Understanding Inductive and Model Biases in Deep Representation Learning

Raahim A Samad Poonawala Muhammad Sabeer Asad Muhammad Bin Tariq

Abstract

This report presents an empirical analysis of fundamental inductive biases across five major deep learning architectures and their impact on out-of-distribution generalization. We investigate inductive and architectural biases across CNNs and Vision Transformers (ViTs), fidelity-diversity tradeoffs in latent generative models, and multimodal contrastive biases in CLIP. Our study reveals that CNNs heavily rely on local cues (e.g. textures), whereas ViTs and CLIP learn domain independent features, increasing their robustness to domain shifts. Our results suggest that mimicking human intuition for classification and incorporating such biases into model architecture or datasets may improve OOD generalization. Moreover, our findings show that generative models' inductive biases are largely characteristic of their inherent architecture, and these biases degrade generalization performance while their strengths can be leveraged for specialized applications.

All of our findings and code can be accessed here: GitHub Repository

1. Introduction

Deep neural networks often fail to generalize when deployed on data that differs from their training distribution—a phenomenon known as distribution shift. A key factor determining generalization behavior is the **inductive bias** encoded in models — assumptions built into model architectures, training objectives, or data distributions— which fundamentally shape what patterns models learn and how they fail when encountering out-of-distribution (OOD) data. This work addresses five key research questions that illuminate different aspects of inductive bias:

- **(1)** Do CNNs and Vision Transformers rely on different visual cues, with CNNs showing texture bias and ViTs favoring shape?
- **(2)** How do architectural assumptions like convolutional locality versus transformer global context affect properties like translation invariance?
- **(3)** What representational and generative differences emerge between VAEs and GANs due to their distinct training objectives?
- **(4)** How do multimodal contrastive models like CLIP develop semantic biases compared to standard vision models?
- **(5)** Which inductive biases most improve out-of-distribution generalization?

Our investigation spans discriminative models (CNNs, ViTs), generative models (VAEs, GANs), and multimodal models (CLIP). We demonstrate that CLIP's training on diverse image-caption pairs instills shape-oriented, conceptual biases that enable remarkable zero-shot generalization and cross-domain robustness—capabilities that pure vision models struggle to achieve without massive datasets or specialized training procedures.

2. Methodology

We conducted a systematic analysis across five model families, examining their inductive biases through carefully designed experiments. Our methodology addresses each research question through targeted evaluations of generalization capabilities, representational structure, and robustness properties.

2.1. Shape vs. Texture Bias in Discriminative Models

We utilized pre-trained ResNet 50 and ViT-B/16 for all our experiments. The models were fine-tuned using the STL10 dataset as the baseline.

2.1.1. Datasets for Style Transfer

In order to assess shape and texture biases, we conducted a cue-conflict experiment on the pre-trained models using the Stylized-ImageNet (SIN) dataset. We induce a covariate (domain) shift that highlights the learned feature biases of each model. Figure 14 show example images the models were shown in this experiment. We presented images with contradictory features for object classification, with the aim being the feature each model used for correct classification.

To minimize the effects of concept shift, we sampled only

classes from an SIN subset that contained the STL10 classes. All our code is available in the GitHub repository. The referred SIN dataset can be retrieved from: https://github.com/rgeirhos/Stylized-ImageNet.

2.1.2. Color Bias Test

The consensus in the literature is that CNNs focus on "local" features and ViTs consider the "global" structure of an image (Geirhos et al. , 2018). We prepared a grayscale version of our baseline dataset by converting the STL10 images to grayscale using `torchvision.transforms.Grayscale`. The primary aim was to see which model is more robust to loss of a localized cue.

2.2. Architectural Bias Analysis of CNNs and ViTs

2.2.1. Locality Bias Experiments

For the purposes of this investigation, we conducted three main experiments along with a control baseline. Fine-tuned ResNet 50 and ViT-B/16 on STL10 were our supervised baseline. The three experiments (samples visualized in Figure 1) included:

- **Translations** Images in our baseline dataset were applied affine transformations to shift the object in the horizontal and vertical directions.
- **Occlusions** Images in our baseline dataset were masked by black patches. The masking ratio was variable.
- **Shuffling** Images in our baseline dataset were divided into patches of fixed sizes; we shuffled the patches randomly.

Each of these shifted domains introduced distributional changes relative to our baseline. We visualized samples and evaluated the models on such covariate shifts.

2.3. Domain Adaptation in Discriminative Models

2.3.1. PACS for Domain Adaptation

To test how well these models generalize to data distributions they have not seen in training, we employ the Photo, Art Painting, Cartoon, and Sketch (PACS) dataset from Hugging Face. The models are fine-tuned on a source domain (consisting of 3 subdomains, namely: Photo, Art, and Cartoon) and tested for inference robustness on the target domain (Sketch).

2.4. Generative Model Bias Comparison

In the generative family of deep representation learning models, we focus our attention on **Variational Autoencoders** (VAEs) and **Generative Adversarial Networks** (GANs). These models have vastly different training objectives and architectural designs, which lead to distinct inductive biases affecting their performance on various tasks.

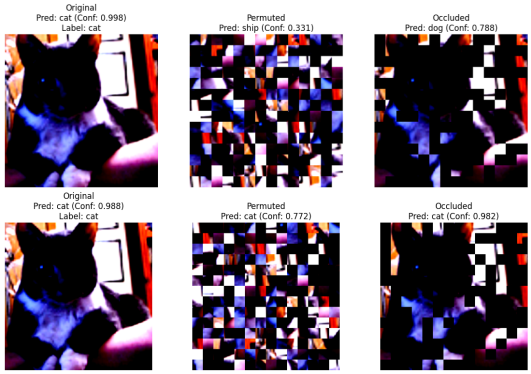


Figure 1. Visualization of results for patch occluded and shuffled images from STL10. Top Row: ResNet 50 predictions on the original, occluded, and shuffled dataset. Bottom Row: Vit-B/16 predictions on the original, occluded, and shuffled dataset respectively.

2.4.1. Datasets, Training Details

We used the **CIFAR-10** dataset for training both VAEs and GANs. CIFAR-10 consists of 60,000 32×32 color images in 10 classes, with 6,000 images per class. We combined both the training and test sets to form a single dataset of 60,000 images for training purposes. Additionally, we evaluated the models on out-of-distribution (OOD) samples collected from random images on the internet to assess their generalization capabilities.

For VAEs, the pixel values were scaled to the range [0, 1] and for GANs, the pixel values were scaled to [-1, 1].

Both models were trained for 50 epochs over the entire dataset with batch size of 128.

VAE was trained using Adam Optimizer with a learning rate of 0.001. The loss function was a combination of reconstruction loss and KL-divergence loss. Models with annealing started with a KL-weight of 0 and linearly increased to 1. Models without annealing used a fixed KL-weight of 0.000075.

GAN was trained using Adam Optimizer with a learning rate of 0.0002 and $\beta_1 = 0.5$ & $\beta_2 = 0.999$. The loss function was the generic adversarial loss except that the generator used the non-saturating loss by minimizing $-\log(D(G(z)))$. Additionally, real labels were smoothed to 0.9 instead of 1 to help in stabilizing training.

2.4.2. Model Architectures

The VAE used 3 convolutional layers to downsample $32 \times 32 \times 3$ dimensional input images into latent mean and logvar vectors of dimension 64, 128 and 256. For each of those, 2 versions were trained: one with KL annealing (gradually increasing the KL-divergence weight) and one without (fixed β-weighting). The decoder mirrored the encoder with 3 `ConvTranspose2d` layers followed by a `Sigmoid` activation to output reconstructed images.

The GAN was based on the DCGAN architecture with a generator consisting of 5 `ConvTranspose2d` layers interleaved with ReLU activation and `BatchNorm` to upsample a latent vector of dimension 64, 128 or 256 into $64 \times 64 \times 3$ images. The discriminator used 5 `Conv2d` layers with LeakyReLU activation and `BatchNorm` to downsample images into a single real/fake prediction via `Sigmoid` activation.

2.4.3. Experiments and Evaluation Metrics

We conducted 3 experiments to probe the inductive biases of VAEs and GANs. Both qualitative and quantitative metrics were employed. Additionally, we monitored the training dynamics of both models to asses convergence.

1. **Reconstruction vs Generation**: we qualitatively compared VAE reconstructions and generations with GAN generations. Quantitatively, we computed FID scores for both models to assess sample quality. To quantify GAN's mode coverage, we computed the Inception Score (IS) for its generated samples.
2. **Latent Space Structure**: we visualized the latent space of all variants of VAE trained using t-SNE plots. We also visualized samples from latent space of both VAE and GANs to assess latent space smoothness and semantic meaning in latent dimensions.
3. **Out-of-Distribution Inputs**: we evaluated VAE on OOD samples from the internet, qualitatively assessing reconstruction quality and quantitatively measuring reconstruction error (MSE). For GANs, we performed an analagous stress-test by feeding OOD latent vectors (away from the standard Gaussian it expects) and qualitatively assessing generation quality.
4. **Training Stability**: we monitored loss curves for both models to assess convergence and stability. For GANs, we also observed for mode collapse by tracking diversity of generated samples over training epochs.

Note: Unless otherwise stated, experiments used the 256-dimensional models. For VAEs, we used the no-anneal variant when reconstructions were needed (to prioritize fidelity) and the annealed variant when sampling (to ensure a smoother, well-structured latent space). Alternate choices did not contribute meaningfully to comparison with GANs.

2.5. Contrastive Multimodal Model Analysis (CLIP)

We focus our analysis on OpenAI's CLIP ViT-B/32, a contrastively trained vision-language model that learns joint representations from 400 million image-text pairs. Unlike traditional supervised models trained on single-label classification, CLIP learns to align images with their natural language descriptions, potentially developing more semantic and human-aligned biases by aligning visual and semantic concepts across modalities (Radford et al., 2021).

2.5.1. Datasets and Experimental Setup

We evaluate CLIP using a pre-trained ViT-B/32 model without any fine-tuning, maintaining its zero-shot capabilities. Our evaluation encompasses multiple datasets: STL-10 (8,000 test images) for zero-shot classification, a curated 500-image retrieval set (50 images per class), cue-conflict stimuli from the texture-vs-shape repository (180 stylized images), and a corruption suite (200 baseline samples with various perturbations for robustness testing). For supervised model comparison, we use ResNet-50 baselines finetuned on STL-10.

2.5.2. Zero-shot Classification Protocol

For each candidate class, we evaluate through the following prompt strategies:

- **Generic:** class name only, e.g. "`airplane, ...`"
- **Descriptive:** "`a photo of a {class}`"
- **Detailed:** "`an amazing, colourful clean textured and shape appropiate photo of a {class}, {class}.`"

CLIP computes image embeddings and selects the class whose text embedding achieves highest cosine similarity in the shared embedding space. We report top-1 accuracy across all evaluations.

2.5.3. Image-Text Retrieval Evaluation

We measure bidirectional retrieval accuracy (Text→Image and Image→Text) across three prompt complexity levels similar to zero-shot classification. This evaluates CLIP's multimodal alignment capabilities.

2.5.4. Domain-Shifted Classification

We evaluate CLIP's domain generalization on a subset of the PACS dataset obtained from the Hugging Face repository (`flwrlabs/pacs`). For this experiment we filter PACS

to seven object classes that appear across all four PACS domains: `dog`, `elephant`, `giraffe`, `guitar`, `horse`, `house`, and `person`. PACS contains four distinct visual domains—`photo`, `art_painting`, `cartoon`, and `sketch`—which together enable controlled tests of cross-style generalization. Evaluation was performed in a zero-shot manner on each domain using 7 different prompting scenarios:

- **generic:** "a {class}".
- **domain:** "a domain of a {class}".
- **artwork:** "artwork of a {class}".
- **painting:** "a painting of a {class}".
- **cartoon:** "a cartoon of a {class}".
- **sketch:** "a sketch of a {class}".
- **drawing:** "a drawing of a {class}".

2.5.5. Representation and Bias Analysis

We extract CLIP image embeddings for mixed in-domain and OOD samples, visualizing embedding structure through t-SNE. For shape vs. texture bias measurement, we use cue-conflict stimuli with dual prompts describing images by shape ("a photo of a {shape}") versus texture ("a photo with the texture of a {texture}").

2.5.6. Robustness Testing

We evaluate performance drops from clean baselines across corruption types including Gaussian noise, blur, and contrast perturbations at multiple severity levels.

3. Results

3.1. Shape vs. Texture Bias in CNNs and Vision Transformers

We established a supervised baseline for our fine-tuned models (refer to table 2) on STL10. Inference on the SIN subset for STL10 yielded an accuracy of 74%. Compared to the CNN model (approximately 50%), we observe that ViTs generalize and adapt better to new domains.

	ResNet 50 (%)	ViT-B/16 (%)
Shape	46.7	48.1
Texture	53.3	51.9

Table 1. Shape and Texture Biases of ResNet 50 and ViT-B/16 on SIN subset corresponding to STL10

We also observed at inference that CNNs perform worse than ViTs when the color information (cue) is removed from an image (Table 2). Figure 2 shows an example where the ResNet 50 failed, and the ViT-B/16 succeeded in correctly classifying the class of the object in our grayscale dataset. This suggests that CNNs rely on local cues like color and texture; may suggest a correlation with their model architecture.

Figure 2. A sample of the grayscale dataset passed through both the models. Left: ResNet 50 classified the grayscale version of the deer as a cat. Right: ViT-B/16 classified that same image with the correct label - deer

3.2. Architectural Biases: Translation Invariance and Context Sensitivity

Table 2 shows the performance of the models at inference time when exposed to the prepared datasets: translated images (with vertical and horizontal shifts), patch occluded images, and shuffled images. Figure 1 shows an example from our experiments. Contrary to literature (Kaiser , 2023), ResNet 50 performed worse than ViT-B/16 when we tested the models for translation invariance (a marginal difference of 5%, but it exists). This may be a result of our training regimen, as during fine-tuning both of these models were exposed to images with random transformations.

	ResNet 50 (%)	ViT-B/16 (%)
Baseline	95.6	98.0
Stylized	51.1	73.9
Grayscale	90.1	96.2
Translated	89.9	94.7
Occluded (ratio = 0.5)	61.8	94.0
Shuffled	18.0	50.3
PACS (Source Domains)	97.0	91.4
PACS (Target Domain)	54.6	36.8

Table 2. Inference Accuracies for ResNet 50 and ViT-B/16 on datasets that highlight their inductive biases

However, on the other two datasets, the ViTs perform much

better by a huge margin (a difference of more than 30% in classification accuracy). For our patch occlusion experiment, we increased the masking ratio in steps from zero to 0.75 and observed that ResNet's performance deteriorated rapidly from 95% to 33% (almost one-third of its original accuracy) whereas the ViT-B/16 showed a smooth degradation in performance from 97% to 80% classification accuracy. On the shuffled patches dataset, ViT-B/16 gave a test accuracy of 50.3% compared to the marginally lower 18% accuracy of ResNet 50.

3.3. PACS for Domain Adaptation

Table 2 highlights that current models fail to generalize well when exposed to domains not seen during training. The accuracy of ResNet 50 drops from 96.95% to 54.57% compared to ViT-B/16 that experienced a larger drop in performance from 91.42% to 36.78%. We further discuss these results in Section 4.3.

3.4. Generative Model Biases: VAE vs. GAN Trade-offs

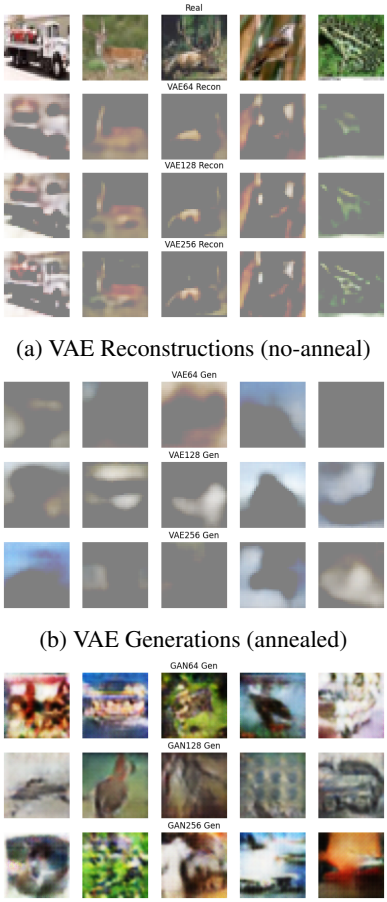
3.4.1. Reconstruction vs Generation Quality

The two VAE variants excelled at expectedly different tasks: the annealed version produced plausible random-sample latent generations although very blurry, and the non-annealed version produced faithful reconstructions of input images, as shown in Figure 3a and 3b. Attempts at random-sample latent generations using the non-annealed version produced unrecognizable noisy outputs, and reconstructions using the annealed VAEs were much blurrier compared to the non-annealed variant's reconstructions (these auxiliary visualizations can be found in the appendix).

GANs, on the other hand, produced sharper and apparently more realistic random-sample generations as shown in Figure 3c. However, these outputs suffered from artifacts such as checkerboard patterns and occasional mode collapse.

Overall, the quality of generations and reconstructions improved with increasing latent dimensionality for both models, with especially pronounced results for VAE reconstructions.

Quantitatively, the FID scores in Table 3 reflect these qualitative observations: high FID scores for VAEs indicate that they generate blurrier outputs (lower fidelity) and low FID scores for GANs because of their sharper images. The Inception Scores for GANs also indicate reasonable diversity and quality, with GAN256 achieving the best score of 5.3628. The MSE reconstruction errors for VAEs were low, confirming their strength in faithful reconstructions which GANs are unable to perform.



(a) VAE Reconstructions (no-anneal)

(b) VAE Generations (annealed)

(c) GAN Generations

Figure 3. Comparison of reconstructions and generations from VAE and GAN models. VAEs reconstruct inputs well but generate blurry samples, whereas GANs produce sharper images.

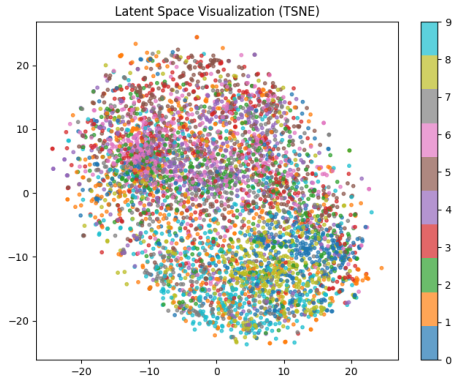
3.4.2. Latent Space Structure

The reconstructions and generations are explained by the latent space of the two models. The non-annealed VAE's t-SNE latent space visualization as show in Figure 4a shows class clustering and separation, indicating a well-structured latent space. However, the latent space is not as smooth or continuous, leading to noisy generations. The annealed VAE's t-SNE plot in Figure 4b shows a strong tendency toward standard Gaussian structure which allows for smooth interpolations and meaningful random-sample latent generations, but there is significant class overlap and not as much clustering so the reconstruction quality suffers.

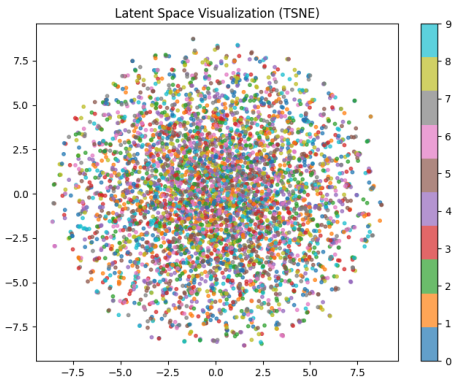
Since GANs do not have an explicit encoder, we cannot visualize their latent space directly. However, we can draw a comparison between VAE and GAN latent dimensions by performing linear interpolations in latent space. The VAE256 interpolations (both annealed VAE and non-annealed VAE) in Figure 5 show smooth transitions between

Table 3. FID, MSE Reconstruction Error and Inception Scores for VAE and GAN models. Lower FID/MSE and higher IS indicate better quality.

Model	FID Score	Inception Score	MSELoss
VAE64 (NA)	168.200	-	0.0378
VAE128 (NA)	182.038	-	0.0366
VAE256 (NA)	215.454	-	0.0359
VAE64 (A)	192.485	-	0.0438
VAE128 (A)	192.208	-	0.0438
VAE256 (A)	197.433	-	0.0438
GAN64	106.410	4.4992	-
GAN128	189.310	3.9236	-
GAN256	72.489	5.3628	-



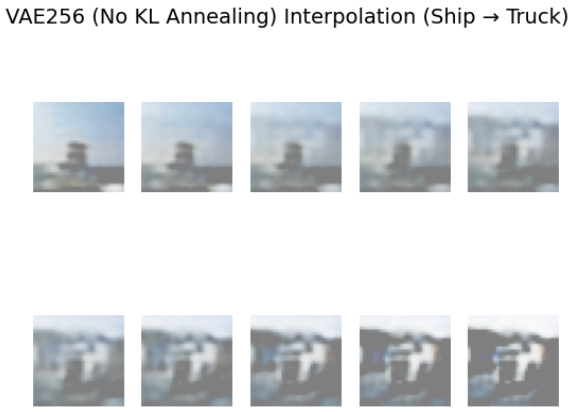
(a) t-SNE of VAE256 latent space (no-anneal)



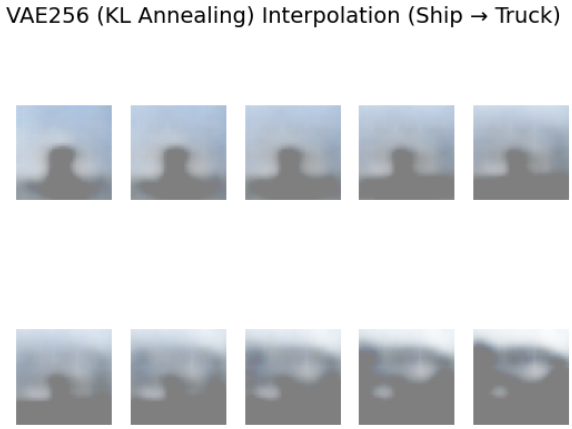
(b) t-SNE of VAE256 latent space (annealed)

Figure 4. t-SNE visualizations of VAE latent spaces. Each color represents a different CIFAR-10 class.

images which reflects their latent space structure as seen in the t-SNE plots (while maintaining respective reconstruction/generation characteristics as discussed earlier). The GAN256 interpolations in Figure 6 show sharper transitions



(a) VAE256 Interpolation (no-anneal)



(b) VAE256 Interpolation (annealed)

Figure 5. Latent interpolations for VAE256 models. Smooth transitions indicate well-structured latent spaces.

and occasional mode collapse, indicating a less structured latent space.

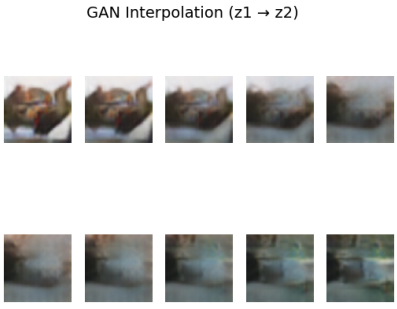
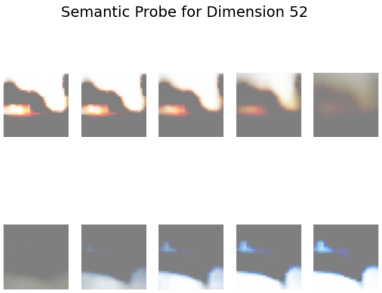


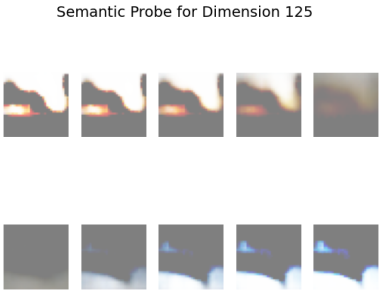
Figure 6. Latent space interpolations for GAN256. Sharp transitions and occasional mode collapse indicate a less structured latent space.

To probe for semantic meaning in the latent space of VAEs, 3 random dimensions of VAE256 latent were sweeped from

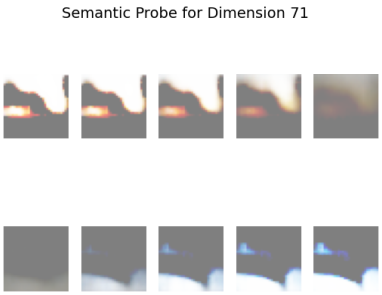
-5 to 5. Results are shown in Figure 7. All 3 interpolations appear almost the same, transitioning from orange hues to blue hues. However, the shape, orientation, etc. remain the same.



(a) Dimension 52 sweep



(b) Dimension 125 sweep



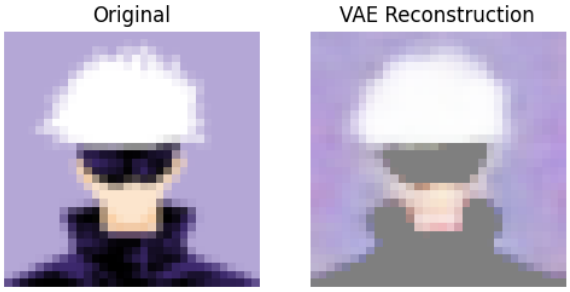
(c) Dimension 71 sweep

Figure 7. Selected semantic traversals for VAE256 (annealed). Sweeping individual latent dimensions produces color shifts but no clear semantic changes.

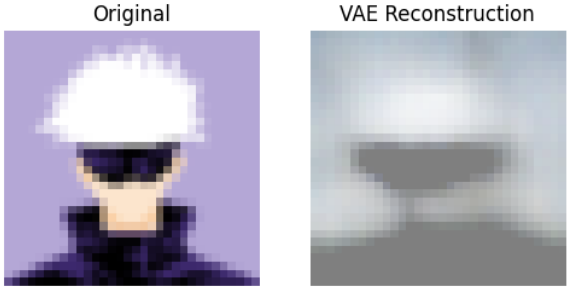
3.4.3. Out-of-Distribution Inputs

The non-annealed VAE is able to reconstruct OOD images from the internet to a great extent (see Figure 8a), although reconstructions are blurry as is characteristic of VAEs. The same image reconstructed by the annealed VAE (Figure 8b) is even blurrier and less faithful to the original image; rather it looks like the model tried to pull the image towards the CIFAR-10 distribution.

For GANs, OOD inputs were emulated by feeding latent vectors not sampled from the standard Gaussian that the generator was trained on. Three types of OOD latent vectors were tested: scaled (latent vectors multiplied by 5), sparse (randomly chosen dimension populated with Gaussian random variable, rest zeroes), and uniform (latent vectors sampled from a uniform distribution spanning $[-10, 10]$ instead of Gaussian). The results (Figure 9) show that higher latent dimensional GAN performed poorest, producing mostly noise. The lower dimensional GANs produced slightly more structured outputs, but these had frequent artifacting and did not resemble natural images.



(a) VAE256 OOD Reconstructions (no-anneal)



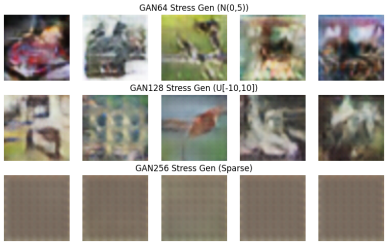
(b) VAE256 OOD Reconstructions (annealed)

Figure 8. Examples of VAE reconstructions on out-of-distribution inputs. Both annealed and no-anneal variants fail to generalize cleanly.

The VAEs performance on OOD data can be used for anomaly detection. Figure 10 shows the reconstruction error (MSE) distribution for in-distribution CIFAR-10 test samples versus OOD internet samples using VAE256 (no-anneal). Although there is significant overlap in the reconstruction error of in-distribution and out-of-distribution reconstructions (due to the reconstructive nature of a non-annealed β-VAE with small KL-weight), the OOD samples still yield a significantly higher MSE on average, allowing anomaly detection thresholds to be set against reconstruction error.

3.4.4. Training Stability

The plots in Figure 11 show the training loss curves for VAE256 (with and without annealing) and GAN256. As expected, the VAAEs showed stable, predictable convergence



(a) Scaled latent input

(b) Sparse latent input

(c) Uniform latent input

Figure 9. GAN256 generations with out-of-distribution latent vectors.

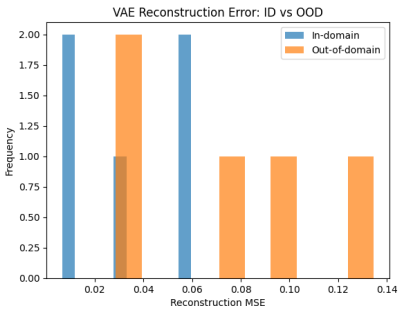
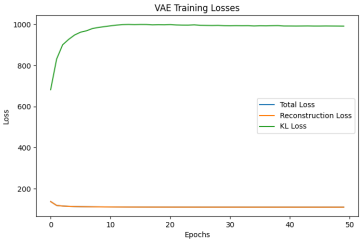


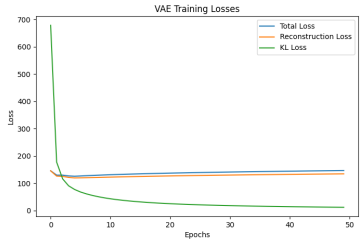
Figure 10. Anomaly detection using reconstruction error (MSE) for VAE256 (no-anneal). OOD samples yield significantly higher errors than in-distribution CIFAR-10.

of their loss functions. The annealed version had a large KL-weight, so its KL-Loss converged fast which caused total loss to start increasing. This is why its reconstruction quality suffers. The non-annealed VAE had a very small KL-weight, so the KL-Loss is not punished as much by the optimizer and it increases, but the total loss decreases as reconstruction error is prioritized.

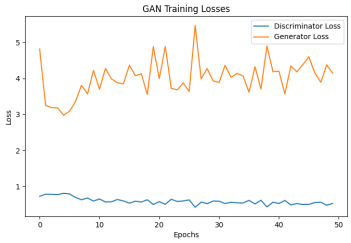
The GAN had a typically unstable loss progression due to adversarial nature of training. However, with techniques like non-saturating loss and label smoothing, we were able to avoid vanishing gradients and mode collapse to a large extent.



(a) VAE256 Loss (no-anneal)



(b) VAE256 Loss (annealed)



(c) GAN256 Loss

Figure 11. Training loss curves for VAE and GAN models. VAE converges stably, whereas GAN shows oscillatory behavior.

3.5. CLIP: Multimodal Contrastive Biases and Semantic Understanding

3.5.1. Zero-shot Classification Performance

CLIP demonstrates remarkable zero-shot capabilities on STL-10, beating the supervised baseline accuracy of **95.55 percent** without any task-specific training. Our **Descriptive** prompt strategy yields the best overall accuracy as shown in Table 4.

	Plain (%)	Descriptive (%)	Detailed (%)
Overall	95.83	96.79	96.43

Table 4. Zero-shot top-1 accuracy on STL-10 by prompt strategy

Class	Plain (%)	Descriptive (%)	DP mistakes
airplane	93.4	97.5	ship (16)
bird	99.5	99.5	monkey (2)
car	95.2	96.6	truck (26)
cat	87.8	88.2	deer (56)
deer	97.0	97.5	horse (12)
dog	96.9	96.0	cat (15)
horse	98.0	98.2	deer (7)
monkey	91.8	95.4	bird (17)
ship	99.9	99.9	horse (1)
truck	98.9	99.0	car (8)

Table 5. Per-class zero-shot accuracy and most common confusion for Descriptive prompts (DP)

Notable patterns from Table 5

- **Cat remains the weakest class** (Plain 87.8%, Descriptive 88.2%) and is most often confused with **deer** under Descriptive prompts.
- **Vehicle/transport confusions** persist (car→truck, airplane→ship) even under the Descriptive prompt strategy, indicating coarse-grained visual overlap.
- **Prompt effects are class-dependent:** some classes (airplane, truck) improve substantially with Descriptive/Detailed prompts, while others (car) can degrade under Detailed prompts.

3.5.2. Image–Text Retrieval Capabilities

We evaluate retrieval on a curated 500-image dataset (50 images per class from the 10 STL-10 classes used here). Three prompt complexity levels were tested: Basic, Descriptive, and Detailed. Retrieval Analysis in CLIP achieves near perfect accuracy in all prompting strategies showing the importance of shared latent space in multi-modal models, except in detailed where extra details might have occluded the image embeddings to get seperated from its text embedding.

Detailed prompts used:

```
airplane: "a commercial airplane flying
in the sky"
bird: "a colorful bird perched on a
branch"
car: "a red car driving on the road"
cat: "a cute cat sitting indoors"
deer: "a deer grazing in a forest"
dog: "a friendly dog playing outside"
horse: "a brown horse galloping in a
field"
monkey: "a monkey swinging in trees"
ship: "a large ship sailing on the ocean"
truck: "a heavy truck carrying cargo"
```

Query type	Text → Image (%)	Image → Text (%)
Basic	100.0	100.0
Descriptive	100.0	100.0
Detailed	100.0	90.0

Table 6. Retrieval Accuracy

3.5.3. Domain-shifted classification

CLIP's ability to generalize through different domains can be explicitly observed from its near perfect accuracy throughout the four different domains and the seven different prompting strategies as shown in Table 7. The individual accuracies for each domain on the respective 7 domain prompts do not show a major shift in accuracies, indicating that prompting strategy can only effect to an extent.

Domain	Optimal prompt (highest accuracy)
photo	generic (100.0%) — all templates tie
art_painting	photo (96.0%)
cartoon	artwork (97.5%)
sketch	artwork (88.5%)

Table 7. Prompt type that yields highest accuracy per PACS domain.

Compared to ResNet's average accuracy of **54.6%** (Table 2), CLIP's **95.5 percent** highlights the advantages CLIP receives from being a multimodal attention based model compared to ResNet's CNN based backbone in domain-shift classification, and its ability to focus on shape more than texture. However, Figure 12 shows one of the many scenarios where CLIP underperforms due to the lack of texture in sketch domain.

3.5.4. Representation Analysis: Semantic Clustering

CLIP's learned representations exhibit strong semantic clustering across visual domains. T-SNE visualization reveals that embeddings group primarily by object category rather than visual style, with photos, sketches, and cartoons of identical objects clustering together in the feature space (Figure 13).

This semantic organization contrasts with supervised models that typically cluster images by dataset or visual appearance, indicating CLIP's bias toward conceptual rather than superficial feature grouping.

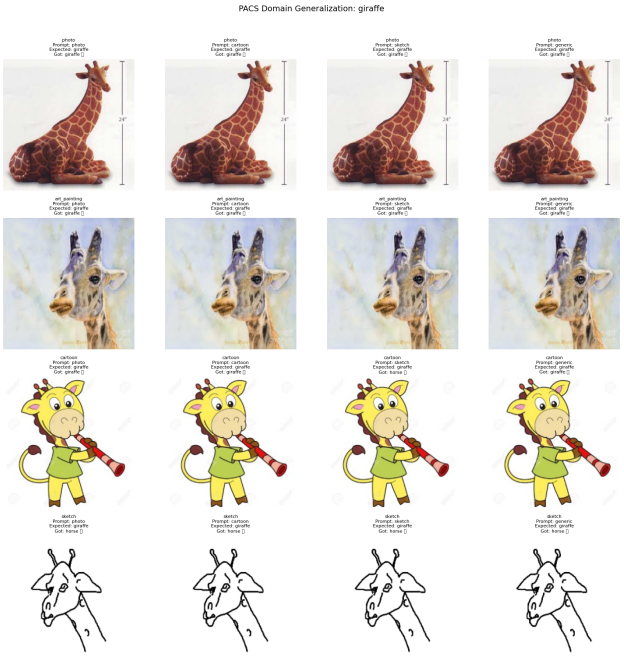


Figure 12. a `giraffe` sketch misclassified as `horse` by CLIP. Sketch domain yields the lowest accuracies for several animal classes.

3.5.5. Shape vs. Texture Bias Quantification

Cue-conflict experiments reveal CLIP's preference for shape-based recognition over texture cues. When presented with images containing conflicting shape and texture information, CLIP chooses shape-consistent labels (Table 8), as compared to ResNet50 which had a greater texture bias due to its behaviour to favour local relationships.

Decision type	CLIP	ResNet50
Shape	60.6%	46.7%
Texture	39.4%	53.3%

Table 8. CLIP vs finetuned ResNet50 accuracy bias in cue-conflict scenarios.

This shape preference supports our hypothesis that multimodal contrastive training encourages semantic, object-centric feature learning over texture-dominated classification strategies.

3.5.6. Robustness to Visual Corruptions

CLIP exhibits differential robustness across corruption types, showing strong resilience to blur but significant sensitivity to additive noise. This pattern suggests robust encoding of high-level semantic structure alongside vulnerability to low-level perturbations.

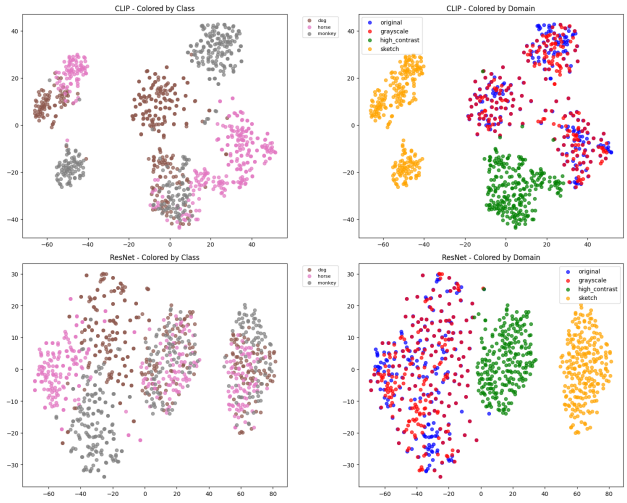


Figure 13. Semantic comparison of embedding space

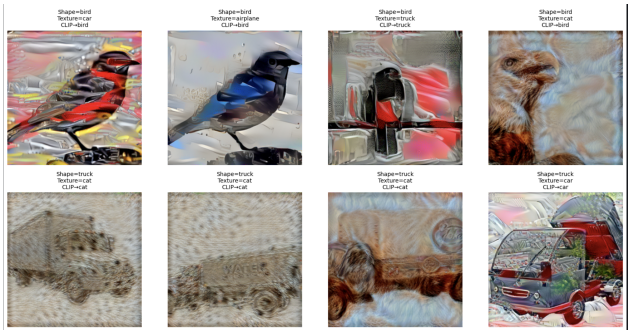


Figure 14. Cue Conflict over bird and truck

The substantial noise sensitivity contrasts sharply with minimal blur impact (Table 9), indicating that CLIP's semantic focus provides natural invariance to some distortions while remaining vulnerable to others.

4. Discussion

4.1. Shape vs. Texture Biases Across Architectures

In line with empirical findings from other literature (Geirhos et al. , 2018), we observe that the ResNet 50 (pre-trained on ImageNet) was more biased towards recognizing textures instead of shapes (Table 1). However, our observations for ViT-B/16 contradict the consensus (Kaiser , 2023) that ViTs are significantly less texture biased than shape biased. The difference is marginal (3.8%) but our ViT-B/16 was more biased towards texture than shape. Figures 14 and **??** display the visualized samples used at inference time. Our results highlight a discrepancy between the common assumption that ViTs are more shape biased than texture biased. A possible reason for this may be the small sample size of our inference dataset. Given we had access to more

Corruption (severity)	Accuracy (%)	Drop (pp)
Clean (baseline)	98.0	0.0
Gaussian noise (light)	74.0	24.0
Gaussian noise (medium)	73.5	24.5
Gaussian blur (light)	97.5	0.5
Gaussian blur (medium)	97.5	0.5
Gaussian blur (heavy)	92.5	5.5
Average noise impact	–	24.3 pp
Average blur impact	–	2.2 pp

Table 9. Robustness to visual corruptions (accuracy and drop from baseline).

stylized images, it is likely we may get results similar to the consensus in literature.

We also observe that the ViT performed better than ResNet 50 on the grayscale dataset. Particularly with regards to convolutional filters capturing information via localized receptive fields. This supports our findings that local color cues impact CNN architecture performance on classification related tasks. On the other hand, ViTs with their attention mechanism between image patches are better able to learn the global context in an image during training; the loss of local cues like color have little impact on their performance.

4.2. Architectural Biases and Their Consequences

Our results correlate with known literature (**?**) - ViTs usually generalize better to new domains since the attention mechanism can correlate distant but semantically relevant cues in different regions of an image that may provide label information. The global receptive field enables ViTs to learn better feature representation; maximizing the mutual information between the input and the embedding space.

On the other hand, the convolutional architecture of ResNet 50 is responsible for the poor generalization across different domains. By design, CNNs focus on local cues like color, texture, edges, etc. that may cause the model to learn spurious correlations between the feature space and label space. Since they tend to focus on local cues only, they are more robust to translations than ViTs that rely on a global image structure for classification. However, unlike ViTs, they suffer from severe performance degradation when exposed to masked images or shuffled patches. This is mainly because these domains remove information regarding local cues (e.g. correlations between color and edges within a patch). The localized receptive fields fail to map useful representations of the given image to the label. Instances of such domain shifts are further discussed in domain generalization using PACS.

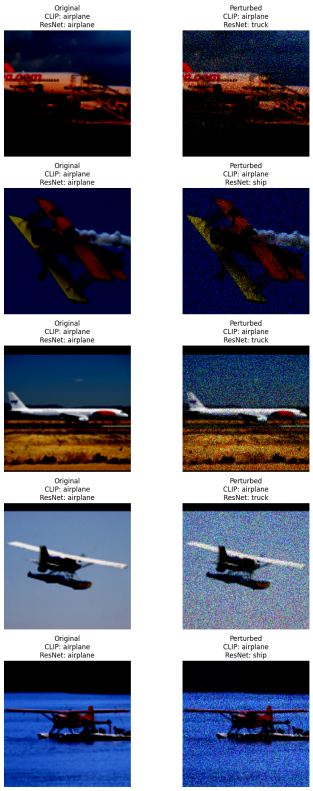


Figure 15. Gaussian noise heavy pertubration

4.3. Domain Adaptation

This multiple domain single task classification problem on the PACS dataset enabled us to explore the problem of domain shifts and domain adaptation. The PACS dataset introduced a covariate shift that changed the distributions across domains, but kept the relationship between input and labels relatively the same.

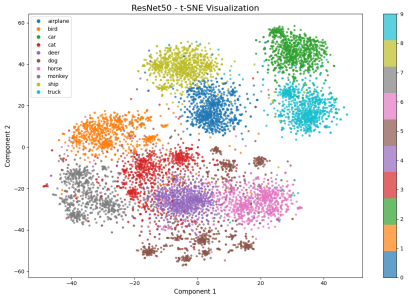


Figure 16. Feature Space t-SNE Visualization for ResNet 50

We visualize the feature representations for both of the models in figures 16 and 17. As observed, the class points for ViT-B/16 form distinct clusters with minimal overlap, whereas there is a continuous spread of class points and significant overlap in labels for ResNet 50 t-SNE map. The

entangled feature space of ResNet 50 suggests that the features are domain dependent and different classes share local features. Compare this to ViT-B/16's t-SNE map and we observe that the learned features are independent of one another, and do not contain spurious correlations (unlike ResNet 50) between the features and labels. The global context in ViTs captures semantically coherent information and domain independent features, which naturally enables it to perform better on generaliztion tasks.

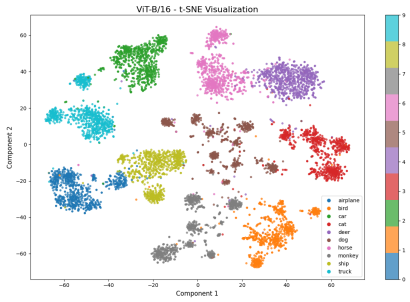


Figure 17. Feature Space t-SNE Visualization for ViT-B/16

However, our results on the PACS dataset showed that ViTs performed poorly compared to CNNs. A reason for this could be improved shape biasness in ResNet 50. We fine-tuned the ResNet model on three separate domains with distinct textures, but the same classes. As discussed in (Geirhos et al. , 2018), we essentially made our model more robust to inductive biases induced by datasets as it learned to correctly map the object shape to the label. The ViTs may have performed poorly mainly due to the small size of our dataset (under 4000 images). This may have caused it to learn spurious correlations in the global structure. But this specific result does go against literature (**?**) that ViTs generalize better.

4.4. Generative Model Biases in Latent Structure

VAEs and GANs exhibit distinct inductive biases in their functionality as generative models due to the difference in how they learn to represent the underlying data distribution. These differences impose specific constraints on the latent spaces of each model, which are reflected in their performance on reconstruction, generation, and out-of-distribution tasks, thereby resulting the observed biases.

4.4.1. Reconstructions vs. Generations

VAEs have an explicit encoder and a well-defined loss objective. These two facts yield VAEs with a latent space that is structured and, to some extent, interpretable. VAEs try to minimize reconstruction error and are regularized by a KL-divergence term that pulls the latent space distribution towards a standard Gaussian prior in order to enable generative sampling. This dual objective places competing pressures on the latent space resulting in a diverse coverage of data distribution at the cost of output fidelity (blurred images).

On the other hand, GANs lack an explicit encoder and are trained adversarially. The sole objective of the generator is to "fool" the discriminator. For this, the model empirically learns to generate outputs that are sharp and high-fidelity, but it may end up relying on the few outputs that successfully fool the discriminator, leading to mode collapse and less diversity in its generations. This results in an unstructured latent space that represents the GANs attempt to reach Nash equilibrium with the discriminator. This equilibrium does not guarantee full data coverage because there is no part of the objective function that explicitly encourages this in the GAN's latent space.

An architecturally specific observation about GANs that illustrates this behaviour is that some outputs displayed "checkerboard" artifacts (repeating patterns within an image). This could be because DCGAN employs convolutional layers that inherently exploit local spatial correlations, and the generator may have learned to exploit these correlations to produce high-frequency patterns that can fool the discriminator, but do not correspond to natural images.

4.4.2. Latent Space

Visualizing the VAE latent space with t-SNE and interpolating between latent vectors reveals the smoothness and structure that is embedded in VAEs learned representations. This structure may also carry semantic meaning in different latent dimensions, as the dimensional sweep plots show a shift frmo orange to blue hues (color semantic), but it is difficult to ascribe much interpretability to the latent because the different dimensions seem to be entangled (hence producing very similar outputs when different dimensions are swept).

An attempt to probe the GAN latent space was by interpolating between two different latent vectors. The results confirmed that GANs lack a structured latent because the transitions were sharp and very similar across 4-5 steps. Moreover, the intermediate images often exhibited mode collapse, producing similar outputs for different latent vectors. This implies that the GAN latent space is characterized by discontinuities and regions of high density separated by areas that do not correspond to realistic images.

4.5. CLIP's Semantic Biases and Multimodal Learning

The consistent high performance across prompt strategies (Table 4) demonstrates CLIP's robust semantic understanding. The systematic misclassification patterns—particularly cat→deer confusions and car→truck errors (Table 5)—reveal CLIP's embedding space organization. These confusions occur between semantically or visually similar cate-

gories, indicating CLIP groups concepts by shared features rather than superficial texture patterns that might distinguish, for example, fur from hide textures, as also seen from the embedding space representation in Figure 13.

4.5.1. Retrieval Performance and Prompt Complexity Effects

The retrieval results illuminate important aspects of CLIP's multimodal alignment. Text→Image retrieval achieves perfect performance (Table 6) across all query complexity levels, demonstrating CLIP's robust ability to select appropriate images given textual descriptions. However, Image→Text retrieval reveals a complexity-dependent trade-off: while basic and descriptive prompts maintain perfect accuracy detailed prompts show degraded performance.

This asymmetric pattern emerges because detailed textual descriptions include specific attributes (actions, colors, contexts) that may not be visually present in every image instance. When an image matches the general class but lacks described attributes, it receives lower similarity scores to detailed text descriptions. For instance, a stationary horse image struggles to match "a brown horse galloping in a field" despite correct class identity. Conversely, Text→Image retrieval succeeds because CLIP can identify images containing the described attributes from a diverse set. This finding has practical implications: for applications involving high image variability, shorter descriptive queries prove more robust for Image→Text retrieval, while detailed prompts benefit scenarios where images consistently contain the specified attributes.

4.5.2. Cross-Domain Adaptation

CLIP's cross-domain performance on PACS reveals a sophisticated understanding of visual style and domain characteristics (Table 7).

Several key factors drive these domain-dependent performance patterns. First, CLIP's pretraining data predominantly consists of natural images with descriptive captions, creating natural alignment with photographic content. Style-dominated domains require domain adaptation beyond typical caption-image co-occurrence patterns. Second, prompt engineering effects reveal CLIP's contextual sophistication: artwork-specific prompts ("artwork of a...") optimize performance on stylized domains by steering text embeddings toward appropriate stylistic representations. Interestingly, "photo" prompts sometimes improve artistic domain performance by emphasizing canonical object shapes over stylistic elements, particularly beneficial when paintings preserve clear object structure.

The sketch domain presents the greatest challenge. Sketches strip away texture and color cues, requiring classification based purely on shape and structural details. When sketches omit distinguishing features—such as giraffe spots or elongated necks—object boundaries and proportional cues alone prove insufficient, leading to systematic giraffe→horse confusions as shown in Figure 12. The artwork prompt partially addresses this limitation by orienting linguistic prototypes toward stylized rather than photorealistic representations, though fundamental challenges remain for highly abstract line drawings.

The model's semantic focus provides substantial advantages over texture-dependent approaches, but domain-specific adaptation through prompt engineering or targeted fine-tuning may be necessary for optimal performance on highly stylized content.

4.6. Inductive Bias and Out-of-Distribution Generalization

4.6.1. VAE vs GAN

The inductive biases embedded in VAE and GAN architectures are especially apparent when they are exposed to out-of-distribution (OOD) inputs. The non-annealed VAE demonstrates a strong reconstructive bias, allowing it to produce reasonable approximations of OOD images, albeit with blurriness characteristic of VAEs. This is possible because, without a strong KL-divergence constraint, the VAE prioritizes reconstruction fidelity, enabling it to generalize to inputs that deviate from the training distribution. On the other hand, the annealed VAE, with its stronger KL regularization, tends to pull OOD inputs towards the learned CIFAR-10 distribution, resulting in even blurrier reconstructions that tend to look like an amalgam of CIFAR-10 images with vague correlations to the original input.

GANs, however, exhibit a different set of biases. When provided with OOD latent vectors (scaled, sparse, or uniformly sampled), GANs struggle to produce coherent images. The higher-dimensional GANs perform the worst, often generating noise rather than structured outputs. This indicates that GANs have learned a more rigid mapping from latent space to image space. The lack of an explicit encoder and the adversarial training process lead to a latent space that does not generalize well to inputs outside the learned distribution. This is because GANs learn the *process* of sampling from a specific distribution rather than the *distribution* of the data itself. This bias is a direct consequence of the adversarial training objective, which focuses on generating high-fidelity samples that can fool the discriminator rather than covering the entire data distribution.

VAEs are better suited for tasks that require a structured latent space, such as interpolation and representation learning, while GANs excel in generating high-fidelity images for applications like image synthesis and super-resolution.

4.6.2. CNN vs CLIP

The cross-domain evaluation on PACS with finetuned ResNet50's accuracy of 95.1% parellels with CLIP's average of 95.5% without any domain-specific training. Results from Table 8 details a key insight that inductive biases can be instilled through training objectives and data diversity, not merely architectural choices. CLIP's success illustrates that the scale and diversity of training data can be more influential than architectural innovations for developing human-aligned inductive biases.

These findings have several implications for model design and deployment: (1) **Multimodal pretraining** as a strategy for improving domain robustness, even when the downstream task is purely visual. (2) **Prompt engineering** as a tool for domain adaptation without fine-tuning. (3) **Semantic inductive biases** as more transferable than visual ones for out-of-distribution scenarios.

However, our findings and recent research also reveals the boundaries of this approach. The persistent challenges in sketch recognition and specific failure modes (Figure 12 and 14) indicate that even large-scale multimodal training cannot fully overcome fundamental distributional gaps. Recent findings suggest that multimodal models require exponentially more data to achieve linear improvements in downstream "zero-shot" performance (Mayilvahanan et al., 2024), indicating that CLIP's generalization capabilities, while impressive, may be fundamentally bounded by pretraining data diversity. Lastly, performance remains sensitive to prompt choice, requiring domain expertise to optimize text descriptions, and the exponential data requirements for performance improvements suggest that scaling alone may not solve generalization challenges.

5. Conclusion

Our comprehensive analysis of inductive biases across deep learning architectures reveals fundamental principles governing model generalization. While we examined CNNs, Vision Transformers, VAEs, and GANs, CLIP's performance particularly illustrates the power of human-aligned biases for robust learning.

Unlike supervised models that often rely on dataset-specific cues, CLIP's representations cluster by conceptual similarity, facilitating recognition across visual domains from photographs to sketches. These findings have important implications for model design: incorporating diverse training data and appropriate learning objectives can be more effective than architectural modifications alone for instilling beneficial biases. CLIP's success with a standard ViT architecture underscores this principle.

The generative models further illustrate how architectural choices induce fundamental trade-offs such as between diversity and fidelity. This prevents generative models from being generalizable in real-world scenarios. However, these limitations have found their own niche applications such as anomaly detection using VAEs. Moreover, although generalization may be more difficult, the biases in these models can make them better suited to specific tasks. A hybrid approach that combines the strengths of both models could potentially yield a generative model that balances diversity and fidelity more effectively, thus allowing for a broader avenue of applications.

The overarching lesson is clear: inductive biases significantly impact model generalization, and human-aligned biases—whether architectural or learned—are essential for robust performance under distribution shifts.

References

Mayilvahanan, P., Wiedemer, T., Rusak, E., Generic, M., Bethge, M., and Brendel, W. (2024). No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv preprint arXiv:2404.04125*.

OpenAI (2021). CLIP: Connecting text and images. Retrieved from https://openai.com/index/clip/.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual representations from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Geirhos, Robert, et al. "ImageNet-trained CNNs Are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness." arXiv.org, 29 Nov. 2018 In arxiv.org/abs/1811.12231.

Kaiser, Nikolas Adaloglou Tim. "Understanding Vision Transformers (ViTs): Hidden Properties, Insights, and Robustness of Their Representations — AI Summer." AI Summer, 21 Feb. 2023, theaisummer.com/vit-properties/#imagenet-pretrained-cnns-are-biased-towards-texture. Retrieved from theaisummer.com/vit-properties/.