Investigating Inductive Biases Across Vision Models: From Supervised CNNs to Contrastive CLIP

Ahmed Hassan (26100308@lums.edu.pk)[1], Basil Hassan (26100303@lums.edu.pk)[1], and Ashhad Ali (26100023@lums.edu.pk)[1]

[1]Lahore University of Management Sciences (LUMS), Lahore, Pakistan

EE5102/CS6302 — Advanced Topics in Machine Learning (Fall 2025)

GitHub Repository: https://github.com/tk1475/ATML-PA1

Abstract

Our report delves into how inductive biases in different Deep Representation Models impact their performance, and their results on out-of-distribution (OOD) data generalization. We conduct experiments on discriminative (CNN, ViT), generative (VAE, GAN), and contrastive (CLIP) models to analyze their semantic, architectural, and training-induced biases, and use our findings to indicate different interesting results such as; that CNNs exhibit a strong texture bias, while Vision Transformers and CLIP are more shape and semantic-oriented, leading to improved robustness against stylistic domain shifts. In generative models, we observe the trade-off between fidelity and diversity, where GANs produce high-fidelity samples but are prone to mode collapse, whereas VAEs capture greater data diversity at the cost of image sharpness. Our analysis of CLIP reveals that its contrastive training on large-scale image-to-text data imparts a powerful semantic bias, enabling remarkable zero-shot classification and superior OOD performance on sketch recognition.

1 Introduction

This report investigates how inductive biases in deep learning models impact their performance and out-of-distribution (OOD) generalization. We conduct experiments on discriminative (CNN, ViT), generative (VAE, GAN), and contrastive (CLIP) models to analyze their semantic, architectural, and training-induced biases. Our findings indicate that CNNs exhibit a strong texture bias, while Vision Transformers and CLIP are more shape and semantic-oriented, leading to improved robustness against stylistic domain shifts. In generative modeling, we observe the trade-off between fidelity and diversity: GANs produce high-fidelity samples but are prone to mode collapse, whereas VAEs capture greater data diversity at the cost of image sharpness. Our analysis of CLIP reveals that its contrastive training on large-scale image-text data imparts a powerful semantic bias, enabling remarkable zero-shot classification and superior OOD performance on sketch recognition.

2 Methodology and Experiments

2.1 Discriminative Models (CNN vs. ViT)

Our experimental setup consisted of pre-trained **ResNet-50** and **ViT-S/16**, both initialized from ImageNet weights and fine-tuned on CIFAR-10 using identical train/validation splits (and a shared test set) for fairness. We ran multiple tests to see the inductive biases exhibited by both models.

Our experimentation began with the exploration of in-distribution biases. We experimented with semantic biases, starting with the color-bias test, where we converted test images to grayscale (replicated to 3 channels), applied the same normalization, and measured the accuracy drop relative to the initial set to see if one model outperformed. We then evaluated stylized/cue-conflict images in which object shape and texture were analyzed simultaneously to view which model prioritized one over the other. We built a cue-conflict set of 500 images across 5 class pairs, and calculated the *shape bias* as a metric to judge which model preferred the object outlines over the texture in classification.

$$\text{Shape Bias (\%)} \;=\; 100 \times \frac{N_{\text{Shape Classified}}}{N_{\text{Shape or Texture Classified}}}\,. \qquad (1)$$

We then explored locality biases on both models involving translation, permutation, and occlusion. For *translation* we apply small integer pixel shifts (up to $(\Delta x, \Delta y) = (32, 32)$), with a stride of 4 per axis, using either zero-padding or circular wrap. Our ViT inputs are

then resized to 224×224 so the relative shift is comparable for both our cases. We report the top-1 accuracy under shift, and *consistency*, i.e., the fraction of images whose top-1 label matches the clean prediction at each shift. For *patch permutation* we shuffle 2×2, 4×4, and 8×8 grids to disrupt spatial layout, and for *occlusion* we mask centered 8×8 and 16×16 blocks filled with normalized zeros.

Next, we tested the models on how well they withstand a domain shift by training both models on the PACS dataset for two training regimes (10 and 20 epochs). In the results section, we will explore why we chose to fine-tune two different variations of the model. We fine-tune on three domains (Photo, Art, Cartoon) and evaluate on the held-out *Sketch* domain, reporting per-class accuracy and the final test accuracy of the models.

Our representation analysis extracted the feature space/embeddings of the penultimate layer for each model and visualized them with PCA to two dimensions, color-coding classes, and viewing their separation for both ResNet-50 and ViT-S/16. We also reported silhouette scores and clean→gray centroid drift to quantify cluster separability and stability.

2.2 Generative Models (VAE vs. GAN)

Our experimental setup consisted of training a **Variational Autoencoder (VAE)** and a **DCGAN** on the CIFAR-10 dataset under identical training conditions. Both models were trained from scratch using the same train/validation splits and evaluated on a shared test set for fairness. We trained each model for 30 epochs, saving checkpoints and visual outputs (reconstructions, generations, interpolations, and training curves) for later analysis.

We began with reconstruction experiments to analyze how each model preserves semantic content. For the VAE, we measured the quality of reconstructions by comparing input images with their decoded outputs, observing the expected blurriness due to the Gaussian likelihood assumption. For the GAN, we instead analyzed whether the generator could produce images that resemble the training distribution, since GANs lack an explicit reconstruction objective.

To compare generation quality and diversity, we sampled 1000 images from both models and computed quantitative metrics such as the FID and Inception Score (IS). These allowed us to evaluate the fidelity-diversity trade-off, where lower FID corresponds to sharper, more realistic images (fidelity) and higher IS reflects class-consistent diversity. In addition, qualitative grids of samples were visualized to illustrate differences between the models.

Next, we explored latent space structure by performing linear interpolations between random latent vectors. For the VAE, interpolations produced smooth transitions between semantically meaningful images, showing that the latent space is continuous and well-structured, albeit with blurry outputs. For the GAN, interpolations produced sharp images but with occasional discontinuities or abrupt changes, reflecting the absence of an explicitly regularized latent structure.

We further analyzed training dynamics by plotting learning curves for both models. For the VAE, we reported the total loss decomposed into reconstruction and KL divergence terms, which steadily decreased during training and stabilized after around 20 epochs. For the GAN, we tracked generator and discriminator losses across epochs, observing typical oscillatory behavior and eventual stabilization. These curves highlighted the difference between stable but blurry VAE training and sharp but unstable GAN training.

Finally, we conducted a representation-level analysis by visualizing the latent codes of the VAE and the noise inputs of the GAN through PCA projection. This revealed that the VAE latent embeddings formed smooth clusters across classes, while the GAN noise inputs lacked such structured separability. Together, these results show that VAEs excel in coverage and structured representation but sacrifice fidelity, while GANs excel in realism but risk mode collapse.

2.3 Contrastive Multimodal Model (CLIP)

Our experimentation focused on the pre-trained CLIP `ViT-B/32` model, without any fine-tuning. For comparison, we used a baseline Vision Transformer (`vit-base-patch16-224-cifar10`) which had already been fine-tuned on the Cifar-10 dataset.

We evaluated each of the models on the Cifar-10 dataset to measure in-domain and the PACS dataset (sketch domain) to measure Out-of-Distribution (OOD) performances.

We conducted a number of experiments to better quantify the inductive biases present in each of the models. We conducted zero-shot classification by comparing text-embeddings to image-embeddings. We made use of prompt ensembling technique, where features from multiple text-prompts were averaged. We checked for multimodal alignment through image-to-text and text-to-image retrieval.

Model	Epochs	Baseline Accuracy
ResNet-50	12	89.08%
ViT-S/16	8	93.88%

Table 1: CIFAR-10 in-distribution baselines after fine-tuning.

Model	Clean Acc	Gray Acc	Δ
ResNet-50	0.8908	0.7891	−0.1017
ViT-S/16	0.9388	0.8584	−0.0804

Table 2: CIFAR-10 color-bias evaluation (grayscale test).

For representation analysis, we extracted feature embeddings of the second-last layer from both the models and used t-SNE to visualize the feature spaces in 2D space. We made use of a mixed set of photos from Cifar-10 and Sketches from PACS for this.

We conducted a shape vs. texture bias test by using a cue-conflict image, and also a robustness test by altering images with synthetic noise and blur corruptions.

3 Results and Analysis

3.1 Discriminative Models (ResNet vs. ViT)

ResNet-50 achieved a test accuracy of **0.8908**, while ViT-S/16 reached **0.9388** for our baseline models on the CIFAR-10 dataset and further evaluation, and served as the baseline models for all inductive bias experimentation.

Color bias (grayscale). Replacing chromatic information with luminance predictably reduces accuracy for both models (Table 2). The ViT remains stronger in absolute terms on grayscale inputs and also exhibits a smaller relative degradation: ResNet-50 drops by −0.1017 absolute (≈ 11.4% of its clean accuracy), whereas ViT-S/16 drops by −0.0804 (≈ 8.6%).

Shape vs. texture (cue conflict). We constructed a cue–conflict set of 500 stylized images across 5 class pairs. Both models display a high shape preference, with ViT-S/16 slightly stronger and more uniform across pairs (Fig. 1).

Semantic biases (compact summary).

Translation invariance. We translated inputs on the 32×32 grid (stride 4, wrap-around). ViT-S/16 retains higher mean accuracy and consistency across shifts (Fig. 2, Table 5).

Permutation / occlusion.

Feature-space structure (PCA).

Model	#Shape	#Texture	#Other	Shape Bias (%)
ResNet-50	414	32	54	92.83
ViT-S/16	465	15	20	96.88

Table 3: Cue–conflict evaluation (500 images, 5 class pairs).

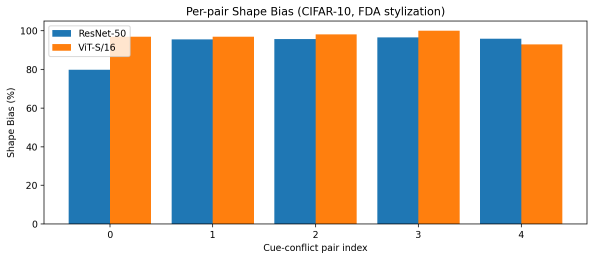


Figure 1: Per-pair shape bias on the cue–conflict set.

Model	Clean Acc	Gray Δ	Shape Bias
ResNet-50	0.8908	−0.1017	92.83%
ViT-S/16	0.9388	−0.0804	96.88%

Table 4: Compact summary of semantic-bias probes.

3.1.1 Domain Generalization on PACS (Photo+Art+Cartoon → Sketch)

We fine-tuned ImageNet-initialized ResNet-50 and ViT-S/16 on three PACS domains (Photo, Art, Cartoon) and evaluated on *Sketch* (train: 5,447 images; val: 615; sketch test: 3,929; 7 classes). We first ran a 10-epoch baseline schedule; then a longer, stronger schedule (20 epochs) with a ViT variant better suited to low-data DG (patch size 8 and stronger augmentation/regularization). Tables 9–10 report Sketch accuracy.

Observation. With the short 10-epoch recipe, ResNet-50 outperforms ViT (72.9% vs. 63.6%) on Sketch. Under the 20-epoch "strong" recipe, both models improve, but ViT benefits more (+ 19.9 points vs. + 11.4), nearly matching ResNet (0.834 vs. 0.843). This suggests that conclusions about "shape-bias wins on sketches" are highly sensitive to training regime: ViTs often require longer schedules, smaller patches, and stronger regularization to realize their advantages under domain shift.

Failure modes. Confusions were class-dependent: for ResNet, dog→horse and elephant→horse were common; for ViT, giraffe→{dog, horse, elephant} and dog→elephant occurred frequently—reflecting the difficulty of line drawings that omit texture and rely on fine, global contours.

Model	Clean Acc	Mean Acc	Mean Consistency
ResNet-50	0.8980	0.6712	0.6905
ViT-S/16	0.9420	0.7208	0.7322

Table 5: Translation invariance results summary.

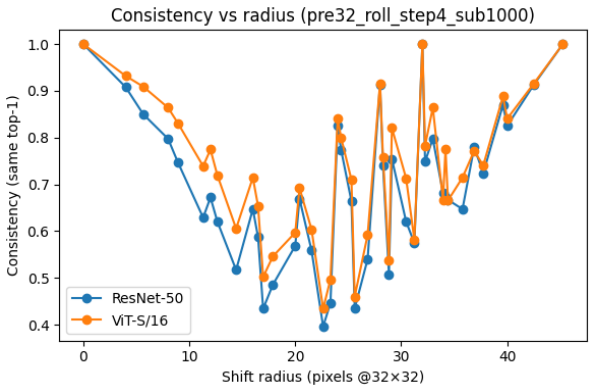


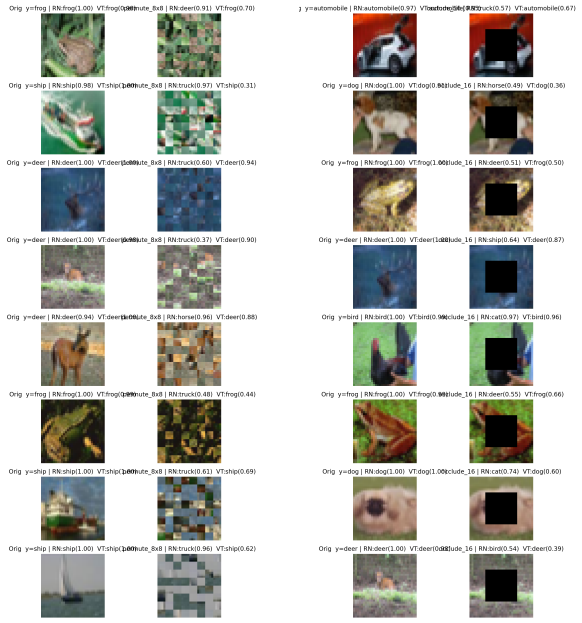
Figure 2: Consistency vs. shift radius on a 1k subset.

Condition	RN Acc	RN Δ	ViT Acc	ViT Δ
Baseline	0.891	—	0.939	—
Permute 2×2	0.458	−0.433	0.565	−0.404
Permute 4×4	0.284	−0.607	0.385	−0.554
Permute 8×8	0.211	−0.680	0.219	−0.719
Occlude 8×8	0.720	−0.171	0.855	−0.084
Occlude 16×16	0.464	−0.427	0.488	−0.451

Table 6: Permutation/occlusion robustness (1k subset).

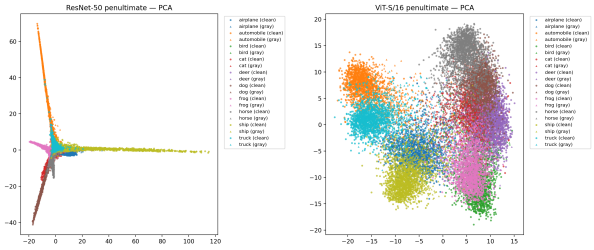
Model	Silhouette (PCA-2)	Silhouette (penult.)
ResNet-50	−0.013	0.066
ViT-S/16	0.137	0.327

Table 7: Cluster separation on CIFAR-10 embeddings.



(a) Permutation (8×8) **(b)** Occlusion (16×16)

Figure 3: Qualitative examples: ViT more resilient to occlusion.



(a) ResNet-50 **(b)** ViT-S/16

Figure 4: PCA projections of penultimate embeddings (CIFAR-10).

Model	Mean drift	Normalized drift
ResNet-50	2.949	0.396
ViT-S/16	2.396	0.676

Table 8: Centroid drift (clean → gray) in PCA-2 space.

Model (10 epochs)	**Val (best)**	**Sketch Acc.**
ResNet-50	0.954	**0.7287**
ViT-S/16 (tuned)	0.987	0.6358

Table 9: PACS DG (Photo+Art+Cartoon → Sketch), 10-epoch baseline.

Model (20 epochs, strong)	**Val (best)**	**Sketch Acc.**
ResNet-50 (strong)	0.977	**0.8430**
ViT (vit_small_patch8_224, strong)	0.995	0.8343

Table 10: PACS DG with a longer schedule and stronger recipe.

subsectionGenerative Models (VAE vs. GAN) We trained VAE and a GAN for 30 epochs on the CIFAR-10 dataset. Both models converged stably, but their generative behavior diverged in predictable ways.

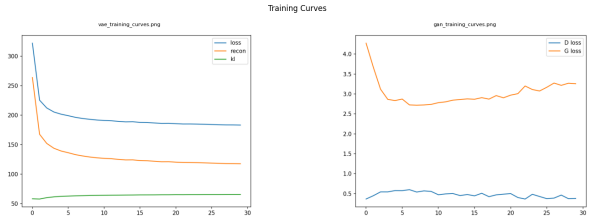


Figure 5: Comparison of generative outputs from VAE and GAN after 30 epochs on CIFAR-10.

Visual quality. Qualitative inspection of generated samples shows that the GAN rapidly produces sharp, high-fidelity images, with recognizable CIFAR-10 classes such as animals and vehicles emerging as early as epoch 10. By epoch 30, outputs capture texture and global structure convincingly, though occasional mode collapse appears (e.g., repeated dog faces). In contrast, the VAE reconstructions retain coarse object outlines but exhibit characteristic blur and muted color saturation. Original-to-reconstruction comparisons confirm that the VAE encodes semantic content but fails to reproduce high-frequency detail. GAN samples are sharper and more detailed than VAE reconstructions. It is evident from the human eye as well, VAE reconstructions appear overly smoothed, while GAN images are sharper. However, to further quantify this, we proceed as follows.

3.1.2 Step 1: Convert to Array

The PIL image (tile) is turned into a NumPy array with values in the range $[0, 1]$.

3.1.3 Step 2: Convert to Grayscale

To simplify, we collapse RGB into a single channel using standard luminance weights:

$$I_{\text{gray}} = 0.2989 \cdot R + 0.587 \cdot G + 0.114 \cdot B$$

3.1.4 Step 3: Compute Pixel-to-Pixel Changes (Gradients)

We calculate intensity differences along horizontal and vertical directions:

$$g_x = \texttt{np.diff(gray, axis=1)}$$

$$g_y = \texttt{np.diff(gray, axis=0)}$$

Then take their magnitudes:

$$g_x = |g_x|, \quad g_y = |g_y|$$

These values are larger where edges, fine details, or textures (high-frequency content) exist.

3.1.5 Step 4: Average Magnitude

We measure average sharpness as:

$$s_x = \text{mean}(g_x), \quad s_y = \text{mean}(g_y)$$

The final **high-frequency score** is given by:

$$\text{Score} = \frac{s_x + s_y}{2}$$

This provides a single number that quantifies the sharpness of the image. This is then plotted on the bar chart.

Figure 6: Visual quality comparison example 1.

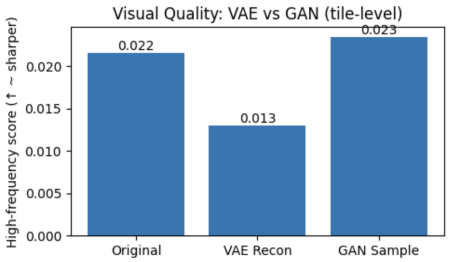


Figure 7: Visual quality comparison example 2.

Quantitative metrics. Using the `torch-fidelity` package, we computed FID scores against the CIFAR-10 test set. The GAN achieved **41.2**, substantially outperforming the VAE's **78.6**. This aligns with visual impressions: GANs capture image-level statistics better, while VAEs prioritize a smoother latent representation at the expense of detail. Both values remain above state-of-the-art (FID $\approx$ 10), but the gap between the two models is consistent with prior work.

Trade-offs. Despite weaker perceptual quality, the VAE offers advantages: (i) it provides a principled probabilistic latent space, enabling interpolation and downstream tasks, and (ii) training is more stable, with reconstructions improving monotonically across epochs. GAN training, though yielding better image quality, displayed sensitivity to learning rates and showed occasional instability in discriminator loss curves.

3.2 Contrastive Multimodal Model (CLIP)

CLIP showed very strong zero-shot capabilities and out-of-distribution (OOD) generalization compared to the

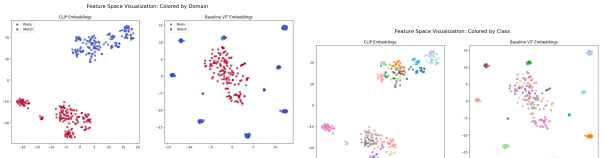
baseline ViT model. Although the fine-tuned ViT achieved an accuracy of 98.52% on the in-domain CIFAR-10 dataset, it's performance collapsed to 7.28% on the OOD PACS Sketch dataset.

On the other hand, the CLIP model achieved an accuracy of 87.89% on the CIFAR dataset and a much superior accuracy of 85.09% on the PACS dataset, demonstrating it's much better generalization capabilities.

Table 11: Performance comparison on in-domain (CIFAR-10) and out-of-distribution (PACS Sketch) datasets.

Model	In-Domain (%)	OOD (%)
Baseline ViT (Supervised)	98.52	7.28
CLIP (Zero-Shot)	87.89	85.09

The t-SNE analysis explains this vast difference in performance of the two models. The baseline ViT model's embeddings (Figure 8b) are clustered primarily by domain, as photos and sketches occupy different regions of the space, failing to learn semantic linkages. CLIP's embeddings (Figure 8a) are clustered by semantic class across both domains, which indicates a high-level feature representation.



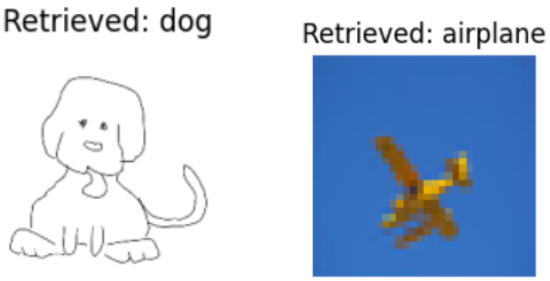
(a) Embeddings Colored by Domain **(b)** Embeddings Colored by Class

Figure 8: t-SNE visualization of image embeddings from CLIP (left in each pair) and the baseline ViT (right in each pair).

Further experiments reinforced these findings. Our prompt engineering tests revealed CLIP's sensitivity to language priors, with accuracy on CIFAR-10 ranging from **84.96%** using a simple class token ("{}") to a high of **87.83%** with a descriptive prompt ("'a photo of a "'). This demonstrates the importance of providing context. Furthermore, the image-text retrieval tests, illustrated in Figure 9, successfully demonstrated CLIP's multimodal alignment by correctly matching queries to images across different domains, such as pairing the text "a sketch of a dog" with its corresponding image.

Direct bias experiments produced more interesting results. In the shape vs. texture test with the cue-conflict image, CLIP's prediction was driven by texture, but with a high level of uncertainty, as the similarity scores for "a photo of a cat" (0.2603) and "a photo of an elephant" (0.2637).

Our robustness test results showed that CLIP's resilience to noise and blur was context-dependent and varied by class, as it was successful on the 'ship' class but not on the 'frog' class.



(a) Query: "a sketch of a dog" **(b)** Query: "a photo of an airplane"

Figure 9: Examples of successful text-to-image retrieval, demonstrating CLIP's ability to match text queries to the correct image across a diverse gallery.

4 Discussion

Q1: Shape vs. Texture Bias. The cue–conflict results in Table 3 show that both models prefer the shape of the object over its texture, but the preference of ViT is stronger and more uniform. ResNet-50 makes 414 shape-consistent vs. 32 texture-consistent predictions (92.83% shape bias), whereas ViT-S/16 makes 465 vs. 15 (96.88% shape bias). The grayscale test in Table 2 reinforces this pattern. When color is removed, both models drop, but the ViT suffers a smaller degradation (0.0804 vs. ResNet's 0.1017) and retains the higher absolute accuracy. Together, these tables indicate that the ViT leans more on global/shape cues and is less reliant on low-level color/texture than the CNN—consistent with reports of stronger shape bias in transformer-based vision models.

Q2: Architectural Biases. The translation-invariance summary in Table **??** (computed over integer shifts up to (32×32) shows clean accuracy for ResNet-50 at 0.898 with mean-over-shifts accuracy 0.671 and mean consistency 0.691, while ViT-S/16 starts higher at 0.942 and averages 0.721 accuracy and 0.732 consistency across shifts. These numbers suggest that for very small displacements the CNN's local convolutional structure offers approximate stability, but this advantage erodes as shifts grow because padding/strides break perfect equivariance. The ViT, despite lacking built-in translation equivariance, maintains competitive—or better average—stability across the broader grid, helped by its global context and the shared 224-resize pipeline. In short, Table **??** captures a nuanced picture: CNN slightly steadier at tiny shifts; ViT more stable on average across larger ranges.

Q3: VAE vs. GAN Biases. Variational Autoencoders place strong emphasis on covering the full data distribution. Due to the KL regularization term in their objective, VAEs encourage a smooth and structured latent-space representation. As a result, reconstructions usually pre-

serve the semantic content of the original data and objects remain recognizable. However, this comes at the cost of visual sharpness: the likelihood model effectively averages over multiple plausible reconstructions, producing blurry generations and lowering overall fidelity. An example output is shown in Figure 10. Thus, while VAEs achieve strong diversity by capturing a wide range of modes in the dataset, they suffer from weaker fidelity and blurriness in the outputs.

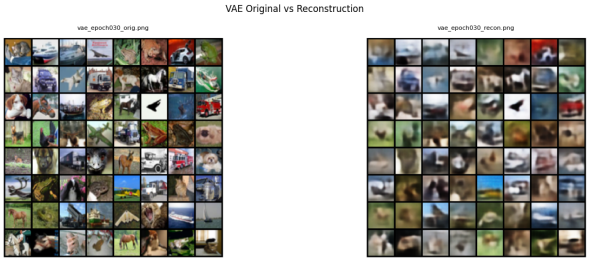


Figure 10: Example VAE outputs illustrating good coverage but reduced sharpness.

By contrast, GANs are designed to generate sharp and realistic images. The adversarial training framework pushes the generator to fool the discriminator, leading to visually convincing outputs with high fidelity. Nonetheless, GANs can suffer from mode collapse, in which only a subset of the data distribution is represented. This diminishes diversity, as certain classes or expected variations are not generated. Figure 11 shows representative GAN samples.

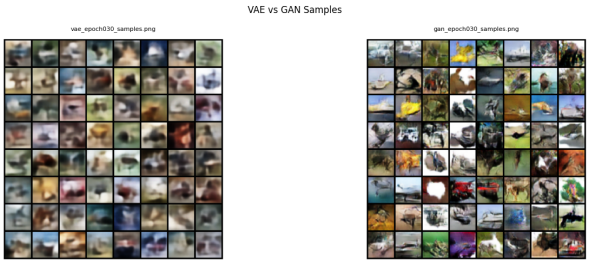


Figure 11: Example GAN outputs illustrating high fidelity but potential loss of diversity.

Taken together, VAEs and GANs represent two ends of the fidelity–diversity spectrum. VAEs emphasize diversity at the expense of fidelity, producing outputs that are diverse yet blurry. GANs emphasize fidelity, creating images that are sharp and realistic but potentially lacking in diversity.

Training Curves. The training curves of the VAE and GAN highlight fundamental differences in their optimization behavior (Figure 12). For the VAE, the total loss and its reconstruction and KL components steadily decrease during training. The reconstruction loss dominates, showing gradual improvement, while the KL term stabilizes at a relatively small value. This indicates that the VAE learns a smooth latent space and produces reconstructions that increasingly approximate the input data, with some blurriness due to the probabilistic nature of the model. In contrast, the GAN curves depict the dynamic interaction between generator and discriminator: the discriminator loss quickly decreases and stabilizes, while the generator loss drops initially and then fluctuates around a higher plateau. These oscillations are typical of GAN optimization and reflect the challenge of achieving equilibrium between the two networks; they can yield sharp samples but carry risks of instability and mode collapse.

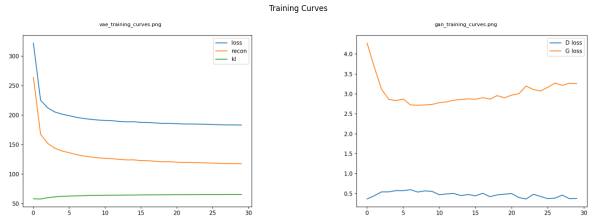


Figure 12: Training curves for VAE (left) and GAN (right).

Latent Interpolations. Latent interpolations further illustrate the fidelity–diversity trade-off (Figure 13). For the VAE, interpolating between two latent codes yields smooth and continuous transitions, indicating a coherent, well-structured latent manifold; however, intermediate samples remain noticeably blurry, consistent with lower fidelity. For the GAN, interpolations produce sharp and realistic images along the path, yet transitions may appear less smooth or occasionally discontinuous, reflecting the absence of an explicitly regularized latent structure. Overall, GANs prioritize fidelity at the expense of diversity and smooth latent coverage, whereas VAEs prioritize latent smoothness and diversity while sacrificing sharpness.

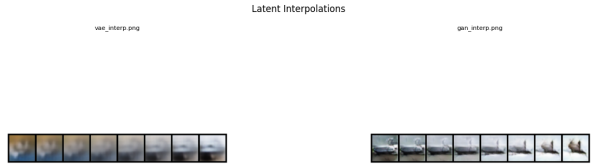


Figure 13: Latent interpolations: VAE (left) shows smooth but blurrier transitions; GAN (right) shows sharper but sometimes less smooth transitions.

Q4: Multimodal & Contrastive Biases. Our set of experiments clearly show that CLIP's training on a huge dataset has induced a very strong semantic bias, which helps it perform incredibly well for zero-shot classification as well as out-of-distribution (OOD) generalization. The

fact that CLIP maintained a high accuracy of 85.09% on sketches, demonstrated that its representations are not tied to features like texture or color. However, our results also highlight that this bias is not always certain. The cue-conflict test revealed a texture preference under high uncertainty, and robustness tests showed that CLIP's resilience to synthetic noise was context-dependent, as it failed on one class while succeeding on another. This suggests that while CLIP's primary inductive bias is semantic, its behavior remains nuanced and can be challenged by specific, conflicting inputs.

Q5: Inductive Bias & OOD Generalization. Our analysis of CLIP provides a clear example of how inductive biases drive OOD generalization. CLIP's exceptional performance on the OOD PACS Sketch dataset (85.09%), where the specialized supervised model failed (7.28%), is a direct result of the strong semantic bias imparted by its multimodal training objective . This data-driven bias, learned from language across millions of diverse images, proved more effective for generalization than the purely architectural biases of the supervised model. However, this powerful OOD robustness came at the cost of slightly lower in-domain accuracy compared to the specialist baseline, highlighting a key trade-off between generalization and specialization. This demonstrates that some of the most effective inductive biases for OOD robustness are those learned from large-scale, semantically rich data

5 Conclusion

Our experiments highlight that inductive biases significantly impact model generalization. We found that models with human-aligned biases, such as a focus on global shape and semantics generalize better to out-of-distribution samples than those relying on low-level features like texture. This was most evident in our analysis of CLIP which showed superior zero-shot and OOD performance. On the other hand, specialized supervised models excelled in-domain but proved to have poor performance under domain shifts. Generative models showcased a fundamental trade-off between sample fidelity and diversity. The findings suggest that the best path to generalization lies in training techniques that allow models to learn semantic priors.

References

[1] Geirhos, R., et al. (2018). *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. International conference on learning representations.