^{(c)}$ is the image of object i at camera perspective c.

In this study, representation equivariance is aggregated across all available objects and camera perspectives. This global aggregation ensures that the resulting metric evaluates the model's generalized response to the domain shift rather than localized artifacts specific to individual object identities. To compare these alignment vectors for pairs of distinct objects ($i \neq j$) undergoing, e.g., the exact same camera transition ($a \to b$) in identical environmental contexts, we measure via Cosine Similarity:

$$\text{Sim}_{i,j}^{(a \to b)} = \frac{\Delta v_i^{(a \to b)} \cdot \Delta v_j^{(a \to b)}}{\|\Delta v_i^{(a \to b)}\|_2 \|\Delta v_j^{(a \to b)}\|_2}$$

The final equivariance score is the expected value of these similarities over all valid object pairs and all possible camera transitions:

$$\mathbb{E}_{a,b}\left[\mathbb{E}_{i \neq j}\left[\text{Sim}_{i,j}^{(a \to b)}\right]\right]$$

To visualize the geometry of these transitions, a PCA is applied to project these high-dimensional difference vectors into a lower-dimensional space.

Interpretation of the Empirical Score The resulting scalar is bounded in the interval $[-1.0, 1.0]$:

- **1.0** implies *Perfect Equivariance*: The domain shift corresponding to camera movement manifests as a universal, object-independent translation vector in the latent space, signifying that the model responds consistently to the factor across the entire dataset.
- **0.0** implies *Orthogonality*: The embedding shifts are entirely uncorrelated and the model's response is entangled with object identity.

An empirical score of e.g., 0.15 indicates *weak representation equivariance*. Rather than learning a generalized spatial concept for a specific shift, the transformations applied by the model are highly entangled with the underlying object identity. Consequently, the two different objects' shifts propagate in a fundamentally different direction.

5.5.3 Explainability and Spatial Attribution

To move beyond aggregate metrics, we integrate a model-agnostic visual explainability framework that traces classification decisions back to specific localized image features.
We generate Saliency Maps via Class Activation Mapping (Grad-CAM), utilizing architecture-specific gradient propagation. For convolutional architectures (e.g., ResNet, ConvNeXt), the gradients are extracted from the final spatial feature map. For ViTs (e.g., ViT, Swin, DINOv2), gradients are aggregated from the final normalization layer prior to the classification token (`layernorm_before`), necessitating spatial reshaping of the flattened 1D patch tokens back into an aligned 2D grid structure.
By overlaying these resulting activation heatmaps onto the original rendered stimuli, we perform qualitative and automated localization analysis. This interpretability layer is critical for validating whether architectures that maintain semantic robustness under domain shift genuinely attend to core object geometry rather than spurious, confounding environmental cues (e.g., leveraging background sky textures to classify an airplane).

5.6 Statistical Rigor and Reproducibility

To guarantee that observed performance disparities are statistically significant and not artifacts of sampling variance, our evaluation incorporates several safeguards. We compute accuracy metrics using 1,000 bootstrap resamples to establish 95% confidence intervals. This quantifies variance in model performance and validates the statistical significance of observed gaps between architectures. While these intervals were utilized during the internal validation phase to ensure that the reported model rankings are statistically robust, we report only the point estimates (mean accuracies) in the primary results tables to prioritize visual legibility and comparative clarity. This reporting choice is justified by substantial performance gaps observed between model families, which render the typical bootstrap variance (typically within $\pm 1.5\%$) negligible for high-level architectural ranking. Furthermore, all experiments use a fully deterministic pipeline with fixed random seeds and reproducible feature extraction to ensure that any reported differences are attributable to model behavior rather than to stochastic variation in the evaluation procedure. We include classical baselines (ResNet-50) to provide a calibration point against which modern ViTs and self-supervised models are measured.
The experiments were conducted using an NVIDIA RTX A5000 GPU.

6 Results

This section presents the empirical results of our evaluation suite across the full spectrum of VFMs (see Table 1) discussed in Section 2.1 under systematically controlled rendering shifts. We first establish a baseline understanding of model performance, comparing absolute accuracies and model rankings on the Multi-Factor and Multi-Color datasets against established external OOD benchmarks. We then investigate the role of semantic hierarchy by evaluating models at different levels of classification granularity (ImageNet-1K vs. ShapeNet superclasses). Next, we examine how the choice of downstream evaluation protocol (comparing logits, k-NN classifiers, and linear probing) affects the measured robustness of frozen representation spaces. This is followed by a stratified factor-level analysis that isolates and ranks the individual rendering dimensions (camera position, materials, backgrounds, lighting, and fog) responsible for classification degradation. Finally, we analyze the impact of model scaling and architectural biases before dissecting the representational mechanisms of robustness through equivariance metrics and spatial explainability visualizations.

6.1 Overall Performance and Baseline Comparisons

We establish a high-level overview of model performance across the two primary evaluation sets introduced in this work: the Multi-Factor Dataset and the Multi-Color Dataset. Before analyzing the specific factors that drive a semantic collapse, it is necessary to quantify the absolute magnitude of the domain gap.
Table 3 and Table 4 summarize the Top-1 classification accuracy for all evaluated models on both the fine-grained (ImageNet-1K) and superclass (ShapeNet) semantic levels.

6.1.1 The Absolute Magnitude of the Domain Gap

A key finding is that all architectures and evaluation paradigms encounter clear challenges when detecting synthetic objects. On the standard ImageNet-1K validation set, foundation models like DINOv2-L and CLIP-L routinely achieve Top-1 accuracies between 80% and 90%. In comparison, on the Multi-Factor Dataset, the highest-performing model (DINOv2-L), evaluated via our k-NN approach, achieves 38.6% at the ImageNet class level and 52.6% at the ShapeNet superclass level.
Similar performance is apparent on the Multi-Color Dataset. Here, the top-performing model (DINOv3-L-KNN) achieves 36.6% fine-grained accuracy and 48.6% superclass accuracy.
The standard supervised baseline, ResNet-50, also faces notable challenges on both datasets, dropping from 74.0% to 7.6% (Multi-Factor) and 8.8% (Multi-Color) at the ImageNet level, respectively.

Model Name	IN1K Validation	IN Top-1	IN Top-5	SN Top-1
DINOv2-L-KNN	81.1%	**38.6%**	**60.6%**	**52.7%**
DINOv3-L-KNN	83.3%	38.3%	55.4%	51.3%
DINOv2-L-IN1K-LOGITS	85.9%	37.6%	59.8%	52.2%
DINOv2-L-IN1K-KNN	81.6%	36.1%	57.2%	49.2%
CLIP-L-IN1K-KNN	**87.5%**	34.4%	49.2%	45.5%
ResNet-50-IN1K-KNN	77.5%	9.2%	18.4%	13.4%
DINOv1-B-KNN	70.2%	8.7%	21.9%	13.5%
CLIP-B-KNN	67.1%	8.6%	19.6%	13.0%
ResNet-50-KNN	74.0%	7.7%	17.4%	12.0%
DINOv1-S-KNN	68.6%	7.3%	18.2%	12.0%

Table 3: Top-5 and Worst-5 Model Rankings on the Multi-Factor Dataset. IN1K is the Top 1 Accuracy for the models run on the ImageNet-1K Evaluation Split Dataset, while IN Top-1 is the Top 1 Accuracy for the models run on our Multi-Factor Dataset, and SN Top-1 is the ShapeNet Top 1 Accuracy.

Model Name	IN1K Validation	IN Top-1	IN Top-5	SN Top-1
DINOv3-L-KNN	83.3%	**36.6%**	**54.4%**	**48.6%**
CLIP-L-IN1K-KNN	**87.5%**	32.3%	45.4%	42.6%
DINOv2-L-KNN	81.1%	30.9%	52.1%	44.6%
CLIP-L-IN1K-LOGITS	87.3%	30.5%	50.9%	40.6%
DINOv2-L-IN1K-KNN	81.6%	29.9%	49.9%	42.6%
ResNet-50-KNN	74.0%	8.8%	18.8%	12.9%
ResNet-50-IN1K-KNN	77.5%	8.6%	17.6%	12.4%
CLIP-B-KNN	67.1%	8.4%	20.4%	13.5%
DINOv3-B-KNN	79.2%	8.2%	17.0%	13.0%
DINOv1-S-KNN	68.6%	7.0%	18.5%	12.3%

Table 4: Top-5 and Worst-5 Model Rankings on the Multi-Color Dataset. IN1K is the Top 1 Accuracy for the models run on the ImageNet-1K Evaluation Split Dataset, while IN Top-1 is the Top 1 Accuracy for the models run on our Multi-Color Dataset, and SN Top-1 is the ShapeNet Top 1 Accuracy.

6.1.2 Ranking Stability Across Datasets

Even with the clear challenges observed in synthetic object detection, the relative ranking of models remains remarkably stable between the two synthetic datasets (Figure 5). Large-scale self-supervised ViTs consistently occupy the top ranks. DINOv2-L and the recently released DINOv3-L dominate both benchmarks, outperforming CLIP models of comparable size.

Conversely, earlier self-supervised models (DINOv1) and smaller supervised CNNs consistently cluster at the bottom of the rankings. However, some rank inversions occur in the mid-tier models. For instance, the representation-only k-NN evaluation of DINOv2 often surpasses its ImageNet-finetuned counterpart, a phenomenon that will be further explored in depth in Section 6.4.

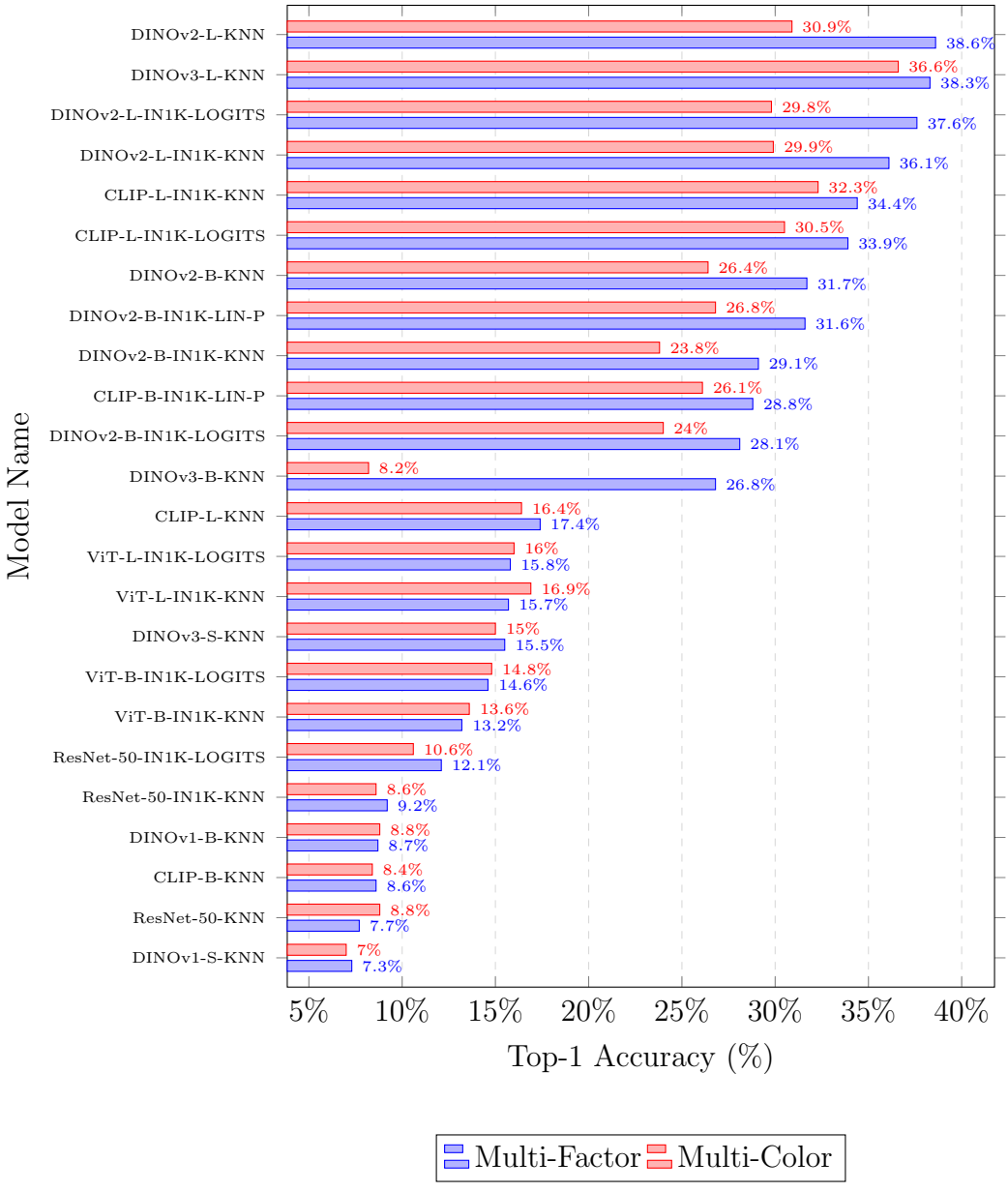


Figure 5: Comparison of ImageNet Top-1 Accuracy (%) across our two datasets for the best 12 and worst 12 models.

6.2 Impact of Domain Shift: All Datasets

Having established the overall performance landscape in the previous section, the present section situates the Multi-Factor and Multi-Color datasets within the broader ecosystem of robustness benchmarks. While Section 6.1 quantified the absolute magnitude of the domain gap, here we ask two complementary questions: *How difficult are our synthetic benchmarks relative to established OOD datasets?* And: *Do models that generalize well on standard benchmarks also generalize well on our novel shifts?* Answering these questions is essential for validating the experimental utility of our datasets before analyzing their fine-grained factor-level structure in subsequent sections.

We evaluate the same suite of model-evaluation combinations on 15 benchmarks spanning diverse distribution shifts. The standard ImageNet-1K validation set, established robustness benchmarks: ImageNet-A (Hendrycks et al., 2021b), ImageNet-Sketch (Wang et al., 2019), ImageNet-V2 (Recht et al., 2019), ImageNet-C (Hendrycks and Dietterich, 2019), ImageNet-D (Zhang et al., 2024), ImageNet-Hard (Taesiri et al., 2023), ImageNet-9 (Xiao et al., 2020), ObjectNet (Barbu et al., 2019), smaller-scale subsets (ImageNet-Mini, Tiny ImageNet), a prior synthetic benchmark (PUG-ImageNet (Bordes et al., 2023)), and our two novel datasets (Multi-Factor Dataset, Multi-Color Dataset).

6.2.1 The Robustness Difficulty Spectrum.

Table 5 ranks all 15 evaluated benchmarks by their mean Top-1 classification accuracy, averaged across all 28 k-NN model configurations. This ranking establishes a *difficulty spectrum* that reveals where our synthetic datasets sit in the landscape of vision robustness challenges.

Several observations emerge from this ranking:
First, the Multi-Factor Dataset (mean accuracy of 21.35%) and Multi-Color Dataset (18.99%) represent severe distribution shifts. They are substantially harder than most other datasets. Notably, the Multi-Factor Dataset closely matches ObjectNet (21.75%) in difficulty.
Second, the Multi-Color Dataset is even harder than the Multi-Factor Dataset (-2.36 *percentage points), despite controlling fewer variables.*
Third, DINOv3-L is consistently ranked as the best model across nine benchmarks with CLIP-L and DINOv2 being best on four and two benchmarks. On our Multi-Factor Dataset DINOv2-L performs best, while its DINOv3 on the Multi-Color Dataset.

Dataset	Mean Acc.	Min Acc.	Max Acc.	Best Model
IN-1K	79.66%	67.13%	87.53%	CLIP-L-IN1K
IN-Mi	78.21%	5.38%	88.86%	CLIP-L-IN1K
IN-C	77.94%	44.54%	96.24%	DINOv3-L
IN-V2	67.11%	52.63%	77.72%	CLIP-L-IN1K
IN-9	59.11%	43.85%	70.97%	DINOv3-L
Ti-IN	51.35%	24.44%	70.55%	DINOv3-L
IN-Ha	45.61%	19.52%	57.62%	DINOv3-L
Avg	45.51%	25.54%	61.24%	DINOv3-L
IN-Sk	41.47%	20.61%	65.43%	DINOv3-L
IN-R	41.47%	21.57%	71.71%	DINOv3-L
PUG-IN	37.46%	19.91%	56.57%	DINOv3-L
IN-A	26.57%	2.07%	62.21%	DINOv2-L
ON	21.75%	8.04%	35.05%	DINOv3-L
MF (ours)	**21.35%**	**7.32%**	**38.60%**	**DINOv2-L**
MC (ours)	**18.99%**	**6.97%**	**36.63%**	**DINOv3-L**
IN-D	14.52%	6.18%	25.44%	CLIP-L-IN1K

Table 5: Robustness benchmarks ordered by difficulty (mean Top-1 accuracy across all 28 k-NN model configurations). The horizontal rule separates moderate shifts from severe shifts. Our synthetic datasets rank among the most challenging benchmarks, on par with ObjectNet. Abbreviations: IN-1K = ImageNet-1K, IN-V2 = ImageNet-V2, IN-R = ImageNet-R, IN-Sk = ImageNet-Sketch, IN-A = ImageNet-A, ON = ObjectNet, PUG IN = PUG ImageNet, Ti-IN = Tiny-Imagenet, IN-Mi = ImageNet-Mini, IN-9 = ImageNet-9, IN-Ha = ImageNet-Hard, IN-D = ImageNet-D, Mf = Multi-Factor Dataset (ours), MC = Multi-Color Dataset (ours)

6.2.2 A Taxonomy of Distribution Shifts.

The 15 benchmarks evaluated in this work span fundamentally different types of distribution shift. To structure the analysis, we group them into four categories based on the nature of their domain gap:

1. **Near-distribution shifts** (ImageNet-V2, ImageNet-Mini, ImageNet-9, Tiny-ImageNet): Datasets that follow the original ImageNet collection methodology or represent subsets thereof. Performance drops are modest (mean accuracies 51–78%), primarily reflecting statistical variation in data collection rather than a fundamental shift in the visual domain.

2. **Artistic and stylistic renditions** (ImageNet-R, ImageNet-Sketch): Images depicting ImageNet categories in non-photographic styles such as paintings, cartoons, and pencil sketches. These benchmarks test whether models encode category-discriminative features that generalize beyond natural image statistics, with mean accuracies around 41%.

3. **Adversarial and geometric real-world shifts** (ImageNet-A, ImageNet-Hard, ObjectNet, ImageNet-D): Datasets curated to expose specific model failure modes: adversarial natural examples (ImageNet-A), challenging viewpoints and backgrounds (ObjectNet), or domain-specific environments (ImageNet-D). Mean accuracies range from 14% to 46%, reflecting the targeted severity of these shifts.

4. **Controlled synthetic shifts** (PUG-ImageNet, Multi-Factor Dataset, Multi-Color Dataset): Rendered 3D scenes with parametric control over individual factors of variation. PUG-ImageNet introduced this paradigm, but operates at a mean accuracy of 37.46%, which is significantly higher than our datasets. The Multi-Factor Dataset (21.35%) and Multi-Color Dataset (18.99%) push the synthetic evaluation paradigm into the *severe shift regime*, comparable to ObjectNet's difficulty.

6.2.3 Rank Correlation and Generalization Consistency.

A crucial validation question is whether model performance on our synthetic datasets correlates with performance on established benchmarks. If models that generalize well on standard OOD datasets also perform well on our rendering-based shifts, this supports the claim that our datasets measure general robustness properties rather than an idiosyncratic, benchmark-specific capability.

Table 6 reports Spearman rank correlation coefficients (ρ) between model performance rankings on our two synthetic datasets and all other benchmarks.

The results reveal three key patterns:
Strong correlation with OOD benchmarks. Performance on our Multi-Factor Dataset is most strongly correlated with PUG-ImageNet ($\rho = 0.893$), ImageNet-Sketch ($\rho = 0.882$), and Tiny-ImageNet ($\rho = 0.850$). Similarly, the Multi-Color Dataset shows exceptionally high correlations with ObjectNet ($\rho = 0.941$), ImageNet-A ($\rho = 0.901$), and ImageNet-D ($\rho = 0.900$).
Weaker correlation with ID data. Correlations with ImageNet-1K ($\rho = 0.596$ / 0.692), ImageNet-V2 ($\rho = 0.594$ / 0.786), and ImageNet-9 ($\rho = 0.511$ / 0.565) are notably lower.
Divergent correlation profiles for Multi-Factor vs. Multi-Color. The two datasets do not correlate identically with all benchmarks, reflecting their different shift characteristics. For example, ImageNet-C shows a moderate correlation with the Multi-Factor Dataset ($\rho = 0.653$) but a weaker correlation with the Multi-Color Dataset ($\rho = 0.485$).

6.2.4 Architecture and Pre-Training Resilience Across Benchmarks.

Table 7 provides a cross-benchmark comparison of representative models from each architecture family for an excerpt of evaluated datasets, allowing us to identify which

Dataset	Spearman ρ (Multi-Factor)	Spearman ρ (Multi-Color)
PUG-IN	0.893	0.871
IN-Sk	0.882	0.898
Ti-IN	0.850	0.840
IN-A	0.839	0.901
ON	0.831	0.941
IN-R	0.797	0.803
IN-Ha	0.733	0.598
IN-D	0.703	0.900
IN-C	0.653	0.485
IN-1K	0.596	0.692
IN-V2	0.594	0.786
IN-Mi	0.585	0.680
IN-9	0.511	0.565

Table 6: Spearman rank correlation coefficients (ρ) between model performance rankings on our synthetic datasets and established benchmarks. The horizontal rule separates OOD benchmarks (strong correlation) from near-distribution datasets (weaker correlation). High positive values indicate model performance rankings remain highly consistent. Abbreviations: IN-1K = ImageNet-1K, IN-V2 = ImageNet-V2, IN-R = ImageNet-R, IN-Sk = ImageNet-Sketch, IN-A = ImageNet-A, ON = ObjectNet, PUG IN = PUG ImageNet, Ti-IN = Tiny-Imagenet, IN-Mi = ImageNet-Mini, IN-9 = ImageNet-9, IN-Ha = ImageNet-Hard, IN-D = ImageNet-D

design and pre-training choices confer the most robust generalization.

Model	ON	IN-V2	IN-R	IN-Sk	IN-A	ON	PUG IN	MF	MC
ResNet-50-IN1K	15.2%	62.5%	21.6%	24.2%	2.1%	15.2%	24.4%	9.2%	8.6%
ViT-L-IN1K	15.7%	66.4%	34.3%	37.3%	14.0%	15.7%	30.9%	15.7%	16.9%
Swin-L-IN1K	27.1%	74.2%	46.2%	46.5%	37.1%	27.1%	44.5%	23.2%	20.5%
CLIP-B-IN1K	26.8%	74.1%	51.4%	54.5%	22.6%	26.8%	46.6%	25.7%	23.4%
CLIP-L-IN1K	32.7%	77.7%	65.7%	64.2%	38.7%	32.7%	53.7%	34.4%	32.3%
DINOv2-B	27.2%	69.5%	50.5%	52.3%	52.5%	27.2%	46.1%	31.7%	26.4%
DINOv2-L	31.6%	71.3%	61.8%	59.4%	62.2%	31.6%	52.7%	38.6%	30.9%
DINOv3-B	10.9%	53.8%	36.9%	38.6%	9.0%	10.9%	40.8%	26.8%	8.2%
DINOv3-L	35.0%	73.1%	71.7%	65.4%	58.1%	35.0%	56.6%	38.3%	36.6%

Table 7: Excerpt comparison of top-performing and baseline vision k-NN models across standard robustness datasets and our synthetic benchmarks (MF, MC). Accuracies are reported as Top-1 accuracy (%). Abbreviations: IN-1K = ImageNet-1K, ON = ObjectNet, IN-V2 = ImageNet-V2, IN-R = ImageNet-R, IN-Sk = ImageNet-Sketch, IN-A = ImageNet-A, ON = ObjectNet, PUG IN = PUG ImageNet, MF = Multi-Factor Dataset (ours), MC = Multi-Color Dataset (ours)

Self-supervised DINO models dominate under severe shifts. DINOv2-L achieves the highest accuracy on the Multi-Factor Dataset (38.6%) and the second-highest on

the Multi-Color Dataset (30.9%), while DINOv3-L leads overall with 38.3% and 36.6%, respectively. Crucially, DINOv2-L also achieves the top score on ImageNet-A (62.2%), a benchmark specifically targeting model failure modes.

CLIP offers strong but secondary robustness. CLIP-L-IN1K achieves the highest ImageNet-1K accuracy (87.5%) and leads on ImageNet-V2 (77.7%), demonstrating strong ID generalization. However, under severe shifts, CLIP consistently trails DINOv2 and DINOv3: on our Multi-Factor Dataset, CLIP-L-IN1K reaches 34.4% compared to DINOv2-L's 38.6% (−4.2pp). This gap widens on ImageNet-A, where CLIP-L-IN1K achieves 38.7% compared to DINOv2-L's 62.2% (−23.5pp).

Supervised baselines exhibit near-catastrophic failure. ResNet-50-IN1K drops to 9.2% on the Multi-Factor Dataset and 8.6% on the Multi-Color Dataset. This is an approximately 88% relative accuracy loss compared to its 77.5% ID performance. ViT-L-IN1K, despite its larger capacity and transformer architecture, fares only marginally better (15.7% / 16.9%).

The DINOv3-B outlier on the Multi-Color Dataset. Perhaps the most striking finding in Table 7 is the behavior of DINOv3-B. On the Multi-Factor Dataset, DINOv3-B achieves a respectable 26.8%, comparable to CLIP-B-IN1K (25.7%). However, on the Multi-Color Dataset, DINOv3-B collapses to 8.2%, a significant −18.6 percentage point drop that is *not* mirrored by any other model of similar ID performance. For comparison, DINOv3-L drops by only −1.7pp from 38.3% to 36.6% between the two datasets.

6.3 Superclass (ShapeNet) vs Fine-grained (ImageNet) Classification: The Role of Semantic Hierarchy

Evaluating robustness purely at a fine-grained, ImageNet-1K level often masks the true capabilities of foundation models under domain shift. When a model misclassifies a rendered plastic `airplane` as a `projectile` rather than a specific `airliner`, it fails the fine-grained metric, but it still successfully recognizes the core shape semantics. To isolate how much performance degradation is due to the loss of fine-grained textural cues versus catastrophic shape failure, we compared model accuracy at both the fine-grained and superclass levels.

Overall Accuracy Gap As shown in Figure 6, evaluating models at the ShapeNet superclass level yields significantly and consistently higher accuracy across all architectures than evaluating them at the ImageNet level. Nearly all models reside comfortably above the $y = x$ parity line. On the Multi-Factor Dataset, top-performing models like DINOv2-L and DINOv3-L achieve over 50% ShapeNet accuracy compared to their 38% ImageNet accuracy. The relative benefit of superclass evaluation is often stronger for weaker models; for instance, ResNet-50 models see their accuracy nearly double when evaluated on ShapeNet superclasses. The average superclass-to-fine-grained classification accuracy ratio per model is 1.418 for both the Multi-Factor and the Multi-Color Dataset (see Table 29).

Per-Class Variance and Vulnerabilities Not all object categories are equally robust to domain shift. Figure 7 and 8 illustrate the ShapeNet superclass accuracy across all models for both datasets. We observe distinct vertical banding, indicating that objects correctly identified by worse-performing models are also nearly always correctly identified by the better-performing models.
To quantify this, Table 8 and 9 detail the average performance across all evaluated models per category. Classes with highly `distinct`, rigid structural templates, such as `piano`, `guitar`, and `helmet`, maintain high superclass accuracy. These objects possess strong geometric signatures that survive extreme material and lighting shifts. Conversely, classes that heavily rely on texture, fine details, or context, such as `bus`, `cabinet`, and `train`, achieve low accuracy.

The Impact of the Plastic Material Shift When comparing the Multi-Factor Dataset to the Multi-Color Dataset, we observe the semantic gap narrowing. The average ShapeNet-to-ImageNet accuracy ratio drops from 1.88 on the Multi-Factor Dataset to 1.40 on the Multi-Color Dataset. While fine-grained ImageNet accuracy degrades when natural textures are removed, ShapeNet accuracy drops even more dramatically in relative terms.

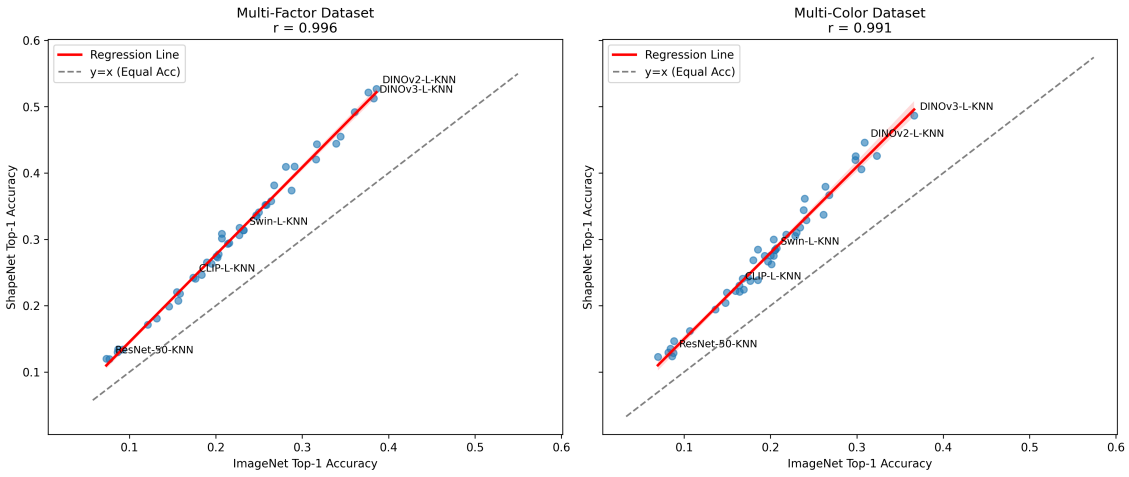


Figure 6: ImageNet Top-1 vs ShapeNet Top-1 per model for Multi-Factor and Multi-Color

6.4 Evaluation Method Comparison: Logits, k-NN, and Linear Probing

The preceding sections established *which* models are most robust to domain shifts. This section examines *how* the choice of evaluation method affects robustness under the domain shifts imposed by the Multi-Factor and Multi-Color datasets. Since different evaluation paradigms make different assumptions about what knowledge is retained in the frozen backbone, understanding the gap between them reveals

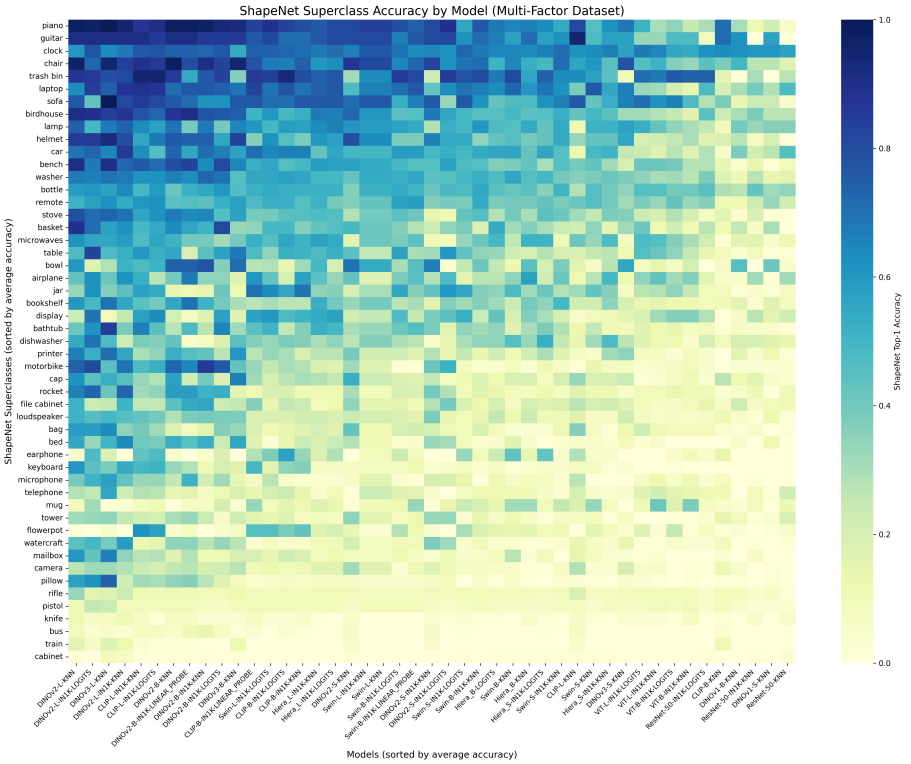


Figure 7: Model accuracy across ShapeNet superclasses (Multi-Factor Dataset).

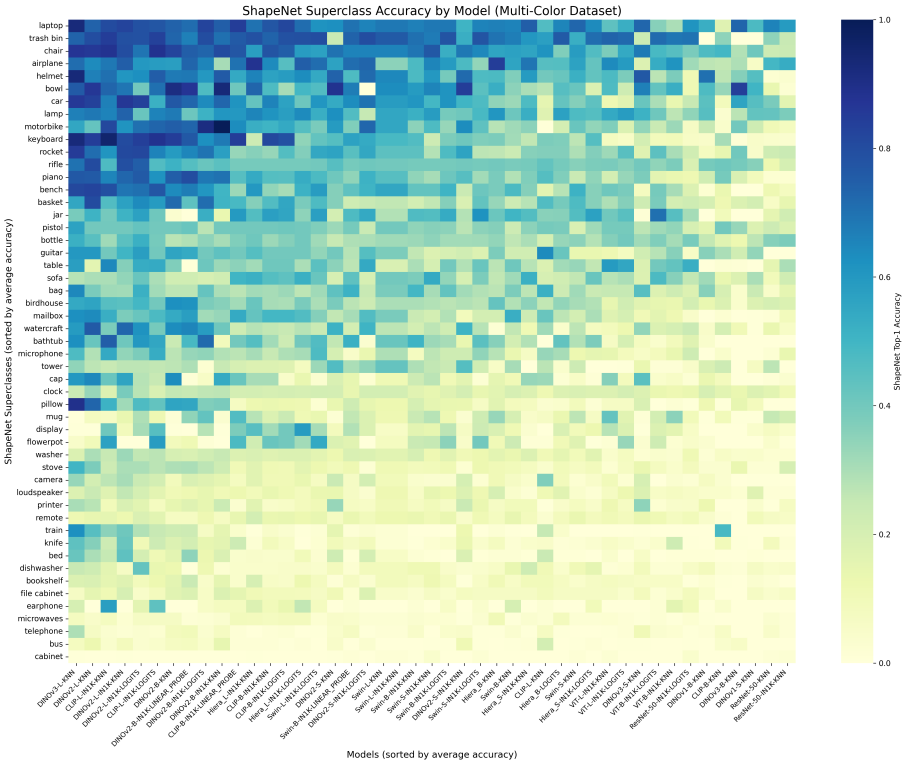


Figure 8: Model accuracy across ShapeNet superclasses (Multi-Color Dataset).

ShapeNet Superclass	Avg IN Top-1	Avg SN Top-1	Ratio (SN/IN)
piano	62.8%	69.6%	1.11x
guitar	61.8%	67.7%	1.10x
clock	28.1%	67.1%	2.39x
chair	33.4%	65.6%	1.96x
trash bin	64.5%	64.5%	1.00x
pistol	8.6%	8.6%	1.00x
knife	2.2%	3.3%	1.52x
bus	0.8%	2.4%	2.88x
train	1.6%	2.2%	1.35x
cabinet	0.1%	1.1%	15.00x

Table 8: Top-5 and Worst-5 average Top-1 ImageNet and ShapeNet accuracy by object category (Multi-Factor Dataset). Ranked by Avg SN Top-1 accuracy descending.

ShapeNet Superclass	Avg IN Top-1	Avg SN Top-1	Ratio (SN/IN)
laptop	44.0%	65.8%	1.50x
trash bin	63.1%	63.1%	1.00x
chair	36.5%	61.3%	1.68x
airplane	39.8%	54.3%	1.37x
helmet	49.4%	53.2%	1.08x
earphone	7.0%	7.0%	1.00x
microwaves	3.8%	3.8%	1.00x
telephone	1.5%	3.5%	2.30x
bus	1.4%	3.1%	2.23x
cabinet	0.0%	0.5%	N/A

Table 9: Top-5 and Worst-5 average Top-1 ImageNet and ShapeNet accuracy by object category (Multi-Color Dataset). Ranked by Avg SN Top-1 accuracy descending.

whether robustness failures originate in the learned feature space or in the classification head.
We evaluate three evaluation paradigms (Logits, k-NN, Linear Probe) across the models for which multiple methods are available (see 5.4). Table 10 presents the Top-1 ImageNet-level accuracy for all models where multiple evaluation methods are available, on the Multi-Factor and Multi-Color datasets, respectively.

Logits vs. k-NN: Classification Head or Representation Space? For models with both a direct classification head (logits) and a k-NN evaluation on the same backbone (IN1K-KNN), we can isolate whether the robustness bottleneck lies in the linear classification head or in the underlying feature space.

Model Family	Method	MF (Top-1)	MC (Top-1)
DINOv2-B-IN1K	logits	28.1%	24.0%
	IN1K-KNN	29.1%	23.8%
DINOv2-L-IN1K	logits	37.6%	29.8%
	IN1K-KNN	36.1%	29.9%
DINOv2-S-IN1K	logits	20.7%	18.5%
	IN1K-KNN	20.7%	18.0%
CLIP-B-IN1K	logits	25.9%	23.0%
	IN1K-KNN	25.7%	23.4%
CLIP-L-IN1K	logits	33.9%	30.5%
	IN1K-KNN	34.4%	32.3%
Swin-B-IN1K	logits	22.7%	19.3%
	IN1K-KNN	21.4%	20.0%
Swin-L-IN1K	logits	26.4%	21.8%
	IN1K-KNN	23.2%	20.5%
ResNet-50-IN1K	logits	12.1%	10.6%
	IN1K-KNN	9.2%	8.6%
ViT-B-IN1K	logits	14.6%	14.8%
	IN1K-KNN	13.2%	13.6%
ViT-L-IN1K	logits	15.8%	16.0%
	IN1K-KNN	15.7%	16.9%

Table 10: Logits vs. IN1K-finetuned k-NN comparison on the Multi-Factor and Multi-Color Datasets (ImageNet Top-1 accuracy). Values within 1pp of each other indicate method parity. Abbreviations: MF = Multi-Factor Dataset (ours), MC = Multi-Color Dataset (ours)

The results (Table 10) reveal a nuanced picture. For the majority of models, logits and IN1K-KNN perform within 2 percentage points of each other on both datasets. However, Self-supervised DINO models favor k-NN on the Multi-Factor Dataset and logits on the Multi-Color Dataset. For DINOv2-B-IN1K, k-NN slightly outperforms logits on the Multi-Factor Dataset (+1.0pp), but the two are near-equal on Multi-Color. For CLIP-L, k-NN outperforms logits on both datasets (+0.5pp and +1.8pp, respectively).
For ResNet-50, logits exceed k-NN by +2.9pp on Multi-Factor and +2.0pp on Multi-Color. For Swin-L, logits exceed k-NN by +3.2pp on Multi-Factor.

The Cost of Fine-Tuning: Raw k-NN vs. IN1K-KNN A central question for evaluating pre-trained vision models is whether fine-tuning on ImageNet-1K improves or harms generalization to OOD data. Table 11 compares the raw pre-

trained backbone (k-NN) against the IN1K fine-tuned variant (IN1K-KNN), so we can directly quantify the cost or benefit of downstream adaptation.

Model Family	Method	MF (Top-1)	MC (Top-1)
DINOv2-B	k-NN	31.7%	26.4%
	IN1K-KNN	29.1%	23.8%
DINOv2-S	k-NN	24.7%	20.4%
	IN1K-KNN	20.7%	18.0%
DINOv2-L	k-NN	38.6%	30.9%
	IN1K-KNN	36.1%	29.9%
CLIP-B	k-NN	8.6%	8.4%
	IN1K-KNN	25.7%	23.4%
Swin-B	k-NN	20.2%	17.4%
	IN1K-KNN	21.4%	20.0%
ResNet-50	k-NN	7.7%	8.8%
	IN1K-KNN	9.2%	8.6%

Table 11: Raw pre-trained k-NN vs. IN1K fine-tuned k-NN on the Multi-Factor and Multi-Color Datasets. Abbreviations: MF = Multi-Factor Dataset (ours), MC = Multi-Color Dataset (ours)

Across all DINOv2 sizes, the raw pre-trained k-NN *outperforms* the IN1K-finetuned k-NN on both datasets. The gap is largest for DINOv2-B: k-NN achieves 31.7% on the Multi-Factor Dataset vs. 29.1% for IN1K-KNN (−2.6pp), and 26.4% vs. 23.8% on the Multi-Color Dataset (−2.6pp).
There is a significant contrast for CLIP-B: k-NN achieves only 8.6% on the Multi-Factor Dataset, while IN1K-KNN reaches 25.7% (+17.1pp).
For Swin-B and ResNet-50, IN1K fine-tuning slightly improves k-NN performance.

Linear Probing: Adapting the Head Without Distorting the Space Linear probing occupies a methodological middle ground between the frozen-head k-NN approach and full fine-tuning of a new classification layer. For comparison, we have trained three models with this method: CLIP-B-IN1K, DINOv2-B-IN1K, and Swin-B-IN1K.

Linear probing produces consistently competitive results across all three model families (see Table 12):
For DINOv2-B-IN1K, linear probing matches k-NN. The linear probing achieves 31.6% on the Multi-Factor Dataset and 26.8% on the Multi-Color Dataset, nearly identical to the raw pre-trained k-NN (31.7% / 26.4%).
For CLIP-B-IN1K, linear probing outperforms both k-NN and logits. The linear probing (28.8% / 26.1%) exceeds IN1K-KNN (25.7% / 23.4%) and logits (25.9% / 23.0%) on both datasets.
For Swin-B-IN1K, linear probing matches or exceeds both logits and IN1K-KNN.

Model Family	Method	MF (Top-1)	MC (Top-1)
DINOv2-B	k-NN	31.7%	26.4%
	Linear Probe	31.6%	26.8%
	IN1K-KNN	29.1%	23.8%
	Logits	28.1%	24.0%
CLIP-B	Linear Probe	28.8%	26.1%
	IN1K-KNN	25.7%	23.4%
	Logits	25.9%	23.0%
	k-NN	8.6%	8.4%
Swin-B	Linear Probe	22.7%	20.7%
	IN1K-KNN	21.4%	20.0%
	Logits	22.7%	19.3%
	k-NN	20.2%	17.4%

Table 12: Four-way evaluation method comparison for models with linear probing variants. Methods are ranked by Multi-Factor Dataset Top-1 accuracy within each model family. Abbreviations: MF = Multi-Factor Dataset (ours), MC = Multi-Color Dataset (ours)

The linear probing achieves 22.7% on Multi-Factor (tied with logits) and 20.7% on Multi-Color, slightly better than both logits (19.3%) and IN1K-KNN (20.0%).

6.5 Factor-Level Robustness Analysis

This section directly addresses which rendering factors (camera position, material, lighting, background, fog) most degrade model performance, and whether their impact differs across semantic levels (ImageNet fine-grained vs. ShapeNet superclass). We report stratified Top-1 accuracy for four representative models spanning the performance spectrum: *DINOv2-L-KNN* (top performer overall), *DINOv3-L-KNN* (top performer on Multi-Color), *CLIP-L-IN1K-KNN* (best language-supervised model), and *ResNet-50-IN1K-logits* (strongest supervised CNN baseline). Each factor stratum contains exactly 160 images, ensuring balanced comparisons across levels.

6.5.1 Viewpoint: Camera Position

The Multi-Factor Dataset renders each object from five camera positions: four lateral views at elevation $Z = 0.2$ (left, right, front, back) and one top-down view at $Z = 0.4$. Table 13 reports Top-1 accuracy for each viewpoint at both semantic levels.

The top-down viewpoint is universally worst. For DINOv2-L, accuracy drops from 43.8% (best lateral view) to 18.1% under the top-down view, a −25.6pp collapse at the ImageNet level and −34.2pp at the ShapeNet level (from 59.7% to 25.5%). For ResNet-50, the collapse is even more severe in relative terms: from 18.6% to just

Camera Position	DINOv2-L		DINOv3-L		ResNet-50	
	IN	SN	IN	SN	IN	SN
Front (X=−1, Z=0.2)	43.8%	59.7%	44.4%	58.4%	18.6%	26.0%
Right (X=0, Y=1, Z=0.2)	35.6%	53.7%	45.0%	59.7%	8.1%	18.8%
Left (X=0, Y=−1, Z=0.2)	37.5%	55.0%	43.1%	57.1%	9.4%	16.8%
Back (X=1, Z=0.2)	28.1%	44.3%	31.3%	44.3%	11.3%	16.8%
Top-down (Z=0.4)	**18.1%**	**25.5%**	**21.9%**	**29.5%**	**1.9%**	**2.7%**

Table 13: Stratified Top-1 ImageNet (IN) and ShapeNet (SN) accuracy by camera position on the Multi-Factor Dataset.

1.9% at the ImageNet level.
Semantic level matters for the viewpoint effect. At the ShapeNet superclass level, the top-down drop is proportionally *larger* than at the ImageNet level for DINOv2-L (−34.2pp vs. −25.6pp).
Among lateral views, front ($X = -1$) is consistently easiest, back ($X = +1$) hardest. DINOv2-L achieves 43.8% from the front view but only 28.1% from the back, meaning a −15.6pp gap. This asymmetry is present across all models. There seems to be no distinct performance difference between the right and left viewpoints.

6.5.2 Material

The Multi-Factor Dataset applies five material conditions to each object in addition to the original texture: Camouflage, Glass, Noise, Titanium, and Wood. Table 14 reports per-material accuracy.

Material	DINOv2-L		DINOv3-L		CLIP-L		ResNet-50	
	IN	SN	IN	SN	IN	SN	IN	SN
Default	43.8%	59.7%	44.4%	58.4%	41.9%	55.0%	18.8%	26.2%
Wood	41.9%	55.0%	41.3%	52.3%	28.8%	37.6%	6.3%	9.4%
Titanium	27.5%	38.9%	31.9%	45.0%	26.3%	34.9%	9.4%	14.1%
Noise	32.5%	43.6%	28.8%	38.3%	**18.1%**	**24.8%**	4.4%	6.0%
Camouflage	34.4%	47.0%	27.5%	37.6%	21.3%	29.5%	**3.1%**	**3.4%**
Glass	**22.5%**	**34.9%**	**21.3%**	**30.2%**	20.0%	29.5%	**3.1%**	6.0%

Table 14: Stratified Top-1 ImageNet (IN) and ShapeNet (SN) accuracy by material on the Multi-Factor Dataset. Default material (original object texture) is the control condition.

Material is the strongest single factor for texture-dependent models. For CLIP-L, the range across material conditions is 41.9% − 20.0% = 21.9pp at the ImageNet level. For ResNet-50, the collapse is significant: from 18.8% on Default to just 3.1% on Glass and Camouflage, a −15.6pp absolute drop.

DINOv2-L shows stronger material resilience than CLIP-L. Both models achieve ∼43% on Default material, but DINOv2-L drops to 22.5% on Glass (−21.3pp) while CLIP-L drops to 20.0% (−21.9pp). More revealingly, on Camouflage: DINOv2-L retains 34.4% while CLIP-L drops to only 21.3%.
Wood is the most benign non-default material. For DINOv2-L (41.9%) and DINOv3-L (41.3%), wooden texture causes only ∼ 2pp degradation from Default.
The material impact is proportionally larger at the ImageNet level than at the ShapeNet level. DINOv2-L loses 21.3pp at ImageNet vs. 24.8pp at ShapeNet level. Although the absolute ShapeNet drop is slightly larger, the relative drop is −48.6% (ImageNet) vs. −41.6% (ShapeNet).

6.5.3 Material Color: Multi-Color Dataset

While the Multi-Factor Dataset varies material types to examine structural appearance shifts, the Multi-Color Dataset isolates a controlled color-hue shift by rendering the same objects in a uniform plastic material across nine colors (Black, Blue, Cyan, Green, Grey, Pink, Red, White, Yellow). This allows us to determine whether model collapse under material shifts is driven by the specific color distribution or by the texture simplification itself. Table 15 reports the stratified accuracies.

Plastic Color	**DINOv2-L**	**DINOv3-L**	**CLIP-L**	**ResNet-50**	**DINOv3-B**
Original	48.1%	47.5%	48.8%	25.6%	10.0%
Black	31.3%	37.5%	33.5%	15.3%	8.4%
Grey	31.2%	38.8%	32.8%	11.3%	8.4%
White	29.7%	36.7%	32.1%	10.3%	7.5%
Yellow	30.3%	36.3%	32.0%	9.5%	8.5%
Pink	31.5%	36.2%	31.2%	9.0%	7.6%
Blue	30.0%	35.8%	31.7%	9.5%	7.9%
Red	31.8%	36.7%	31.3%	9.9%	7.6%
Cyan	30.6%	35.5%	31.0%	8.9%	8.1%
Green	29.9%	34.8%	30.4%	9.3%	7.9%
Color Range	**2.1pp**	**4.0pp**	**3.1pp**	**6.4pp**	**1.0pp**

Table 15: Stratified ImageNet-level Top-1 accuracy across plastic colors on the Multi-Color Dataset. Original is the original object texture. Color range measures the performance difference between the easiest and hardest plastic color hues.

Our color-level breakdown reveals three big findings:
Color hue has very little impact on strong vision models. Across the nine plastic colors, the accuracy of DINOv2-L varies by only 2.1pp (from 29.7% on White to 31.8% on Red). Similarly, DINOv3-L exhibits a small range of 4.0pp, and CLIP-L varies by 3.1pp. This indicates that once a model learns robust, shape-biased representation manifolds, its performance is highly invariant to simple shifts in color hue. The model's classification decision does not depend on whether a chair is

rendered in green, pink, or yellow plastic.
The primary bottleneck is the texture shift, not the color shift. The transition from the multi-colored, detailed Default texture to *any* uniform plastic color causes a performance collapse. For DINOv2-L, the drop from Default (48.1%) to the average plastic color performance (∼30.8%) is an absolute loss of 17.3pp. For CLIP-L, the drop is 16.9pp.
ResNet-50 is uniquely sensitive to color-based silhouetting. ResNet-50 exhibits the biggest (6.4pp) range across plastic colors. Notably, it performs far better on Black plastic (15.3% Top-1) and Grey plastic (11.3%) than on Green (9.3%) or Cyan (8.9%).
DINOv3-B exhibits color-invariant collapse. Reflecting the outlier behavior noted in previous sections, the base DINOv3-B model remains relatively evenly collapsed across all colors (ranging from 7.5% to 8.5%). This flat profile reflects the same bad performance on the Multi-Color Dataset as the general Top-1 Accuracy in Figure 5 shows compared to the Multi-Factor Dataset.

Color and Viewpoint Association With both the Metrics for the Camera Viewpoints (see Section 6.5.1) and our color-level breakdown above, we can now compare the performance of the models when both color and viewpoint are changed. Table 16 highlights some counterintuitive metrics.

Color & Viewpoint	**DINOv2-L**		**DINOv3-L**		**ResNet-50**	
	MF	MC	MF	MC	MF	MC
Front (X=−1, Z=0.2)	43.8%	45.22%	44.4%	45.22%	18.6%	24.84%
Right (X=0, Y=1, Z=0.2)	35.6%	35.22%	45.0%	45.28%	8.1%	12.58%
Left (X=0, Y=−1, Z=0.2)	37.5%	38.8%	43.1%	46.06%	9.4%	13.33%
Back (X=1, Z=0.2)	28.1%	28.3%	31.3%	34.59%	11.3%	15.72%
Top-down (Z=0.4)	**18.1%**	**16.9%**	**21.9%**	**19.38%**	**1.9%**	**1.9%**

Table 16: Stratified Top-1 accuracy by camera viewpoint comparing the metrics from the base material on the MF (Multi-Factor Dataset) with the Top-1 accuracy for the viewpoints averaged over all color variations on the MC (Multi-Color Dataset).

Across the bigger models, there seems to be no significant average accuracy difference between the images with original textures and the color variations. ResNet, on the other hand, improves its Top-1 Accuracy by 3-7pp across standard lateral viewpoints with our colors.

6.5.4 Background Scene

Six background environments were used: a plain Unreal Engine map (Base, control condition), a plain desert (Desert), forest hills (Hills), a living room (Living), a mixed rocky desert (M-Desert), and a darker artificially lighted wooden indoor room (Basement). Table 17 reports per-background accuracy.

Background	DINOv2-L		DINOv3-L		ResNet-50		DINOv3-B	
	IN	SN	IN	SN	IN	SN	IN	SN
Basement	46.3%	60.4%	43.1%	59.7%	16.9%	20.1%	30.6%	43.0%
Base (control)	43.8%	59.7%	44.4%	58.4%	18.8%	26.2%	31.9%	45.6%
Desert	41.9%	55.0%	46.3%	60.4%	17.5%	23.5%	36.3%	50.4%
M-Desert	41.3%	55.0%	43.8%	55.7%	21.9%	27.5%	35.6%	48.3%
Hills	40.6%	54.4%	37.5%	49.7%	21.3%	28.9%	29.4%	37.6%
Living	**38.8%**	**53.7%**	**37.5%**	**52.3%**	**7.5%**	**10.1%**	**30.6%**	**40.9%**

Table 17: Stratified Top-1 accuracy by background scene on the Multi-Factor Dataset.

Background impact is moderate for strong models, severe for weak ones. For DINOv2-L, the range across backgrounds is only 7.5pp at the ImageNet level (46.3% to 38.8%), substantially smaller than the viewpoint or material effects. DINOv3-L shows a similar range of 8.8pp. In contrast, ResNet-50 exhibits a dramatic 14.4pp background effect: it achieves 21.9% in the M-Desert setting but collapses to 7.5% in Living. DINOv3-B displays the smallest range with 6.9pp.
Living is the hardest background on average. This environment places objects inside a domestic room without furniture, but with wall textures and specular reflections that create complex object-background interactions. For DINOv2-L, accuracy drops to 38.8% in Living. DINOv3-L also reaches its lowest accuracy in Living at 37.5%. For ResNet-50, it is particularly challenging, with only 7.5% accuracy.
Basement is the easiest background. Across all models, Basement has a performance close to or even better than UE5's default background (control). DINOv2-L achieves 46.3% on Basement and only 43.8% in Base. ResNet-50 for comparison achieves 16.9% and 18.8% respectively.

6.5.5 Lighting Color

Six lighting conditions were evaluated: natural light (control), dim near-dark light, and four chromatic conditions (red, green, blue, yellow). Table 18 reports per-lighting accuracy.

Lighting color has minimal impact on strong foundation models. For DINOv2-L, the total range across lighting conditions is only 4.4pp at the ImageNet level (48.1% to 43.8%). For DINOv3-L, it is 5.0pp.
Dim near-dark light is the easiest *lighting color.* This result appears consistently: DINOv2-L achieves its best performance (48.1%) under dim lighting, and ResNet-50 also achieves its best performance (21.3%) under dim light, higher than under natural light (18.8%).
Red light is the most disruptive chromatic condition. For ResNet-50, red light reduces accuracy to 10.0% at the ImageNet level, 11.3pp below dim lighting and 8.8pp below white light. Foundation models are more robust to this perturbation; DINOv2-L

Light Color	DINOv2-L		DINOv3-L		ResNet-50	
	IN	SN	IN	SN	IN	SN
Dim (near-dark)	**48.1%**	**62.4%**	**45.0%**	**59.1%**	**21.3%**	**28.9%**
Blue	47.5%	62.4%	41.9%	55.7%	14.4%	20.1%
Green	46.9%	61.7%	42.5%	57.1%	13.1%	16.1%
Yellow	46.9%	60.4%	40.0%	55.7%	15.6%	20.8%
Natural (control)	43.8%	59.7%	44.4%	58.4%	18.8%	26.2%
Red	**45.0%**	**60.4%**	**41.3%**	**55.0%**	**10.0%**	**16.1%**

Table 18: Stratified Top-1 accuracy by lighting color on the Multi-Factor Dataset. Dim light ranks consistently best.

drops only 3.1pp from best (dim) to worst (natural at 43.8%), with all chromatic conditions falling between 45% and 48%.

6.5.6 Fog

Fog is a binary factor, meaning each image either has atmospheric fog enabled or disabled, and Table 19 illustrates the effect of fog on the images.

Fog	DINOv2-L		DINOv3-L		ResNet-50	
	IN	SN	IN	SN	IN	SN
No Fog	43.8%	59.7%	44.4%	58.4%	18.8%	26.2%
Fog	45.6%	62.4%	48.8%	63.8%	19.4%	27.5%
Δ (Fog − No Fog)	+1.9pp	+2.7pp	+4.8pp	+4.8pp	+0.6pp	+1.3pp

Table 19: Stratified Top-1 accuracy with and without fog on the Multi-Factor Dataset.

Fog counterintuitively improves performance. For DINOv2-L, fog increases ImageNet accuracy by +1.9pp and ShapeNet accuracy by +2.7pp. DINOv3-L improves exactly the same by 4.8pp and 4.8pp for ImageNet and ShapeNet. ResNet-50 also improves slightly under fog (+0.6pp / +1.3pp).

6.5.7 Cross-Factor Importance Ranking

To synthesize the per-factor analyses, Table 20 ranks the five rendering factors by their impact on model performance, measured as the Top-1 accuracy range (maximum minus minimum accuracy across factor levels) for DINOv2-L and ResNet-50 on the Multi-Factor Dataset.

The factor ranking is: Camera Position > Material > Background > Lighting > Fog. Camera position and material are by far the dominant factors, together accounting

Factor	N Levels	Accuracy Range (pp)		Direction
		DINOv2-L	ResNet-50	
Camera Position	5	25.6	16.7	Top-down ↓↓↓
Material	6	21.3	15.6	Glass/Camo ↓↓, Wood ≈
Background	6	7.5	14.4	Home Interior ↓↓
Lighting	6	4.4	11.3	Red ↓, Dim ↑
Fog	2	1.9	0.6	Fog ↑

Table 20: Factor importance ranking by Top-1 accuracy range (max − min across levels) on the Multi-Factor Dataset. Arrows indicate the direction of the most impactful level.

for performance ranges of 21–26pp. Background is a secondary factor (∼7–14pp), lighting has minimal impact on strong models, and fog is slightly beneficial.
The relative importance of factors shifts with model capacity. For ResNet-50, the accuracy range induced by Background (14.4pp) is nearly as large as that of Material (15.6pp), a near-parity that does not hold for DINOv2-L (7.5pp vs. 21.3pp). Conversely, both models are equally vulnerable to the top-down viewpoint in proportional terms.
No factor is irrelevant. Even lighting produces an 11.3pp range for ResNet-50, and the fog effect, while small, is statistically consistent across all models.

6.6 Scaling Trajectories and Architectural Performance

This section examines whether scaling model capacity (small, base, and large variants) and changing the underlying architectural family (convolutional networks, standard transformers, hierarchical transformers) provides a systematic path toward resolving previously shown vulnerabilities. This analysis is critical for understanding whether OOD robustness is a natural byproduct of scaling up compute and parameters, or whether it requires specific pre-training objectives and structural inductive biases.

We compare five model families across their available parameter scales: *DINOv2* (S, B, L), *DINOv3* (S, B, L), *Swin* (S, B, L), *Hiera* (S, B, L), *CLIP* (B, L), and *ViT* (B, L), alongside the convolutional baseline *ResNet-50*. To ensure architectural fairness, all comparisons use ImageNet-1K fine-tuned k-NN evaluation (IN1K-KNN) unless specified otherwise.

6.6.1 Scaling Trajectories Across Model Families

Figure 9 visualizes the ImageNet-level Top-1 accuracy of each model family as a function of parameter scale on both the Multi-Factor and Multi-Color datasets. The quantitative scaling jumps are summarized in Table 21.

Family	Scale Jump	MF (Top-1)		MC (Top-1)	
		Accuracy	Δ	Accuracy	Δ
DINOv2	Small → Base	24.7% → 31.7%	+7.0pp	20.4% → 26.4%	+6.0pp
	Base → Large	31.7% → 38.6%	+6.9pp	26.4% → 30.9%	+4.5pp
DINOv3	Small → Base	15.5% → 26.8%	+11.3pp	15.0% → 8.2%	**-6.8pp**
	Base → Large	26.8% → 38.3%	+11.5pp	8.2% → 36.6%	**+28.4pp**
Hiera	Small → Base	17.6% → 20.1%	+2.5pp	18.5% → 20.1%	+1.6pp
	Base → Large	20.1% → 25.0%	+4.9pp	20.1% → 24.2%	+4.1pp
Swin	Small → Base	19.5% → 21.4%	+1.9pp	20.4% → 20.0%	-0.4pp
	Base → Large	21.4% → 23.2%	+1.8pp	20.0% → 20.5%	+0.5pp
CLIP	Base → Large	25.7% → 34.4%	+8.7pp	23.4% → 32.3%	+8.9pp
ViT	Base → Large	13.2% → 15.7%	+2.5pp	13.6% → 16.9%	+3.3pp

Table 21: Robustness scaling jumps across model families (ImageNet-level Top-1 accuracy). Abbreviations: MF = Multi-Factor Dataset (ours), MC = Multi-Color Dataset (ours)

The data reveals a significant difference between the model families:
Steep robustness scaling. Self-supervised models (DINOv2, DINOv3) and language-supervised contrastive models (CLIP) exhibit substantial, consistent gains with scale. For DINOv2, scaling from Small to Large yields a cumulative gain of +13.9pp on the Multi-Factor Dataset and +10.5pp on the Multi-Color Dataset. CLIP gains +8.7pp (Multi-Factor) and +8.9pp (Multi-Color) when moving from Base (86M parameters) to Large (307M).
Flat robustness scaling. In contrast, supervised models (Swin, ViT) show almost no robustness improvement with scale. For Swin, scaling from Small (50M parameters) to Large (197M) gains +3.7pp on the Multi-Factor Dataset, and is virtually flat on the Multi-Color Dataset (20.4% → 20.5%, a +0.1pp change across three orders of parameter scale). ViT gains +2.5pp (Multi-Factor) and +3.3pp (Multi-Color) from Base to Large. Lastly, DINOv1 shows small improvements of +1.4pp and +1.8pp on both Datasets respectively.

6.6.2 The DINOv3-B Multi-Color Anomaly: Non-Monotonic Scaling

As seen in the previous sections, DINOv3 exhibits a striking, non-monotonic scaling behavior on the Multi-Color Dataset. While DINOv3-S achieves 15.0% and DINOv3-L achieves the highest overall accuracy of 36.6%, the intermediate DINOv3-B collapses to just 8.2%, a drop of −6.8pp below its smaller counterpart, followed by a massive +28.4pp recovery from Base to Large. To understand this collapse, we examine the stratified behavior of DINOv3-B across plastic colors (Section 6.5.3). DINOv3-B's performance remains uniformly low ($\approx 7.5\% - 8.5\%$) across all nine plastic hues.

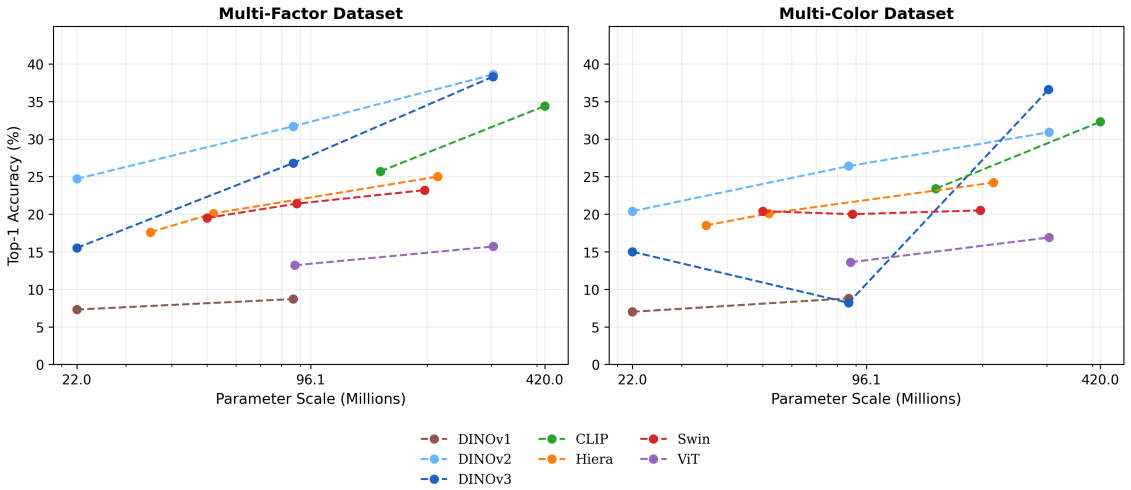


Figure 9: Top-1 accuracy as a function of model parameter scale on the Multi-Factor (left) and Multi-Color (right) datasets.

6.6.3 Inductive Biases: CNNs, ViTs, and Hierarchical ViTs

Beyond model scale, the structural architecture of the model determines its inductive biases, which dictate how it processes visual space. We compare three distinct architectural families:

1. **Convolutional Neural Networks** (ResNet-50): Characterized by local receptive fields, translation weight sharing, and a rigid spatial hierarchy.
2. **Standard ViTs** (ViT-B/L): Characterized by global self-attention from the first layer, with no built-in spatial hierarchy. This family includes specialized cases like self-supervised DINOv2 and language-supervised CLIP.
3. **Hierarchical ViTs** (Swin-B/L, Hiera-B/L): Characterized by windowed local self-attention, with a spatial hierarchy that partitions and downsamples visual tokens across stages (similar to CNNs).

Comparing these architectures under domain shift reveals a clear progression in robustness:
CNNs are highly vulnerable to spatial and textural changes. ResNet-50-IN1K collapses to 9.2% (Multi-Factor) and 8.6% (Multi-Color). As shown in Section 6.5, ResNet-50 is highly sensitive to background clutter and lighting changes, and fails completely under material overrides.
Hierarchical Transformers outperform CNNs and Standard Transformers under supervision. When fine-tuned on ImageNet-1K, Swin-L achieves 23.2% (Multi-Factor) and 20.5% (Multi-Color), substantially outperforming ViT-L-IN1K (15.7% / 16.9%) and ResNet-50 (9.2% / 8.6%). Similarly, Hiera-L achieves 25.0% (Multi-Factor) and 24.2% (Multi-Color).
Self-supervised standard ViTs are the most robust. While supervised standard ViTs

perform poorly (ViT-L-IN1K: 15.7% on Multi-Factor), the self-supervised standard ViTs (DINOv2-L: 38.6%, DINOv3-L: 38.3%) achieve the top overall scores.

6.7 Equivariance and Explainability

The preceding sections quantified model robustness in terms of classification accuracy under domain shift. However, accuracy alone does not fully describe how a model's internal representation manifold responds to physical rendering changes. Does a robust model succeed by ignoring rendering factors entirely (invariance), or by systematically encoding them separately from the object's identity (equivariance)?

6.7.1 Equivariance Across Rendering Factors

To measure how sensitive model representations are to specific rendering factors, we calculate an equivariance score across the variations in the Multi-Factor and Multi-Color datasets. A higher score indicates that the model's feature space changes significantly and predictably as the rendering factor changes (high sensitivity/equivariance), whereas a lower score indicates that the representation remains relatively static (invariance).
Table 22 reports the equivariance scores for base-scale models on the Multi-Factor Dataset.

Model	Background	Material	Light Color	Camera	Fog
DINOv2-L (IN1K)	**0.909**	0.534	0.680	0.480	0.790
DINOv2-L (raw)	0.408	0.184	0.164	0.106	0.248
DINOv3-S (raw)	0.649	0.368	0.389	0.219	0.522
DINOv3-B (raw)	0.563	0.294	0.222	0.149	0.388
DINOv3-L (raw)	0.518	0.290	0.237	0.158	0.427
CLIP-L (IN1K)	0.333	0.150	0.193	0.090	0.121
CLIP-L (raw)	0.580	0.368	0.532	0.389	0.443
ResNet-50 (IN1K)	0.626	0.183	0.372	0.237	0.415

Table 22: Feature equivariance scores across rendering factors on the Multi-Factor Dataset with models finetuned on ImageNet (IN1K) or not (raw). Higher values indicate greater sensitivity/representational change in response to the factor.

The equivariance analysis reveals three dynamics of model representation:
Supervised Fine-Tuning destroys invariance. The most striking pattern in Table 22 is the big gap between raw, self-supervised representations and their fine-tuned counterparts for some models. Raw DINOv2-L exhibits strong invariance (low scores) across all factors: 0.164 for lighting, 0.184 for material, and 0.408 for background. However, once DINOv2-L is fine-tuned on ImageNet-1K, its sensitivity across all

variations explodes. Notably, background equivariance jumps to 0.909, fog to 0.790, and lighting to 0.680.
To ground these abstract equivariance scores in tangible classification performance, Table 23 directly compares the equivariance scores against the Top-1 Accuracy Swing for each factor for both models.

Factor	**DINOv2-L (IN1K)**		**DINOv2-L (Raw)**	
	Eq. Score	**Acc. Swing**	**Eq. Score**	**Acc. Swing**
Background Level	0.909	14.4pp	0.408	7.5pp
Fog	0.790	1.9pp	0.248	1.8pp
Light Color	0.680	5.0pp	0.164	4.3pp
Camera Position	0.480	33.1pp	0.106	25.7pp
Material	0.534	30.6pp	0.184	21.3pp

Table 23: Comparison of Equivariance Scores and Top-1 Accuracy Swings (Max − Min across factor strata) between IN1K-finetuned and raw DINOv2-B.

This comparison reveals a small dichotomy. For *environmental context factors* (Background, Fog, Light), the high equivariance scores of the IN1K model predict large accuracy instability. The model's extreme sensitivity to the background (0.909) causes a 14.4pp swing in performance, whereas the raw model remains much more invariant (0.408 score, 7.5pp swing).
However, for *object-centric factors* (Camera Position, Material), both models suffer massive accuracy drops (∼21–33pp) despite the raw model having much lower equivariance scores.

DINOv3 exhibits elevated structural equivariance. Compared to raw DINOv2-L, raw DINOv3-L shows higher equivariance scores across the board (e.g., 0.518 vs. 0.408 for background; 0.290 vs. 0.184 for material).
We observe a distinct inverse scaling trend for DINOv3: as model capacity increases, equivariance scores consistently decrease (e.g., background equivariance scales as $S = 0.649$, $B = 0.563$, $L = 0.518$).

The Multi-Color Dataset isolates geometric and material variation. Table 24 reports the scores on the Multi-Color Dataset.

On the Multi-Color Dataset, raw CLIP-B shows high sensitivity to material color variations (0.416) and camera position (0.238), significantly higher than DINOv2-B (0.169 and 0.085). This aligns with the accuracy findings from Section 6.5.3: CLIP's language-aligned features are heavily disrupted by uniform textures and boundary-breaking colors, leading to high representational variance. DINOv2, by contrast, maintains stable, shape-centric invariance even when texture is stripped away.

Model	Material	Camera
DINOv2-B (raw)	0.169	0.085
DINOv3-B (raw)	0.299	0.109
CLIP-B (raw)	0.416	0.238
Swin-B (IN1K)	0.228	0.137
ResNet-50 (IN1K)	0.174	0.193

Table 24: Feature equivariance scores on the Multi-Color Dataset.

6.7.2 Explainability and Localization under Nuisance Factors (Grad-CAM and Segmentation-Grounded Evaluation)

To move beyond aggregate accuracy, we employ Grad-CAM to trace classification decisions back to specific localized features. By aggregating heatmaps across nuisance factors (e.g., averaging focus across all 'Living Room' backgrounds), we quantify the degree of context bias.
Figure 10 provides qualitative evidence that DINOv2 generally concentrates its Grad-CAM responses on the object region, whereas ResNet-50 is more susceptible to context-dependent shifts. In particular, under harder settings, ResNet-50 either fails to localize the target object or reallocates saliency toward background structures, indicating that its decision-making is less robust to nuisance variation.

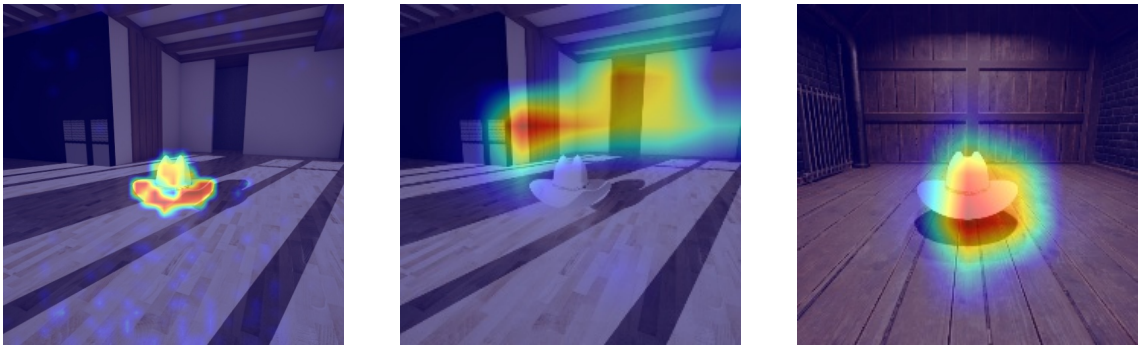
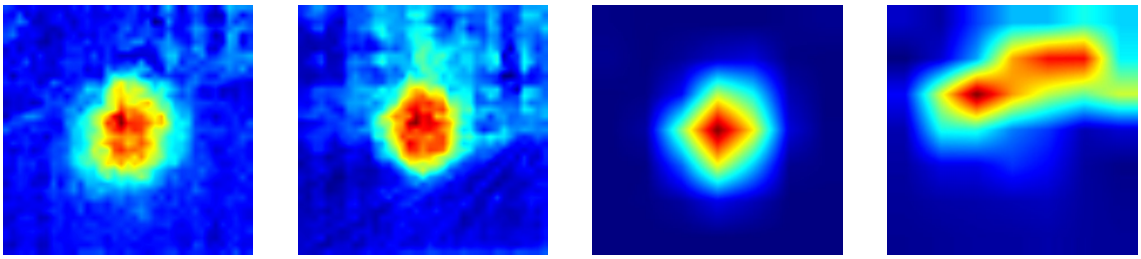


Figure 10: Qualitative Grad-CAM results for DINOv2-L-IN1K and ResNet-50. The DINOv2 visualization (Living Room; left); ResNet-50 visualization (Living Room; middle); ResNet-50 visualization (Basement; right).

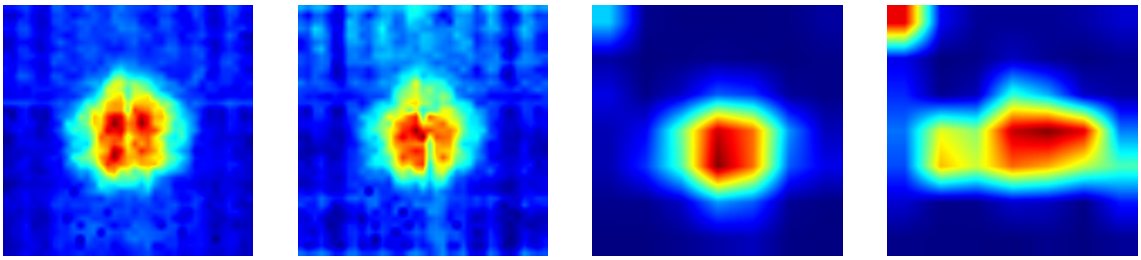
To quantify these effects, we compute average Grad-CAM activations for each model under matched nuisance levels. Across backgrounds, both models show relatively centered and compact activations in easier conditions, but the localization quality diverges under challenging contexts (see Figure 11). DINOv2 maintains comparatively stable attention maps, while ResNet-50 exhibits a pronounced increase in spatial drift and broader activation footprints. Analogous trends are observed across material variations: for Wood, both models localize reasonably well, but ResNet-50 introduces localized artifacts unrelated to the object. When the material becomes

more challenging (Glass), DINOv2 still preserves a tighter focus on object-relevant regions, whereas ResNet-50 produces substantially more diffuse heatmaps and increased activation in non-object areas (see Figure 12).



(a) DINOv2-L: Forest Hills (b) DINOv2-L: Living Room (c) ResNet-50: Forest Hills (d) ResNet-50: Living Room

Figure 11: Comparison of average Grad-CAM activations for the specified models across backgrounds.



(a) DINOv2-L: Wood (b) DINOv2-L: Glass (c) ResNet-50: Wood (d) ResNet-50: Glass

Figure 12: Comparison of average Grad-CAM activations for the specified models across materials.

7 Discussion

This section contextualizes and synthesizes the empirical findings documented in the previous sections, exploring the theoretical and practical implications of model behavior under severe domain shifts. We first analyze the discrepancy between ID performance and OOD robustness, framing the observed failures through the lens of representation misspecification and spurious shortcuts. Furthermore, we investigate the role of semantic hierarchy in mitigating or revealing failures across different levels of category granularity. We then evaluate the role of different evaluation protocols in combination with the various VFM architectures. Next, we examine the specific impact of camera position and material variation, detailing the lack of true 3D spatial understanding and the capacity-driven, non-monotonic scaling trajectories of texture bias. Furthermore, we dissect the impact of model scaling and architectural inductive biases on resilience. We then discuss the role of Supervised Fine-Tuning

(SFT) in introducing context bias, comparing different evaluation methods for freezing or adapting pre-trained representations. Finally, we reconcile our findings with theoretical expectations regarding self-supervised and hierarchical learning.

7.1 The Illusion of Robustness: Misspecifications and Spurious Correlations

The empirical findings presented in Section 6.1 reveal a stark discrepancy between ID performance and OOD robustness. While state-of-the-art VFMs achieve near-perfect classification scores on the ImageNet-1K validation set (ranging between 80% and 90% Top-1 accuracy), their performance degrades sharply when evaluated on the synthetic domains of the Multi-Factor and Multi-Color datasets. Even the highest-performing architecture evaluated in this work, DINOv2-L under a k-NN evaluation protocol, collapses to just 38.6% Top-1 accuracy at the ImageNet-1K level under multi-factor rendering shifts (Table 3). The convolutional baseline, ResNet-50, suffers a near-total collapse, dropping to 7.7% accuracy on the Multi-Factor Dataset and 8.8% on the Multi-Color Dataset.
The smaller gap between the weakest and strongest models on our Multi-Factor Dataset, compared to the other datasets, indicates that all models in our dataset struggle considerably, rather than just a subset.

This big performance gap suggests that the high accuracy reported on standard closed-world validation sets is, to a large extent, an *illusion of robustness*. Under the closed-world i.i.d. assumption, model evaluation over-indexes on low-level statistical correlations that happen to co-occur consistently in the training distribution, but are not causally linked to the semantic identity of the target classes. In the terminology of causal representation learning and OOD generalization, this is a clear manifestation of model misspecification (Yang et al., 2024; Miyai et al., 2025). The models optimize their empirical risk by exploiting spurious shortcuts, such as localized texture statistics, dominant background contexts, and canonical lighting angles, which are computationally easier to learn than the complex, invariant geometric structure of the 3D objects themselves (Geirhos et al., 2020; Xiao et al., 2020).
Our Grad-CAM analysis further clarifies these findings. DINOv2-L consistently maintains a tighter, more object-focused attention distribution in both qualitative visualizations and quantitative spatial metrics, while ResNet-50 demonstrates marked susceptibility to nuisance factors. Under challenging backgrounds and material shifts, ResNet-50's Grad-CAM heatmaps often drift away from the object, activating irrelevant background regions or broadly dispersing attention. These explainability-driven insights, supported by segmentation-grounded evaluation, show that DINOv2-L's internal representations are anchored to core object geometry, whereas ResNet-50 is more vulnerable to spurious environmental cues. While DINOv2-L does occasionally exhibit some sensitivity to distractions, it does so significantly less than ResNet-50. Thus, explainability tools not only reveal the source of predic-

tive failures but also highlight the limitations of conventional accuracy metrics in capturing true model robustness.

Our synthetic datasets further expose these misspecifications by systematically decoupling these co-occurring factors. When the physical appearance of an object is perturbed through non-standard viewpoints, artificial material overrides, and out-of-context backgrounds, the models' statistical shortcuts collapse. The fact that the performance drop is so severe across all evaluated models indicates that current pre-training paradigms (supervised, self-supervised, or language-aligned) do not natively resolve this reliance on spurious correlations, highlighting a fundamental limitation in how DNNs construct their visual decision boundaries.

The observed divergence in correlation patterns across benchmarks further clarifies the types of robustness being measured. For example, ImageNet-C primarily tests for robustness against pixel-level corruptions such as blur, noise, and compression artifacts. Forms of perturbation that challenge perceptual invariance but do not fundamentally alter the material appearance of objects. In contrast, the Multi-Color Dataset and ObjectNet both feature shifts that drastically change surface appearance, demanding true shape and semantic recognition. The strong correlations between our Multi-Factor Dataset, PUG-ImageNet, ImageNet-Sketch, and Tiny-ImageNet, as well as between our Multi-Color Dataset, ObjectNet, ImageNet-A, and ImageNet-D, confirm that our datasets reward the same general robustness properties as established benchmarks: models with robust, shape-biased representations consistently outperform texture-dependent models across all severe distribution shifts. The fact that our datasets correlate with different benchmarks is to be expected, as they focus on different types of robustness.

7.2 The Hierarchy of Failure: Fine-Grained Precision vs. Semantic Recognition

Evaluating models purely on fine-grained metrics (such as ImageNet-1K Top-1 accuracy) often masks their true visual capabilities under domain shift. When a model misclassifies a rendered `airliner` as a `warplane` or `projectile`, it is penalized under the standard fine-grained metric, even though it has successfully recognized the core shape semantics of the object. Our comparison between ImageNet-level and ShapeNet-level accuracy (Section 6.1 and Section 6.4) reveals that evaluating models at the ShapeNet superclass level yields consistently higher accuracy, with an average superclass-to-fine-grained classification accuracy ratio of 1.418 across both datasets. Weaker models, such as ResNet-50, see their performance nearly double under superclass evaluation, demonstrating that while they lose fine-grained precision, they retain baseline semantic recognition.
The models consistently classified the ShapeNet classes accurately. This means that objects correctly identified by models that performed poorly were also identified correctly by those that performed better. Moreover, the presence of distinct vertical

banding over the classes (see Figures 7 and 8) with very few outliers provides indirect evidence that the observed superclass–fine-grained discrepancies are not dominated by systematic evaluation artifacts or biased object sampling. Instead, the pattern supports the interpretation that the ranking of failures primarily reflects genuine representational differences rather than noise in the evaluation pipeline.

However, our results also show that under severe shift, this semantic structure itself is vulnerable to collapse. Specifically, on the Multi-Color Dataset, the average ShapeNet-to-ImageNet accuracy ratio drops from 1.88 (on Multi-Factor) to 1.40 (on Multi-Color), indicating that the semantic gap narrows when natural textures are removed. When objects are rendered as solid plastic, the resulting unnatural specular highlights and lack of realistic surface details disrupt the models' ability to parse fundamental geometric contours, causing superclass accuracy to plummet closer to baseline fine-grained performance.

Furthermore, we observe distinct class-wise variations in robustness (Table 8). Categories with highly distinct, rigid structural templates, such as pianos, guitars, and helmets, maintain high superclass accuracy because they possess strong geometric signatures that survive extreme material and lighting shifts. Conversely, categories that heavily rely on texture, fine details, or context, such as buses, cabinets, and trains, experience catastrophic failure. This highlights a semantic hierarchy of failure: models do not fail uniformly across classes, but rather collapse along specific semantic dimensions that are highly dependent on the object's structural properties and its reliance on contextual cues.

7.3 The Role of Evaluation Protocols in Measuring Robustness

A comparative analysis of evaluation protocols, including classification using model logits, k-NN retrieval in feature space, and linear probing, shows that the choice of evaluation method can substantially affect conclusions about model robustness under distribution shift (see Section 6.4).
For most models evaluated, accuracy achieved with logits and IN1K-KNN methods on the Multi-Factor and Multi-Color synthetic datasets is closely matched. This similarity indicates that, after fine-tuning on ImageNet-1K, both the classification head and the feature space structure contribute similarly to performance under distribution shift. The feature manifold and the linear decision boundaries created by the head are therefore equally adapted or maladapted to the new synthetic domains.

In high-performing self-supervised models (e.g., DINOv2 and DINOv3), k-NN sometimes outperforms logits, especially under moderate geometric shifts. This result suggests that the feature space in these models retains generalizable structure that the classification head does not fully use. Conversely, in architectures like ResNet-50 and Swin-L, logits consistently outperform k-NN, supporting the view that super-

vised pre-training produces feature spaces optimized for linear decision boundaries rather than metric-based retrieval, with classification heads encoding class-boundary information not directly accessible via nearest-neighbor methods.

Further examination of the effects of fine-tuning reveals important differences between self-supervised and supervised models. In the case of DINOv2, subsequent fine-tuning on ImageNet-1K results in a measurable, though modest, reduction in generality. Linear probing experiments demonstrate that training a new linear classifier on frozen, fine-tuned DINOv2 features can largely restore the OOD robustness that is diminished by fine-tuning. This finding suggests that the principal limitation imposed by fine-tuning arises at the level of the classification head, which becomes overly specialized to the natural image domain, while the underlying feature representation remains broadly transferable.

This contrast is especially evident in models like CLIP, where the feature space is structured around broad semantic concepts due to the language-image contrastive objective. Without further alignment to specific class labels (such as through fine-tuning or linear probing), k-NN produces suboptimal results for fine-grained classification. Applying a trained linear head or fine-tuning to ImageNet categories allows both k-NN and logits to achieve competitive performance, highlighting the need to align the feature space with the target label structure for accurate evaluation.

For fully supervised architectures, fine-tuning provides only a marginal improvement in k-NN performance, indicating that their feature spaces are inherently well-aligned with their classification heads and less transferable relative to their self-supervised counterparts.

7.4 Viewpoint Sensitivity and the Lack of 3D Understanding

The cross-factor importance ranking in Section 6.5.7 identifies camera position as the single most disruptive factor affecting model performance, responsible for an accuracy range of 25.6 percentage points for DINOv2-L and 16.7 percentage points for ResNet-50 (Table 20). In particular, the top-down perspective emerges as a near-universal failure mode across all architectural families and pre-training objectives. Additionally, for some classes, the shape signature is most disrupted by a vertical camera: `chair`, `table`, or `dishwasher` become nearly indistinguishable from one another when viewed from above, whereas lateral views preserve category-distinctive silhouettes. These results demonstrate that despite being pre-trained on millions of diverse images, modern VFMs fail to acquire a robust, viewpoint-invariant 3D spatial representation of objects, remaining highly sensitive to the 2D projective geometry of the camera angle.

This vulnerability is historically well-documented as standard training distributions

are heavily biased toward canonical horizontal perspectives, leaving unusual angles like the top-down view severely underrepresented (Alcorn et al., 2019; Geirhos et al., 2020). However, our analysis of feature space equivariance reveals a deeper, more troubling representational dissociation. As shown in Table 23, raw DINOv2-L exhibits high mathematical stability across camera position variations, achieving a very low camera equivariance score of 0.106. Yet, despite this smooth representation, the model suffers a massive 25.7pp accuracy drop.

This finding demonstrates that *latent representation stability does not guarantee downstream generalization*. Although the raw self-supervised feature space groups different viewpoints of the same object together in a relatively compact manifold, these manifolds are not aligned with the classifier's decision boundaries. When the viewpoint deviates from the canonical training perspectives, the feature vectors, though stable, drift across the decision boundaries established by the classification head. A true 3D physical bias during pre-training would enforce representation consistency across the transformation. Without that bias, the models partition the representation space based on 2D statistical patterns rather than the underlying 3D structure of the physical world.

7.5 Impact of Individual Factors: Material, Lighting, Background, and Fog

The findings in Section 6.5 illustrate the significantly different impact on performance.

Material overrides, such as glass, camouflage, wood, titanium, and noise, emerge as some of the most disruptive factors for model performance. The difference between the most benign and the most challenging materials is substantial for all models. Notably, materials that disrupt both local texture and global shape cues, such as transparent or highly patterned surfaces, lead to the most severe failures. In contrast, materials that preserve surface structure, like wood, result in only minor degradations. These findings underscore a persistent reliance on surface appearance and texture. When typical visual cues are replaced with challenging materials, even robust models lose much of their fine-grained recognition capacity.

Lighting manipulations, contrary to our expectations, have a surprisingly small, sometimes even positive, effect on accuracy. For example, dim lighting consistently produces higher accuracy for all models compared to natural or colored lighting conditions. This effect likely arises because dim lighting suppresses background and surface texture, thereby emphasizing global shape cues and facilitating silhouette-based recognition. Chromatic lighting, particularly red light, can be disruptive for some models, possibly due to a mismatch with the color statistics present during training. However, larger, more robust models generally handle illumination changes with minimal degradation, indicating a degree of learned invariance.

Background changes tend to produce moderate but model-dependent effects. Certain backgrounds, especially those with complex textures and specular surfaces,

are consistently challenging for all models, while simpler, low-clutter backgrounds improve performance. The results suggest that background suppression, whether achieved through color, lighting, or environmental conditions, can facilitate object-centric recognition. Stronger models with attention mechanisms are more resistant to context bias, but all models experience some degradation in cluttered or realistic scenes.

Fog may be the most counterintuitive phenomenon. Rather than degrading performance, the addition of moderate fog generally leads to a marginal improvement in accuracy for all models. This effect is likely attributable to the suppression of background detail, which enables models to focus more effectively on the foreground object. Since we only utilized one fog intensity, there may be adversarial effects for a stronger fog setting. This finding echoes the result that simpler backgrounds benefit classification, and it demonstrates that perceptual degradations that simplify the scene or remove distractors can actually improve model accuracy.

7.6 Color Bias

By overriding the natural materials of ShapeNet objects with uniform, solid-colored plastic, the Multi-Color Dataset isolates texture simplification from other rendering factors. This dataset directly probes the trade-off between surface appearance and global geometric form, a central theme in the robustness literature (Geirhos et al., 2022). On average, models had a harder time with the Multi-Color Dataset than with the Multi-Factor Dataset (see Section 6.5.3). This finding suggests that severe textural distribution shifts alone can be more destructive to model performance than the averaged effect of geometric and environmental perturbations.

Furthermore, our results show that, for strong vision models, the color itself has very little impact. This suggests that once a model learns robust, shape-biased representation manifolds, its performance is highly invariant to simple shifts in color hue. The model's classification decision does not depend on whether a chair is rendered in green, pink, or yellow plastic. Instead, overall texture simplification and the removal of local surface cues drive performance collapse, not the specific color applied.

There appears to be no substantial difference in model accuracy across different viewpoints when comparing objects with original textures to those with color variations. This finding is counterintuitive, as one might expect that combining an altered viewpoint and challenging plastic texture would further degrade performance, especially given that both factors individually led to sizable drops. The absence of a compounding negative effect suggests that the difficulties introduced by the different viewpoints may already saturate the models' vulnerability, leaving little room for additional degradation. Alternatively, it could indicate that the models' representations are equally brittle to either factor in isolation, and that their mechanisms for recognizing objects do not further deteriorate when these challenging conditions are combined.

7.7 Implications of Model Scale and Inductive Biases

As observed in Section 6.6, increasing model size generally leads to improvements in top-1 accuracy across all benchmarks. This trend is especially pronounced for self-supervised ViTs, such as DINOv2 and DINOv3, which exhibit consistent performance gains with each architectural increment from base to large variants. These findings support the prevailing view that larger models are better able to capture abstract, transferable representations, allowing for enhanced generalization under challenging distribution shifts.
However, the scaling benefit is not uniform across all architectures. While both DINOv2-L and DINOv3-L consistently outperform their base counterparts, scaling effects are less pronounced or even inconsistent for supervised models such as ResNet-50 and Swin-B. This suggests that the inductive biases and training objectives inherent to each architecture play a critical role in determining the effectiveness of scaling, particularly when faced with synthetic domain challenges.
Notably, the performance gap between self-supervised and supervised models widens as model size increases. Large self-supervised models maintain or even enhance their robustness under severe distribution shifts, whereas supervised architectures continue to exhibit substantial accuracy declines, especially under material and texture perturbations. These results indicate that self-supervised objectives, which encourage instance-level discrimination and invariance to superficial cues, are more effective at leveraging increased capacity for robust visual understanding.

An unexpected observation arises in the behavior of certain scaled models under specific synthetic perturbations. The scaling trajectories of the DINOv3 family on the Multi-Color dataset reveal a non-monotonic anomaly: while DINOv3-S achieves 15.0% accuracy and DINOv3-L reaches the highest accuracy of 36.6%, the intermediate DINOv3-B collapses to just 8.2% Top-1 accuracy, dropping significantly below its smaller counterpart (Table 21).
We hypothesize that this non-monotonic progression indicates a capacity-driven phase change in representation alignment during self-supervised pre-training:
Low-Capacity Shape Bias (DINOv3-S). With limited capacity, the small model prioritizes coarse shape boundaries and low-frequency structures, maintaining a baseline recognition capability when evaluated on the Multi-Color Dataset.
Shortcut Transition Collapse (DINOv3-B). At the intermediate capacity scale, this model can optimize its self-supervised instance-discrimination loss by focusing on high-frequency textural patterns. This representation is highly effective for natural, ID images but collapses when these diagnostic textures are replaced with plastic, relying too heavily on local cues.
High-Capacity Shape Fallback (DINOv3-L). With a larger capacity, this model encodes both high-frequency texture and global shape invariants. When texture is removed, it still performs well by relying on shape-based invariants.
However, alternative explanations must be considered. This anomaly could also stem from *optimization instability* specific to the Base scale during pre-training. It is possible that the learning rate schedules or batch sizes optimized for the Small and

Large variants were suboptimal for the Base configuration, leading to convergence in a sharper, less generalizable minimum. Without access to the original training loss curves and hyperparameter sweeps for DINOv3-B, we cannot definitively distinguish between a fundamental representational phase transition and a transient optimization artifact. Regardless of the root cause, this finding serves as a critical caution: intermediate-scale models may exhibit unique vulnerabilities not present in the original counterparts, making single-scale evaluations insufficient for assessing family-wide robustness.

7.8 Context Bias and Supervised Fine-Tuning

A major finding of our equivariance analysis is that downstream SFT can change the representation space of robust foundation models, sacrificing their general OOD robustness to maximize ID classification performance (see Section 6.7). As reported in Table 22, the raw pre-trained DINOv2-L backbone exhibits strong invariance to environmental factors, with low equivariance scores of 0.408 for background, 0.164 for lighting, and 0.248 for fog. However, once the backbone undergoes SFT on ImageNet-1K, its sensitivity to these factors explodes, with background equivariance jumping to 0.909, fog to 0.790, and lighting to 0.680.

This representational shift has direct consequences for OOD classification stability. As compared in Table 23, the high background equivariance score (0.909) of the fine-tuned model translates directly into a large 14.4pp accuracy swing across background levels, whereas the raw model remains much more stable (7.5pp swing). As demonstrated in Table 11, the accuracy metrics for both models exhibit negligible divergence. While an ideal robust model would yield an equivariance score approximating zero, current state-of-the-art models display no substantial disparity in this regard that would definitively favor or disfavor SFT, with the exception of more pronounced accuracy fluctuations observed in the latter. SFT likely forces the model to encode spurious background correlations present in the ImageNet-1K training data, exploiting the context as a shortcut to distinguish classes. Regardless of whether SFT is applied, stronger models, such as DINOv2, consistently maintain a tightly localized focus on object-relevant regions while worse performing models like ResNet-50 shift their attention maps away from the core geometry toward background structures in challenging contexts. This representational distortion is visualized in our Grad-CAM analysis (Figures 10–12).

Our results suggest that linear probing is the most reliable evaluation method for assessing the OOD robustness of a frozen backbone, as it avoids both the metric geometry assumptions of k-NN and the fine-tuning distortion of pre-fitted logit heads. Table 12 shows that training a linear probe on frozen DINOv2-B-IN1K features achieves 31.6% accuracy on the Multi-Factor Dataset, matching the raw pre-trained k-NN (31.7%) and significantly outperforming fully fine-tuned logits (28.1%) and fine-tuned k-NN (29.1%). By adapting only the classification head while keep-

ing the feature space frozen, linear probing preserves the generalizable geometry learned during self-supervised pre-training, avoiding the context bias introduced by full fine-tuning.

7.9 Theoretical Expectations and Empirical Findings

Our empirical findings largely support the theoretical distinctions established regarding how different architectural families process visual information. The results demonstrate that a model's structural architecture determines its inductive biases, which fundamentally dictate its resilience to domain shifts.

Texture Bias and Local Receptive Fields. The failure of CNNs, specifically ResNet-50, confirms theoretical concerns regarding their limited ability to generalize OOD. The local receptive fields and translation weight sharing inherent in CNNs encourage the memorization of local pixel statistics, or textures. While effective for standard benchmarks, this bias proved highly fragile under OOD rendering shifts, such as background clutter, lighting changes, and material overrides. These observations suggest that a purely local spatial hierarchy is insufficient for capturing the robust features necessary to survive severe domain collapse.

Hierarchical Attention as a Spatial Regularizer. Hierarchical ViTs, such as Swin and Hiera, occupied a middle ground in our evaluation, outperforming both CNNs and supervised standard transformers. This performance aligns with the hypothesis that a local-to-global windowed attention structure provides a more balanced inductive bias than purely local or purely global architectures. By partitioning and downsampling visual tokens across stages, these models employ the windowed partitioning as a spatial regularizer. This mechanism prevents the model from over-relying on either extremely local textures (CNNs) or fragile global pixel dependencies (supervised standard transformers).

Global Attention and the Impact of Self-Supervision. The Standard ViT architecture presents a double-edged sword regarding its global self-attention bias. In a supervised setting, this lack of built-in spatial hierarchy leads to poor generalization. However, when paired with self-supervised objectives that discourage shortcut memorization, these models learn highly robust shape representations. Our results for the DINO family confirmed the expectation of superior invariance to spurious factors. Notably, however, robustness does not always scale monotonically; the scaling anomaly observed in DINOv3-B indicates that increased model capacity can sometimes introduce unexpected vulnerabilities.

The Limits of Language Alignment. Finally, while CLIP models benefited from the robustness conferred by language alignment, they remained more sensitive to texture and material variation than initially anticipated. This indicates that while language supervision provides a strong foundation for general visual understanding,

it does not fully mitigate vulnerabilities to specific OOD factors.

In summary, while hierarchical structures offer a robust middle ground, the combination of global attention and self-supervised objectives currently provides the most resilient framework for mitigating domain shift.

8 Limitations and Future Work

Limitations The findings of this work are bounded by several technical and methodological constraints. First, the semantic scope is restricted to a specific subset of ShapeNet categories mapped to ImageNet-1K. Second, while the rendering pipeline achieves high photorealism, it focuses on static, single-object scenes. This does not account for the complexities of multi-object occlusions or dynamic environmental changes that occur in natural images. Third, the nuisance factors are sampled at discrete intervals. For instance, viewpoint variation is limited to five principal axes, omitting the bottom view and more granular intermediate rotations (see Section 4.4.1). Similarly, atmospheric effects like fog are treated as binary (on/off), which may mask more complex performance degradations under varying visibility intensities. Finally, despite the use of UE5's Lumen and Nanite, a simulation-to-reality gap persists, as evidenced by the significant performance drop for models transitioning from natural training data to synthetic test distributions.

Future Work Building on the diagnostic framework established in this thesis, several avenues for future research are proposed:
Exploiting Segmentation Masks for Enhanced Training: The current pipeline generates binary (black-and-white) segmentation masks for every rendered image, providing pixel-level ground truth. Future work should investigate if these masks can be used in mask-informed linear probing or SFT to explicitly penalize models for attending to background features. This could mitigate the context bias observed in supervised models like ResNet-50, which often shift attention toward background structures in challenging environments.
Dataset Expansion and Factor Granularity: To further bridge the gap between controlled evaluation and real-world complexity, the dataset should be expanded to include a wider range of ImageNet classes and more diverse 3D assets. Research should also move toward continuous sampling of factors, such as varying intensities of illumination and weather effects. This would enable the identification of exact failure points along a spectrum of environmental degradation.
Bigger Model Scaling (XS to Huge): The discovery of a non-monotonic scaling anomaly in DINOv3, where the Base model collapsed on the Multi-Color Dataset while Small and Large variants remained robust, suggests that robustness is not a linear function of scale. Future studies should evaluate models at further ends of the spectrum, including Extra Small (XS) and Huge sizes. Testing *Huge* variants could determine if increased capacity eventually resolves the representation phase changes

observed at the Base scale, while *XS* variants could confirm if minimal capacity inherently forces a stronger shape-based bias.
Incorporating 3D Inductive Biases: Given the near-universal failure of models to recognize objects from top-down viewpoints, there is a need for pre-training objectives that enforce 3D geometric awareness. Future work could explore training paradigms that align representation manifolds with physical 3D transformations, ensuring that latent stability translates into true viewpoint invariance.
Synthetic-to-Real Transfer Learning: Researchers could use the controlled nature of this synthetic pipeline to create targeted curriculum datasets. By pre-training or fine-tuning models on specific factors (e.g., shape-only plastic objects or varied lighting), it may be possible to improve their robustness on natural adversarial benchmarks like ObjectNet or ImageNet-A.
Investigation of a Maximum Sensitivity to Domain Shifts: The color bias result highlights a potential ceiling in the models' sensitivity to domain shifts. Although viewpoint changes and plastic-texture/color variations each substantially reduce accuracy, their combination does not lead to additive degradation, suggesting that future research should investigate whether the lack of additive degradation is due to representational collapse or a form of invariance to multiple simultaneous perturbations. Understanding this could lead to robust mechanisms for overcoming challenges posed by environmental changes and adversarial conditions.

9 Conclusion

This thesis set out to investigate the OOD robustness of modern VFMs by addressing the limitations of existing benchmarks. Traditional benchmarks rely on web-scraped natural images or diffusion-generated images, which fail to provide programmatic control over physical factors such as camera position, material properties, lighting conditions, and background context. To overcome these limitations, we designed and implemented a controlled, procedurally driven synthetic image rendering pipeline in UE5. By leveraging ShapeNet 3D models mapped to corresponding ImageNet leaf categories, we synthesized two novel evaluation datasets: the Multi-Factor and Multi-Color datasets under systematic, factor-by-factor environmental shifts. This controlled setup allowed us to examine the performance boundaries of a diverse suite of vision architectures (including CLIP, DINOv2, DINOv3, Swin, ViT, Hiera, and ResNet-50) at different levels of semantic abstraction.
Beyond controllability, our synthetic datasets provide a clean diagnostic tool for robustness studies under severe distribution shift. The rendering pipeline removes common biases (e.g., spatial center bias), by producing exact spatial annotations and segmentation masks. Automatic access to ground-truth object attributes and segmentation enables analyses, such as Grad-CAM attribution and feature-space equivariance, that are impractical on large-scale natural datasets. These properties establish synthetic evaluation as a controlled benchmark for causal interpretation and deeper model understanding.

A central finding of this research is that self-supervised pre-training (specifically DINOv2 and DINOv3) produces latent feature representations that are inherently more robust and stable under domain shifts than downstream SFT can preserve. Our representational analysis reveals that while the raw pre-trained backbones exhibit remarkable invariance to environmental parameters such as background variations, lighting shifts, and weather corruptions like fog, subsequent full SFT on ImageNet-1K distorts this geometry. Under SFT, the models adapt their features to maximize ID classification accuracy by exploiting spurious, dataset-specific shortcuts such as background textures or canonical lighting. Consequently, when evaluated OOD, their sensitivity to these nuisance factors spikes, leading to bigger accuracy swings. To address this potential representational instability, we compared the logits-based evaluation with non-parametric k-NN, and linear probing. The results demonstrate that linear probing serves as the most reliable evaluation methodology for OOD robustness. By training only a linear classifier on top of a frozen backbone, we preserve the robust representation space of self-supervised pre-training, outperforming both fully fine-tuned models and native classification heads.

Furthermore, our factor-wise analysis identifies the camera position as the single most critical factor affecting model robustness across all evaluated architectural and pre-training configurations. In particular, unusual perspectives such as top-down camera views trigger near-universal recognition failures, highlighting that current vision models struggle to construct true, viewpoint-invariant 3D representations of physical objects. Intriguingly, our equivariance analysis reveals that while the raw DINOv2 backbone maintains highly stable latent embeddings across camera rotations, this stability does not translate into classification generalization. When camera angles deviate from the canonical viewpoints dominant in web-scale training data, representation vectors drift across classifier decision boundaries. This representational drift suggests that stability in the latent space alone is insufficient, and models require a stronger geometric or physical inductive bias during pre-training to align their representation manifolds with semantic boundaries across the full 3-dimensional transformations.

Our experiments on the Multi-Color Dataset, which isolates texture simplification by replacing natural materials with uniform colored plastic, uncovered a striking non-monotonic scaling anomaly within the DINOv3 model family. While the small-capacity DINOv3-S and large-capacity DINOv3-L models maintained relatively stable, shape-biased recognition, the base-capacity DINOv3-B suffered a significant collapse, performing worse than its smaller counterpart. We theorize that this represents a capacity-driven representation phase change during self-supervised pre-training: at small scales, models can't memorize high-frequency textures and rely on low-frequency shapes; at intermediate scales (Base), they memorize discriminative textural shortcuts; and at large scales (Large), they can learn both high-frequency textures and global shape invariants. While optimization instability remains a possible alternative explanation, this non-monotonic trajectory warns against evaluating robustness at a single model scale, showing that intermediate-capacity backbones

can develop localized representational vulnerabilities.

By analyzing predictions at both the exact ImageNet-1K level and the ShapeNet superclass level, this work highlights the critical role of semantic hierarchy in robustness evaluation. Standard classification metrics that require exact fine-grained leaf class matches systematically underestimate a model's true semantic recognition capability under shift. Under severe domain shifts, models often misclassify objects into semantically related sibling classes (e.g., predicting an `office chair` as a `folding chair`), retaining baseline semantic recognition. However, our findings also reveal that this hierarchical buffer is itself vulnerable to collapse under severe domain shifts. Specifically, the plastic surface textures on the Multi-Color Dataset significantly narrow the accuracy gap between superclass and leaf-class predictions. Moreover, class-wise analysis shows that classes with rigid, highly distinctive geometric templates (such as `pianos`, `helmets`, and `guitars`) exhibit far greater semantic resilience than classes heavily reliant on texture and contextual clues (such as `buses`, `trains`, and `cabinets`), indicating a non-uniform collapse across semantic categories.

Ultimately, this work validates programmatic 3D synthetic rendering in engines like UE5 as a rigorous and necessary methodology for the next generation of computer vision benchmarks. While web-scraped OOD datasets suffer from the irreconcilable entanglement of visual factors and uncontrolled spatial and selection biases, our procedurally generated datasets offer very high causal granularity. By varying a single factor at a time while holding all others constant, we can isolate and diagnose specific failure modes in modern foundation models. Furthermore, the automatic generation of precise, pixel-level segmentation masks and ground-truth metadata enables advanced representational diagnostics (e.g., Grad-CAM spatial attribution and equivariance metrics) that are impossible to execute at scale on natural datasets.
Our pipeline thus bridges the gap between aggregate classification accuracy and deep, representation-level explainability, providing a scalable and highly controllable framework to evaluate and guide the development of future, physically grounded vision models.

A Appendix

This appendix provides auxiliary data, comprehensive model rankings, and extended qualitative examples to support the analysis presented in the main text. It includes the full semantic mapping between ShapeNet and ImageNet, complete performance tables for all evaluated model configurations, and a gallery of the synthetic images and segmentation masks generated by the pipeline.

A.1 ShapeNet

To ensure the reproducibility of the hierarchy-aware evaluation, Table 25 provides the complete mapping between ShapeNetCore unique synset IDs and their corresponding ImageNet-1K indices. This mapping serves as the basis for the superclass (ShapeNet) and fine-grained (ImageNet) evaluation protocols described in Section 5.3.

Table 25: Complete mapping of the ShapeNet 3D objects with their ImageNet indexes.

Model Name	**ShapeNet ID**	**ImageNet Index**
Old Plane	02691156	404
Pas Plane	02691156	404
Fighter Jet	02691156	895
Trash Can 1	02747177	412
Trash Can 2	02747177	412
Trash Can 3	02747177	412
Backpack 1	02773838	414
Backpack 2	02773838	414
Plastic Bag 1	02773838	728
Crate 1	02801938	519
Shopping Basket 1	02801938	790
Shopping Basket 2	02801938	790
Bathtub 1	02808440	435
Bathtub 2	02808440	435
Bathtub 3	02808440	435
Bed 1	02818832	564
Bed 2	02818832	564
Day Bed 1	02818832	831
Bench 1	02828884	703
Bench 2	02828884	703
Bench 3	02828884	703
Birdhouse 1	02843684	448
Birdhouse 2	02843684	448

Model Name	**ShapeNet ID**	**ImageNet Index**
Birdhouse 3	02843684	448
Bookshelf 1	02871439	453
Bookshelf 2	02871439	453
Bookshelf 3	02871439	453
Pill Bottle 1	02876657	720
Water Bottle 1	02876657	898
Wine Bottle 1	02876657	907
Mixing Bowl 1	02880940	659
Soup Bowl 1	02880940	809
Soup Bowl 2	02880940	809
Bus 1	02924116	779
Bus 2	02924116	779
Bus 3	02924116	779
China Cabinet 1	02933112	495
China Cabinet 2	02933112	495
China Cabinet 3	02933112	495
Reflex Camera 1	02942699	759
Reflex Camera 2	02942699	759
Reflex Camera 3	02942699	759
Can 1	02946921	737
Can 2	02946921	737
Can 3	02946921	737
Cowboy Hat 1	02954340	515
Cowboy Hat 2	02954340	515
Cowboy Hat 3	02954340	515
Limousine 1	02958343	627
Race Car 1	02958343	751
Sports Car 1	02958343	817
Folding Chair 1	03001627	559
Folding Chair 2	03001627	559
Toilet Seat 1	03001627	861
Analog Clock 1	03046257	409
Wall Clock 1	03046257	892
Wall Clock 2	03046257	892
Computer Keyboard 1	03085013	508
Computer Keyboard 2	03085013	508
Computer Keyboard 3	03085013	508
Dishwasher 1	03207941	534
Dishwasher 2	03207941	534
Dishwasher 3	03207941	534
Monitor 1	03211117	664
Monitor 2	03211117	664

Model Name	ShapeNet ID	ImageNet Index
Monitor 3	03211117	664
Earphone 1	03261776	000
Earphone 2	03261776	000
iPod 1	03261776	605
Faucet 1	03325088	000
Faucet 2	03325088	000
Faucet 3	03325088	000
File Cabinet 1	03337140	553
File Cabinet 2	03337140	553
File Cabinet 3	03337140	553
Acoustic Guitar 1	03467517	402
Acoustic Guitar 2	03467517	402
Electric Guitar 1	03467517	546
Crash Helmet 1	03513137	518
Crash Helmet 2	03513137	518
Football Helmet 1	03513137	560
Vase 1	03593526	883
Vase 2	03593526	883
Vase 3	03593526	883
Knife 1	03624134	623
Knife 2	03624134	623
Knife 3	03624134	623
Table Lamp 1	03636649	846
Table Lamp 2	03636649	846
Table Lamp 3	03636649	846
Laptop 1	03642806	620
Laptop 2	03642806	620
Laptop 3	03642806	620
Loudspeaker 1	03691459	632
Loudspeaker 2	03691459	632
Loudspeaker 3	03691459	632
Mailbox 1	03710193	637
Mailbox 2	03710193	637
Mailbox 3	03710193	637
Microphone 1	03759954	650
Microphone 2	03759954	650
Microphone 3	03759954	650
Microwave 1	03761084	651
Microwave 2	03761084	651
Microwave 3	03761084	651
Moped 1	03790512	665
Moped 2	03790512	665

Model Name	ShapeNet ID	ImageNet Index
Moped 3	03790512	665
Mug 1	03797390	504
Mug 2	03797390	504
Mug 3	03797390	504
Grand Piano 1	03928116	579
Upright Piano 1	03928116	881
Upright Piano 2	03928116	881
Pillow 1	03938244	721
Pillow 2	03938244	721
Pillow 3	03938244	721
Revolver 1	03948459	763
Revolver 2	03948459	763
Revolver 3	03948459	763
Flowerpot 1	03991062	738
Flowerpot 2	03991062	738
Flowerpot 3	03991062	738
Printer 1	04004475	742
Printer 2	04004475	742
Printer 3	04004475	742
Remote 1	04074963	761
Remote 2	04074963	761
Remote 3	04074963	761
Assault Rifle 1	04090263	413
Rifle 1	04090263	764
Rifle 2	04090263	764
Missile 1	04099429	657
Missile 2	04099429	657
Space Shuttle 1	04099429	812
Skateboard 1	04225987	000
Skateboard 2	04225987	000
Skateboard 3	04225987	000
Sofa 1	04256520	000
Sofa 2	04256520	000
Sofa 3	04256520	000
Stove 1	04330267	827
Stove 2	04330267	827
Stove 3	04330267	827
Dining Table 1	04379243	532
Dining Table 2	04379243	532
Dining Table 3	04379243	532
Mobile Phone 1	04401088	487
Mobile Phone 2	04401088	487

Model Name	ShapeNet ID	ImageNet Index
Mobile Phone 3	04401088	487
Mosque 1	04460130	668
Obelisk 1	04460130	682
Water Tower 1	04460130	900
Bullet Train 1	04468005	466
Electric Locomotive 1	04468005	547
Steam Locomotive 1	04468005	820
Ocean Liner 1	04530566	628
Speedboat 1	04530566	814
Yawl 1	04530566	914
Washer 1	04554684	897
Washer 2	04554684	897
Washer 3	04554684	897

A.2 Datasets

Figures 13–14 provide an extended qualitative gallery of the synthetic datasets. These examples illustrate the diversity of the ShapeNet object subset and the systematic application of nuisance factors, such as viewpoint, material overrides, and background environments across both the Multi-Factor and Multi-Color datasets.

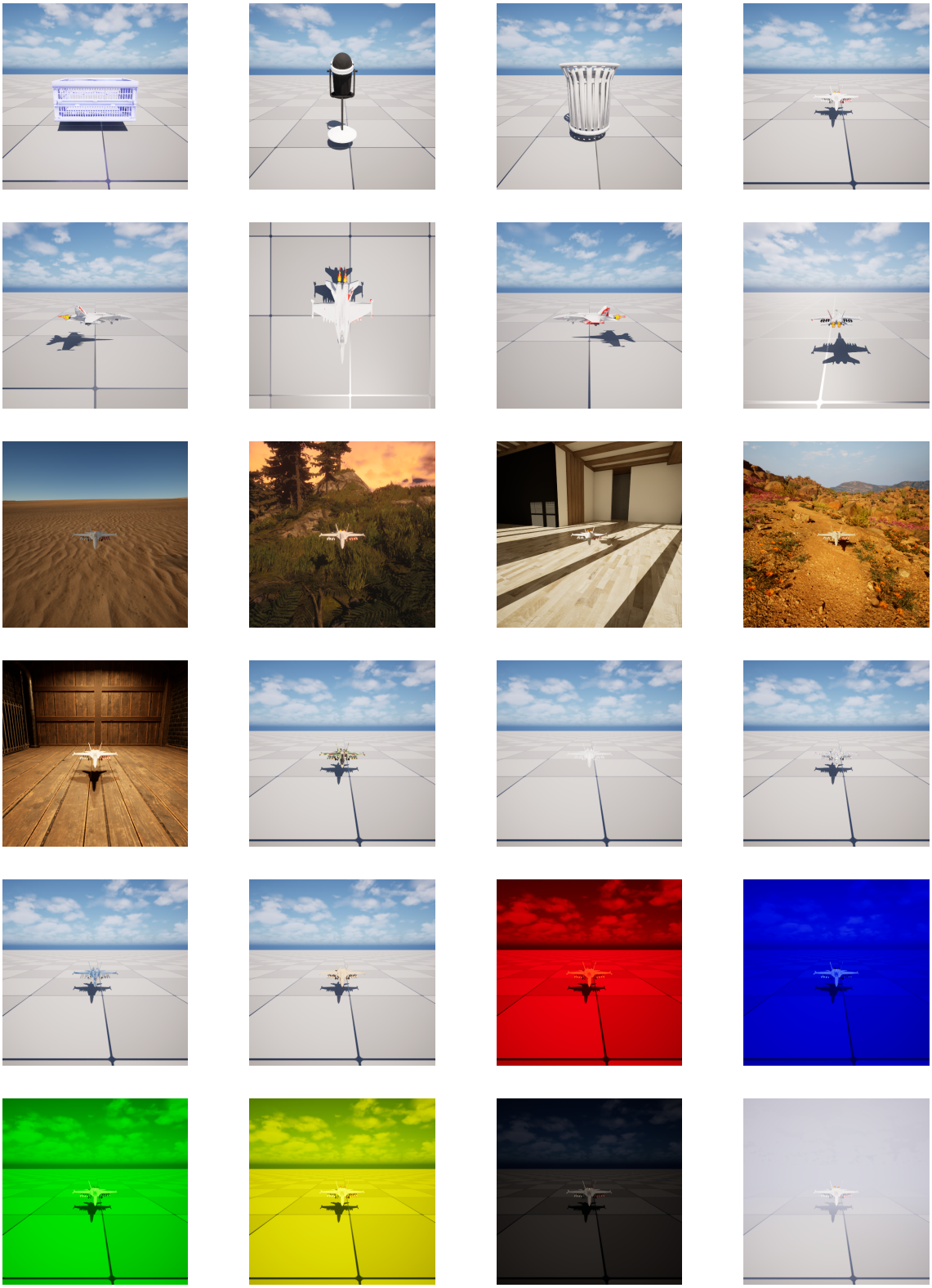


Figure 13: Qualitative examples from the Multi-Factor Dataset (full factor list). Each image shows the original, unzoomed inputs across different objects (Basket, Microphone, Trash Bin, Fighter Jet), viewpoints (Front, Left, Top, Right, Back), backgrounds (Plain Desert, Forrest Hills, Living Room, Rocky Desert, Basement), surface materials (Camouflage, Glass, Noise, Titanium, Wood), lighting conditions (Red, Blue, Green, Yellow, Dim Black), and atmospheric effects (fog).

Figure 14: Qualitative examples from the Multi-Color Dataset (full factor list). Each image shows the original, unzoomed inputs across different objects (Basket, Microphone, Trash Bin, Fighter Jet), viewpoints (Front, Left, Top, Right, Back), and plastic colors (blue, cyan, green, pink, red, yellow, black, grey, light grey).

A.3 Results

While Section 6 provides a summarized view of model performance, Tables 26 and 27 present the exhaustive performance rankings for all 45 evaluated model-method combinations. These tables include Top-1 and Top-5 ImageNet accuracy, as well as superclass-level results, providing a granular view of how different architectures, scaling sizes, and evaluation modes (logits, k-NN, linear probing) behave under severe domain shift.

This section contains all the tables and figures that are displayed shortened (Best 5 and Worst 5) in the Results section.

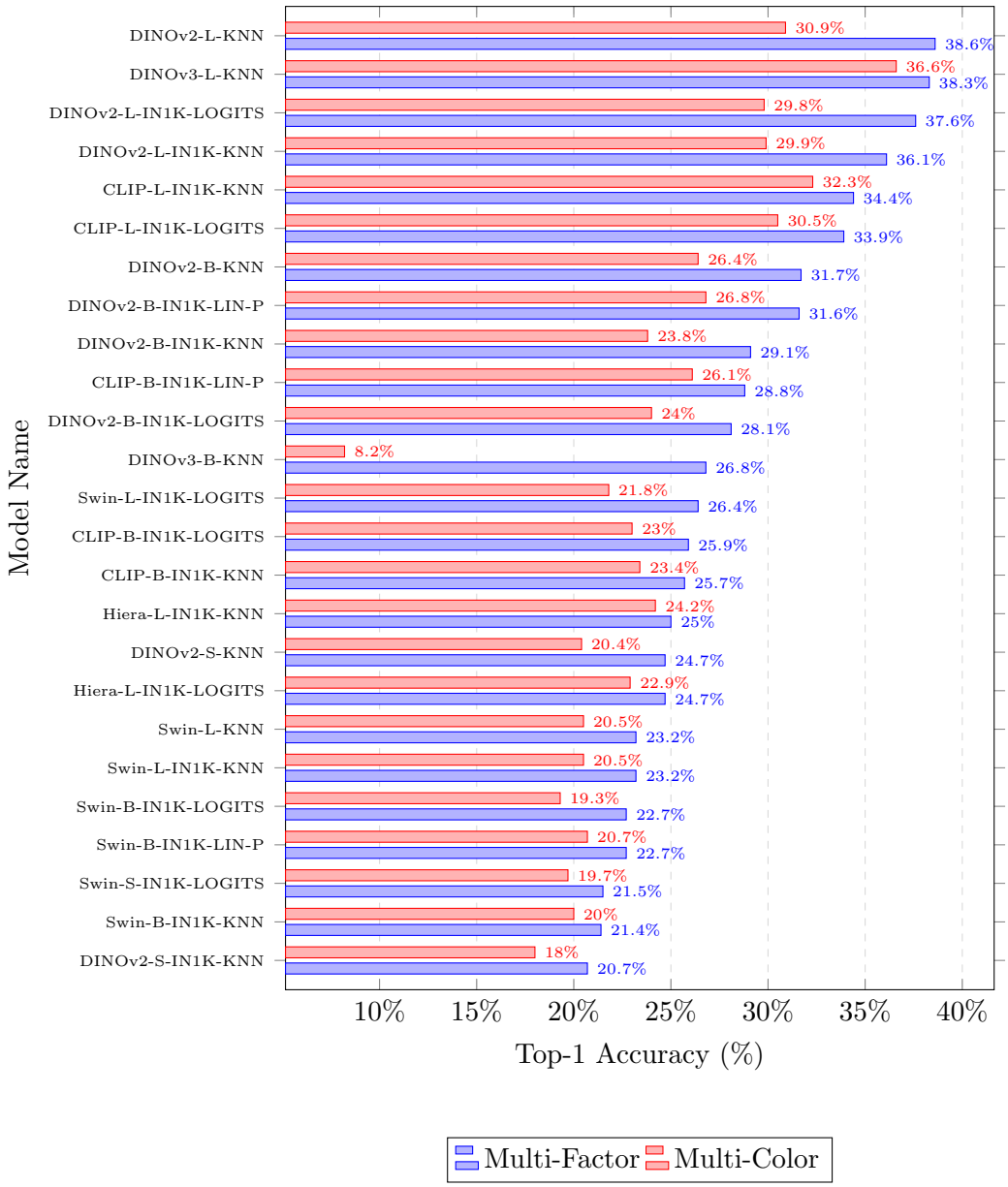


Figure 15: Comparison of ImageNet Top-1 Accuracy (%) across datasets for the top 25 models.

Model Name	IN1K Validation	Imagenet Top-1	Imagenet Top-5	Shapenet Base Top-1	Shapenet Top-1
DINOv2-L-KNN	81.1%	**38.6%**	**60.6%**	59.7%	**52.7%**
DINOv3-L-KNN	83.3%	38.3%	55.4%	58.4%	51.3%
DINOv2-L-IN1K-LOGITS	85.9%	37.6%	59.8%	**65.1%**	52.2%
DINOv2-L-IN1K-KNN	81.6%	36.1%	57.2%	61.7%	49.2%
CLIP-L-IN1K-KNN	**87.5%**	34.4%	49.2%	55.0%	45.5%
CLIP-L-IN1K-LOGITS	87.3%	33.9%	55.8%	52.3%	44.4%
DINOv2-B-KNN	79.7%	31.7%	53.2%	53.7%	44.4%
DINOv2-B-IN1K-LIN-P	78.1%	31.6%	55.5%	50.3%	42.1%
DINOv2-B-IN1K-KNN	80.5%	29.1%	51.3%	53.7%	41.0%
CLIP-B-IN1K-LIN-P	83.8%	28.8%	46.3%	47.0%	37.4%
DINOv2-B-IN1K-LOGITS	84.5%	28.1%	51.8%	53.0%	40.9%
DINOv3-B-KNN	79.2%	26.8%	48.0%	45.6%	38.2%
Swin-L-IN1K-LOGITS	86.3%	26.4%	45.5%	47.0%	35.8%
CLIP-B-IN1K-LOGITS	85.5%	25.9%	44.9%	45.6%	35.2%
CLIP-B-IN1K-KNN	84.6%	25.7%	41.5%	48.3%	35.2%
Hiera-L-IN1K-KNN	86.2%	25.0%	37.6%	48.3%	34.1%
DINOv2-S-KNN	76.1%	24.7%	46.5%	39.6%	33.5%
Hiera-L-IN1K-LOGITS	86.0%	24.7%	44.3%	47.7%	33.7%
Swin-L-KNN	85.3%	23.2%	37.1%	39.6%	31.4%
Swin-L-IN1K-KNN	85.3%	23.2%	37.1%	39.6%	31.4%
Swin-B-IN1K-LOGITS	85.3%	22.7%	41.1%	43.6%	31.7%
Swin-B-IN1K-LIN-P	83.9%	22.7%	41.4%	44.3%	30.6%
Swin-S-IN1K-LOGITS	83.3%	21.5%	40.6%	38.9%	29.5%
Swin-B-IN1K-KNN	84.4%	21.4%	35.8%	40.9%	29.3%
DINOv2-S-IN1K-KNN	76.3%	20.7%	41.3%	38.3%	30.1%
DINOv2-S-IN1K-LOGITS	82.1%	20.7%	44.4%	41.6%	30.8%
Hiera-B-LOGITS	84.5%	20.3%	38.3%	36.9%	27.8%
Swin-B-KNN	83.6%	20.2%	32.6%	38.9%	27.3%
Hiera-B-KNN	84.3%	20.1%	33.7%	39.6%	27.5%
Swin-S-IN1K-KNN	81.9%	19.5%	34.3%	34.2%	26.3%
Hiera-S-IN1K-LOGITS	83.9%	19.0%	36.4%	37.6%	26.5%
Swin-S-KNN	82.8%	18.4%	30.8%	36.2%	24.6%
Hiera-S-IN1K-KNN	83.3%	17.6%	32.0%	35.6%	24.1%
CLIP-L-KNN	74.4%	17.4%	32.1%	36.2%	24.3%
ViT-L-IN1K-LOGITS	82.0%	15.8%	31.9%	30.9%	21.8%
ViT-L-IN1K-KNN	80.5%	15.7%	27.6%	29.5%	20.7%
DINOv3-S-KNN	72.6%	15.5%	32.5%	30.2%	22.1%
ViT-B-IN1K-LOGITS	80.3%	14.6%	29.6%	32.2%	19.9%
ViT-B-IN1K-KNN	78.4%	13.2%	24.1%	24.2%	18.1%
ResNet-50-IN1K-LOGITS	80.4%	12.1%	26.9%	26.2%	17.1%
ResNet-50-IN1K-KNN	77.5%	9.2%	18.4%	21.5%	13.4%
DINOv1-B-KNN	70.2%	8.7%	21.9%	16.8%	13.5%
CLIP-B-KNN	67.1%	8.6%	19.6%	16.8%	13.0%
ResNet-50-KNN	74.0%	7.7%	17.4%	19.5%	12.0%
DINOv1-S-KNN	68.6%	7.3%	18.2%	16.8%	12.0%

Table 26: Model Rankings on the Multi-Factor Dataset. IN1K is the Top 1 Accuracy for the models run on the ImageNet-1K Evaluation Split Dataset, while ImageNet Top-1 is the Top 1 Accuracy for the models run on our Multi-Factor Dataset.

Model Name	IN1K Validation	Imagenet Top-1	Imagenet Top-5	Shapenet Base Top-1	Shapenet Top-1
DINOv3-L-KNN	83.3%	**36.6%**	**54.4%**	58.4%	**48.6%**
CLIP-L-IN1K-KNN	**87.5%**	32.3%	45.4%	55.0%	42.6%
DINOv2-L-KNN	81.1%	30.9%	52.1%	59.7%	44.6%
CLIP-L-IN1K-LOGITS	87.3%	30.5%	50.9%	52.3%	40.6%
DINOv2-L-IN1K-KNN	81.6%	29.9%	49.9%	61.7%	42.6%
DINOv2-L-IN1K-LOGITS	85.9%	29.8%	52.3%	**65.1%**	42.0%
DINOv2-B-IN1K-LIN-P	78.1%	26.8%	48.4%	50.3%	36.7%
DINOv2-B-KNN	79.7%	26.4%	46.4%	53.7%	38.0%
CLIP-B-IN1K-LIN-P	83.8%	26.1%	41.6%	47.0%	33.7%
Hiera-L-IN1K-KNN	86.2%	24.2%	37.6%	48.3%	32.9%
DINOv2-B-IN1K-LOGITS	84.5%	24.0%	45.7%	53.0%	36.2%
DINOv2-B-IN1K-KNN	80.5%	23.8%	44.2%	53.7%	34.4%
CLIP-B-IN1K-KNN	84.6%	23.4%	37.7%	48.3%	31.8%
CLIP-B-IN1K-LOGITS	85.5%	23.0%	40.1%	45.6%	31.0%
Hiera-L-IN1K-LOGITS	86.0%	22.9%	41.1%	47.7%	30.5%
Swin-L-IN1K-LOGITS	86.3%	21.8%	40.0%	47.0%	30.7%
Swin-B-IN1K-LIN-P	83.9%	20.7%	39.3%	44.3%	28.7%
Swin-L-KNN	85.3%	20.5%	36.1%	39.6%	28.4%
Swin-L-IN1K-KNN	85.3%	20.5%	36.1%	39.6%	28.4%
Swin-S-IN1K-KNN	81.9%	20.4%	35.7%	34.2%	27.5%
DINOv2-S-KNN	76.1%	20.4%	41.7%	39.6%	30.0%
Hiera-B-KNN	84.3%	20.1%	32.7%	39.6%	26.3%
Swin-B-IN1K-KNN	84.4%	20.0%	34.8%	40.9%	27.6%
Swin-S-IN1K-LOGITS	83.3%	19.7%	37.0%	38.9%	26.7%
Swin-B-IN1K-LOGITS	85.3%	19.3%	37.8%	43.6%	27.5%
DINOv2-S-IN1K-LOGITS	82.1%	18.5%	39.1%	41.6%	28.5%
Hiera-S-IN1K-KNN	83.3%	18.5%	31.8%	35.6%	23.9%
DINOv2-S-IN1K-KNN	76.3%	18.0%	38.1%	38.3%	26.9%
Hiera-B-LOGITS	84.5%	17.7%	35.1%	36.9%	23.8%
Swin-B-KNN	83.6%	17.4%	31.5%	38.9%	24.3%
ViT-L-IN1K-KNN	80.5%	16.9%	29.7%	29.5%	22.4%
Swin-S-KNN	82.8%	16.8%	29.8%	36.2%	24.1%
Hiera-S-IN1K-LOGITS	83.9%	16.4%	31.5%	37.6%	22.1%
CLIP-L-KNN	74.4%	16.4%	30.8%	36.2%	23.1%
ViT-L-IN1K-LOGITS	82.0%	16.0%	34.3%	30.9%	22.2%
DINOv3-S-KNN	72.6%	15.0%	33.2%	30.2%	22.0%
ViT-B-IN1K-LOGITS	80.3%	14.8%	31.1%	32.2%	20.4%
ViT-B-IN1K-KNN	78.4%	13.6%	26.6%	24.2%	19.4%
ResNet-50-IN1K-LOGITS	80.4%	10.6%	25.1%	26.2%	16.2%
DINOv1-B-KNN	70.2%	8.8%	22.1%	16.8%	14.7%
ResNet-50-KNN	74.0%	8.8%	18.8%	19.5%	12.9%
ResNet-50-IN1K-KNN	77.5%	8.6%	17.6%	21.5%	12.4%
CLIP-B-KNN	67.1%	8.4%	20.4%	16.8%	13.5%
DINOv3-B-KNN	79.2%	8.2%	17.0%	45.6%	13.0%
DINOv1-S-KNN	68.6%	7.0%	18.5%	16.8%	12.3%

Table 27: Model Rankings on the Multi-Color Dataset. IN1K is the Top 1 Accuracy for the models run on the ImageNet-1K Evaluation Split Dataset, while ImageNet Top-1 is the Top 1 Accuracy for the models run on our Multi-Color Dataset.

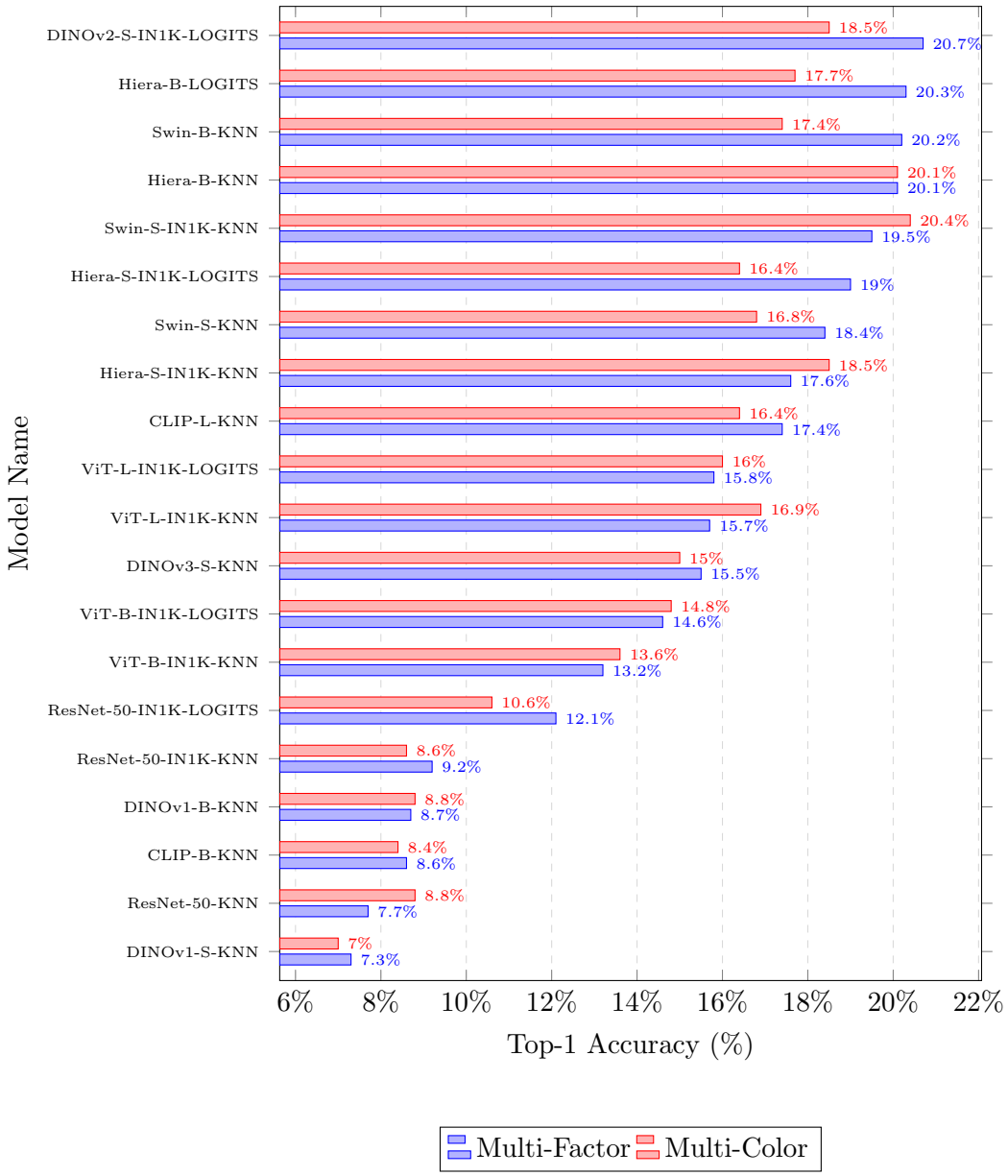


Figure 16: Comparison of ImageNet Top-1 Accuracy (%) across datasets for the remaining models.

Model	IN-1K	IN-V2	IN-R	IN-Sk	IN-A	ON	PUG IN	MF	MC
ResNet-50-IN1K	77.5%	62.5%	21.6%	24.2%	2.1%	15.2%	24.4%	9.2%	8.6%
ViT-L-IN1K	80.5%	66.4%	34.3%	37.3%	14.0%	15.7%	30.9%	15.7%	16.9%
Swin-L-IN1K	85.3%	74.2%	46.2%	46.5%	37.1%	27.1%	44.5%	23.2%	20.5%
CLIP-B-IN1K	84.6%	74.1%	51.4%	54.5%	22.6%	26.8%	46.6%	25.7%	23.4%
CLIP-L-IN1K	87.5%	77.7%	65.7%	64.2%	38.7%	32.7%	53.7%	34.4%	32.3%
DINOv2-B	79.7%	69.5%	50.5%	52.3%	52.5%	27.2%	46.1%	31.7%	26.4%
DINOv2-L	81.1%	71.3%	61.8%	59.4%	62.2%	31.6%	52.7%	38.6%	30.9%
DINOv3-B	79.2%	53.8%	36.9%	38.6%	9.0%	10.9%	40.8%	26.8%	8.2%
DINOv3-L	83.3%	73.1%	71.7%	65.4%	58.1%	35.0%	56.6%	38.3%	36.6%

Table 28: Comparison of top-performing and baseline vision k-NN models across standard robustness datasets and our synthetic Multi-Factor/Multi-Color benchmarks. Accuracies are reported as Top-1 accuracy (%). Abbreviations: IN-1K = ImageNet-1K, IN-V2 = ImageNet-V2, IN-R = ImageNet-R, IN-Sk = ImageNet-Sketch, IN-A = ImageNet-A, ON = ObjectNet, PUG IN = PUG ImageNet, MF = Multi-Factor Dataset (ours), MC = Multi-Color Dataset (ours)

Model	Multi-Factor-Dataset	Multi-Color-Dataset
CLIP-B-KNN	1.503	1.606
CLIP-B-IN1K-KNN	1.367	1.357
CLIP-B-IN1K-LIN-P	1.301	1.290
CLIP-B-IN1K-LOGITS	1.362	1.349
CLIP-L-KNN	1.396	1.407
CLIP-L-IN1K-KNN	1.322	1.317
CLIP-L-IN1K-LOGITS	1.309	1.329
DINOv1-B-KNN	1.548	1.660
DINOv1-S-KNN	1.641	1.764
DINOv2-B-KNN	1.400	1.439
DINOv2-B-IN1K-KNN	1.408	1.443
DINOv2-B-IN1K-LIN-P	1.331	1.368
DINOv2-B-IN1K-LOGITS	1.457	1.509
DINOv2-L-KNN	1.364	1.444
DINOv2-L-IN1K-KNN	1.364	1.426
DINOv2-L-IN1K-LOGITS	1.385	1.407
DINOv2-S-KNN	1.353	1.472
DINOv2-S-IN1K-KNN	1.457	1.491
DINOv2-S-IN1K-LOGITS	1.491	1.536
DINOv3-B-KNN	1.426	1.588
DINOv3-L-KNN	1.339	1.328
DINOv3-S-KNN	1.424	1.469
Hiera-B-KNN	1.369	1.305
Hiera-B-LOGITS	1.366	1.344
Hiera-L-IN1K-KNN	1.366	1.361
Hiera-L-IN1K-LOGITS	1.364	1.334
Hiera-S-IN1K-KNN	1.364	1.288
Hiera-S-IN1K-LOGITS	1.399	1.348
ResNet-50-KNN	1.557	1.467
ResNet-50-IN1K-KNN	1.451	1.437
ResNet-50-IN1K-LOGITS	1.414	1.523
Swin-B-KNN	1.353	1.400
Swin-B-IN1K-KNN	1.373	1.376
Swin-B-IN1K-LIN-P	1.350	1.382
Swin-B-IN1K-LOGITS	1.396	1.425
Swin-L-KNN	1.352	1.383
Swin-L-IN1K-KNN	1.352	1.383
Swin-L-IN1K-LOGITS	1.355	1.407
Swin-S-KNN	1.342	1.435
Swin-S-IN1K-KNN	1.351	1.351
Swin-S-IN1K-LOGITS	1.367	1.354
ViT-B-IN1K-KNN	1.375	1.425
ViT-B-IN1K-LOGITS	1.363	1.381
ViT-L-IN1K-KNN	1.325	1.328
ViT-L-IN1K-LOGITS	1.379	1.394
Average Ratio	1.418	1.418

Table 29: Superclass to Fine-grained classification accuracy ratio per model.

ShapeNet Superclass	Avg IN Top-1	Avg SN Top-1	Ratio (SN/IN)
piano	62.8%	69.6%	1.11x
guitar	61.8%	67.7%	1.10x
clock	28.1%	67.1%	2.39x
chair	33.4%	65.6%	1.96x
trash bin	64.5%	64.5%	1.00x
laptop	34.7%	61.2%	1.76x
sofa	60.8%	60.8%	1.00x
birdhouse	54.6%	54.6%	1.00x
lamp	50.2%	50.2%	1.00x
helmet	47.5%	49.2%	1.03x
car	8.6%	49.1%	5.71x
bench	48.2%	48.2%	1.00x
washer	47.3%	47.3%	1.00x
bottle	27.1%	46.7%	1.72x
remote	42.9%	42.9%	1.00x
stove	38.7%	38.7%	1.00x
basket	31.9%	37.9%	1.19x
microwaves	37.4%	37.4%	1.00x
table	17.5%	36.0%	2.06x
bowl	6.1%	34.4%	5.67x
airplane	25.0%	34.2%	1.37x
jar	31.8%	34.0%	1.07x
bookshelf	32.5%	32.5%	1.00x
display	21.6%	31.5%	1.46x
bathtub	30.8%	30.8%	1.00x
dishwasher	28.9%	28.9%	1.00x
printer	25.9%	25.9%	1.00x
motorbike	12.3%	25.5%	2.07x
cap	23.8%	24.4%	1.03x
rocket	10.7%	22.5%	2.10x
file cabinet	22.1%	22.1%	1.00x
loudspeaker	21.2%	21.2%	1.00x
bag	10.1%	18.4%	1.84x
bed	3.9%	18.2%	4.65x
earphone	15.3%	15.3%	1.00x
keyboard	14.7%	14.7%	1.00x
microphone	14.7%	14.7%	1.00x
telephone	12.6%	14.5%	1.16x
mug	14.5%	14.5%	1.00x
tower	5.8%	14.3%	2.47x
flowerpot	14.2%	14.2%	1.00x
watercraft	3.1%	14.1%	4.51x
mailbox	13.5%	13.5%	1.00x
camera	12.0%	13.2%	1.11x
pillow	11.4%	11.4%	1.00x
rifle	6.6%	10.5%	1.59x
pistol	8.6%	8.6%	1.00x
knife	2.2%	3.3%	1.52x
bus	0.8%	2.4%	2.88x
train	1.6%	2.2%	1.35x
cabinet	0.1%	1.1%	15.00x

Table 30: Average Top-1 ImageNet and ShapeNet accuracy by object category (Multi-Factor Dataset).

ShapeNet Superclass	Avg IN Top-1	Avg SN Top-1	Ratio (SN/IN)
laptop	44.0%	65.8%	1.50x
trash bin	63.1%	63.1%	1.00x
chair	36.5%	61.3%	1.68x
airplane	39.8%	54.3%	1.37x
helmet	49.4%	53.2%	1.08x
bowl	11.9%	52.1%	4.38x
car	14.5%	50.0%	3.46x
lamp	49.8%	49.8%	1.00x
motorbike	26.9%	44.2%	1.64x
keyboard	43.5%	43.5%	1.00x
rocket	19.5%	42.7%	2.19x
rifle	22.9%	42.0%	1.84x
piano	36.0%	42.0%	1.17x
bench	41.8%	41.8%	1.00x
basket	34.2%	37.2%	1.09x
jar	35.1%	37.2%	1.06x
pistol	34.6%	34.6%	1.00x
bottle	21.9%	33.2%	1.52x
guitar	28.3%	33.0%	1.17x
table	22.3%	32.4%	1.45x
sofa	32.1%	32.1%	1.00x
bag	19.3%	31.6%	1.63x
birdhouse	31.2%	31.2%	1.00x
mailbox	30.3%	30.3%	1.00x
watercraft	10.2%	29.8%	2.93x
bathtub	29.3%	29.3%	1.00x
microphone	22.6%	22.6%	1.00x
tower	6.1%	19.9%	3.24x
cap	19.3%	19.9%	1.03x
clock	8.8%	19.7%	2.24x
pillow	19.0%	19.0%	1.00x
mug	16.8%	16.8%	1.00x
display	12.5%	16.4%	1.31x
flowerpot	15.9%	15.9%	1.00x
washer	14.8%	14.8%	1.00x
stove	13.9%	13.9%	1.00x
camera	10.5%	11.5%	1.09x
loudspeaker	11.2%	11.2%	1.00x
printer	11.1%	11.1%	1.00x
remote	10.7%	10.7%	1.00x
train	8.7%	9.8%	1.13x
knife	7.4%	9.7%	1.32x
bed	3.0%	9.1%	3.04x
dishwasher	8.8%	8.8%	1.00x
bookshelf	8.3%	8.3%	1.00x
file cabinet	8.1%	8.1%	1.00x
earphone	7.0%	7.0%	1.00x
microwaves	3.8%	3.8%	1.00x
telephone	1.5%	3.5%	2.30x
bus	1.4%	3.1%	2.23x
cabinet	0.0%	0.5%	N/A

Table 31: Average Top-1 ImageNet and ShapeNet accuracy by object category (Multi-Color Dataset).

A.4 Segmentation Masks

As detailed in the synthetic rendering pipeline (Section 4, every photorealistic image is paired with a binary segmentation mask. Figure 17 displays representative masks for various object categories, demonstrating the high-fidelity, pixel-level ground truth used to isolate foreground geometry from complex backgrounds for spatial attribution and future dense prediction tasks.

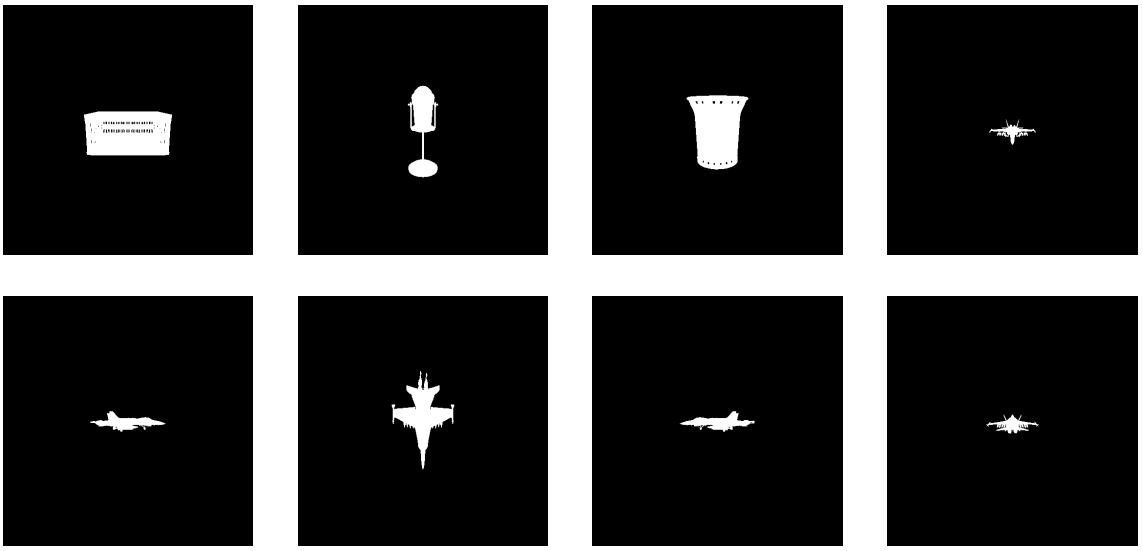


Figure 17: Qualitative examples from the generated Segmentation Masks in the Multi-Factor Dataset (full factor list). Each image shows the original mask inputs across different objects (Basket, Microphone, Trash Bin, Fighter Jet), viewpoints (Front, Left, Top, Right, Back).

Bibliography

Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects, April 2019. URL `http://arxiv.org/abs/1811.11553`. arXiv:1811.11553 [cs].

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation Models Defining a New Era in Vision: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, April 2025. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3506283. URL `https://ieeexplore.ieee.org/document/10834497/`.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic Data from Diffusion Models Improves ImageNet Classification, April 2023. URL `http://arxiv.org/abs/2304.08466`. arXiv:2304.08466 [cs].

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9448–9458. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/9142-objectnet-a-large-scale-bias-controlled-dataset-for-pushing-the-limits-of-object-recognition-models.pdf`.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krish-

nan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, August 2021. URL `https://ui.adsabs.harvard.edu/abs/2021arXiv210807258B`. ADS Bibcode: 2021arXiv210807258B.

Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari S. Morcos. PUG: Photorealistic and Semantically Controllable Synthetic Data for Representation Learning, December 2023. URL `http://arxiv.org/abs/2308.03977`. arXiv:2308.03977 [cs].

Mike Bostock. ImageNet Hierarchy / Mike Bostock, 2018. URL `https://observablehq.com/@mbostock/imagenet-hierarchy`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. URL `http://arxiv.org/abs/2104.14294`. arXiv:2104.14294 [cs].

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository, December 2015. URL `http://arxiv.org/abs/1512.03012`. arXiv:1512.03012 [cs].

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.00276. URL `https://ieeexplore.ieee.org/document/10205297/`.

Cycu5. Fantasy ruins demo unreal engine 5 - standard pipeline vs nanite/lumen rtx 4080 graphics comparison, 2024. URL `https://youtu.be/JPPp-_PBhzk`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. URL `https://ieeexplore.ieee.org/document/5206848/`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL `http://arxiv.org/abs/2010.11929`. arXiv:2010.11929 [cs].

Olaf Dünkel, Artur Jesslen, Jiahao Xie, Christian Theobalt, Christian Rupprecht, and Adam Kortylewski. CNS-Bench: Benchmarking Image Classifier Robustness Under Continuous Nuisance Shifts, September 2025. URL http://arxiv.org/abs/2507.17651. arXiv:2507.17651 [cs].

Inc. Epic Games. Unreal Engine 5.7 is here! Find out what's new and download today., 2025. URL https://www.unrealengine.com/news/unreal-engine-5-7-is-now-available.

Inc. Epic Games. Fab, Retrieved May, 2026a. URL https://www.fab.com/.

Inc. Epic Games. Unreal Engine 5 is now available!, Retrieved May, 2026b. URL https://www.unrealengine.com/blog/unreal-engine-5-is-now-available.

Inc. Epic Games. Unreal Engine 5.7 Documentation | Unreal Engine 5.7 Documentation | Epic Developer Community, Retrieved May, 2026c. URL https://dev.epicgames.com/documentation/unreal-engine/unreal-engine-5-7-documentation.

Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks, October 2020. URL http://arxiv.org/abs/1808.08750. arXiv:1808.08750 [cs].

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, November 2022. URL http://arxiv.org/abs/1811.12231. arXiv:1811.12231 [cs].

Florian Gutbier. Design and Implementation of an Unreal Engine 5 Plugin for Generating Photorealistic and Semantically Controllable Synthetic Data to Evaluate the Robustness of ImageNet Classifier. 2025.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition?, February 2023. URL http://arxiv.org/abs/2210.07574. arXiv:2210.07574 [cs].

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, March 2019. URL http://arxiv.org/abs/1903.12261. arXiv:1903.12261 [cs.LG].

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, October 2018. URL http://arxiv.org/abs/1610.02136. arXiv:1610.02136 [cs].

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness:

A Critical Analysis of Out-of-Distribution Generalization, July 2021a. URL http://arxiv.org/abs/2006.16241. arXiv:2006.16241 [cs.CV].

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples, March 2021b. URL http://arxiv.org/abs/1907.07174. arXiv:1907.07174 [cs.LG].

Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. ImageNet-X: Understanding Model Mistakes with Factor of Variation Annotations, November 2022. URL http://arxiv.org/abs/2211.01866. arXiv:2211.01866 [cs].

Brian Karis, Rune Stubble, and Graham Wihidal. Nanite-A Deep Dive, 2021. URL https://advances.realtimerendering.com/s2021/Karis_Nanite_SIGGRAPH_Advances_2021_final.pdf.

Steven Lang, Felipe Bravo-Marquez, Christopher Beckham, Mark Hall, and Eibe Frank. IMAGENET 1000 Class List - WekaDeeplearning4j, Retrieved January, 2026. URL https://deeplearning.cms.waikato.ac.nz/user-guide/class-maps/IMAGENET/.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. URL http://arxiv.org/abs/2103.14030. arXiv:2103.14030 [cs].

Hashmat Shadab Malik, Muhammad Huzaifa, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. ObjectCompose: Evaluating Resilience of Vision-Based Models on Object-to-Background Compositional Changes, October 2024. URL http://arxiv.org/abs/2403.04701. arXiv:2403.04701 [cs].

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, Ziwei Liu, Toshihiko Yamasaki, and Kiyoharu Aizawa. Generalized Out-of-Distribution Detection and Beyond in Vision Language Model Era: A Survey, June 2025. URL http://arxiv.org/abs/2407.21794. arXiv:2407.21794 [cs].

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

Ameya Prabhu, Vishaal Udandarao, Philip H S Torr, Matthias Bethge, Adel Bibi, and Samuel Albanie. Efficient Lifelong Model Evaluation in an Era of Rapid Progress. 2024. URL https://arxiv.org/abs/2402.19472.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet?, June 2019. URL http://arxiv.org/abs/1902.10811. arXiv:1902.10811 [cs.CV].

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, January 2015. URL http://arxiv.org/abs/1409.0575. arXiv:1409.0575 [cs].

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ILSVRC/imagenet-1k · Datasets at Hugging Face, October Retrieved April, 2025. URL https://huggingface.co/datasets/ILSVRC/imagenet-1k.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles, June 2023. URL https://arxiv.org/abs/2306.00989v1.

Carlos N. Silla and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, January 2011. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-010-0175-9. URL http://link.springer.com/10.1007/s10618-010-0175-9.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025. URL http://arxiv.org/abs/2508.10104. arXiv:2508.10104 [cs].

Mohammad Reza Taesiri, Giang Nguyen, Sarra Habchi, Cor-Paul Bezemer, and Anh Nguyen. ImageNet-Hard: The Hardest Images Remaining from a Study of the

Power of Zoom and Spatial Biases in Image Classification, October 2023. URL http://arxiv.org/abs/2304.05538. arXiv:2304.05538 [cs.CV].

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. ID and OOD Performance Are Sometimes Inversely Correlated on Real-world Datasets, May 2023. URL http://arxiv.org/abs/2209.00613. arXiv:2209.00613 [cs].

Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning Robust Global Representations by Penalizing Local Predictive Power, November 2019. URL http://arxiv.org/abs/1905.13549. arXiv:1905.13549 [cs.CV].

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or Signal: The Role of Image Backgrounds in Object Recognition, June 2020. URL http://arxiv.org/abs/2006.09994. arXiv:2006.09994 [cs].

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized Out-of-Distribution Detection: A Survey. *International Journal of Computer Vision*, 132(12):5635–5662, December 2024. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-024-02117-4. URL https://link.springer.com/10.1007/s11263-024-02117-4.

Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, and Chengzhi Mao. ImageNet-D: Benchmarking Neural Network Robustness on Diffusion Synthetic Object. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21752–21762, June 2024. doi: 10.1109/CVPR52733.2024.02055. URL https://ieeexplore.ieee.org/document/10655614/.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Bamberg, 03.08.2026
Place, Date

Signature