An Empirical Study of Understanding Deep Learning Architectures

Raahim A Samad Poonawala [1]

GitHub Repository: https://github.com/Raahim58/ATML

Abstract

This report presents an empirical analysis of four fundamental deep learning architectures: ResNet-152, Vision Transformer (ViT), Generative Adversarial Networks (GAN), and Variational Autoencoders (VAE). I investigate transfer learning effectiveness, attention mechanisms, generative modeling dynamics, and multimodal alignment through hands-on experiments. Key findings include the critical importance of residual connections for deep networks, ViT's interpretable attention patterns, common GAN training pathologies, and the modality gap in CLIP embeddings.

1. Introduction

Deep learning has revolutionized computer vision and generative modeling through specialized architectures. Understanding how these models work in practice - their strengths, failure modes, and training dynamics - is crucial for effective deployment. This study explores four key architectures through controlled experiments to provide practical insights for practitioners.

The scope encompasses transfer learning with ResNet-152, attention visualization in Vision Transformers, deliberate pathology induction in GANs, posterior collapse analysis in VAEs, and modality gap investigation in CLIP. Each experiment targets specific architectural properties that influence real-world performance and deployment considerations.

2. Task 1: ResNet-152 Transfer Learning

2.1. Methodology

I employed a pre-trained ResNet-152 from PyTorch (torchvision.models.resnet152(pretrained=True)) with the final classification layer replaced for CIFAR-10's 10 classes. The backbone remained frozen while training only the classification head. Images were resized to 224×224 to match ImageNet training dimensions. Training used Adam optimizer with learning rate 1e-3 for 5 epochs with batch size 128 and num workers=2 for parallel computation efficiency. For residual connection analysis, I disabled skip connections in all blocks of layer2 by replacing F.relu(out + self.shortcut(x)) with F.relu(out) in the forward pass. Feature extraction utilized forward hooks to capture activations from layer1, layer2-3, layer4, and average pooling layers for t-SNE and UMAP visualization with default parameters.

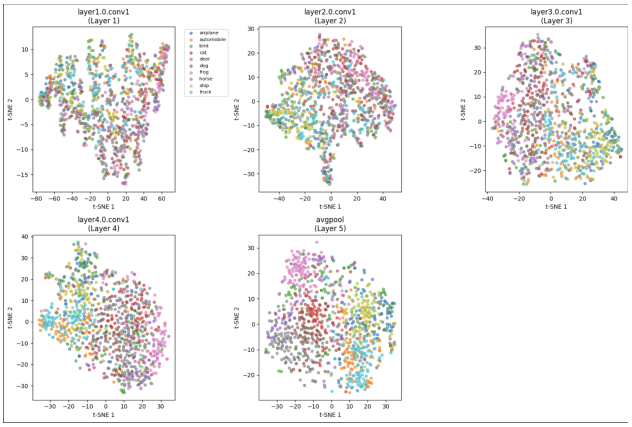

Figure 1. Visualization of ResNet features

2.2. Results

Training ResNet-152 from scratch on small datasets proves unnecessary due to three critical factors. First, the model contains approximately 60 million parameters requiring massive computational resources and weeks of GPU training time. Second, CIFAR-10's limited 50,000 training samples create severe overfitting risks with such parameter-heavy architectures. Third, pre-trained ImageNet features already capture universal visual patterns like edges, textures, and shapes that transfer effectively across domains.

Residual connections demonstrate fundamental importance for deep network training. My experiments showed baseline accuracy of 81.80 percent with intact skip connections versus 16.20 percent when disabled - a devastating 65.60 percent performance drop. This dramatic degradation illustrates how residual pathways enable gradient flow through very deep architectures, preventing the vanishing gradient problem that plagued early deep networks.

Feature hierarchy analysis reveals clear progression from

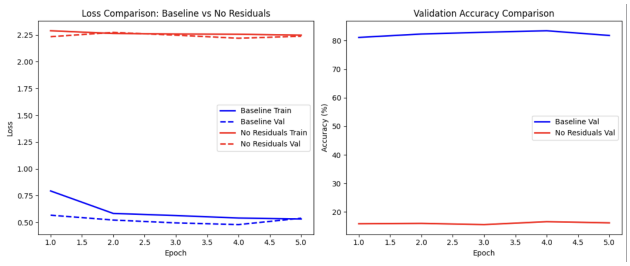

Figure 2. ResNet Baseline vs no skip loss

general to specific representations across network depth. Early layers (layer1) learn low-level features like edges and textures with high class overlap, making them universally transferable. Middle layers (layer2-3) develop intermediate patterns with emerging class separation. Late layers (layer4 and average pool) capture high-level semantic features with strong class separability optimized for the specific task.

Transfer learning comparison validates the efficiency of pre-trained approaches. Pretrained weights with head-only fine-tuning achieved 81.80 percent accuracy with minimal computational cost, training only a fraction of total parameters. Random initialization severely degraded performance, confirming that ImageNet features provide crucial initialization benefits even for different domains like CIFAR-10.

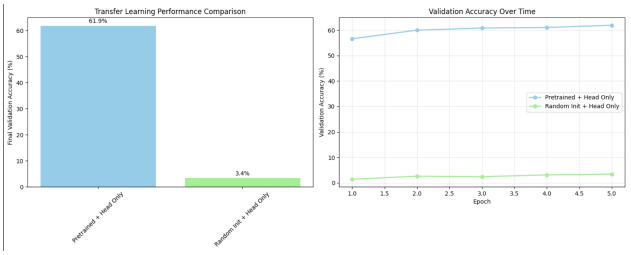

Figure 3. ResNet transfer learning

2.3. Discussion

Skip connections fundamentally solve the vanishing gradient problem in deep networks by providing direct pathways for gradient flow to early layers. Without these connections, gradients diminish exponentially through backpropagation, leaving early layers unable to learn effectively. The 65 percent performance drop demonstrates their critical role in enabling very deep architectures like ResNet-152.

The observed feature hierarchy validates transfer learning effectiveness by showing how different network depths serve distinct representational purposes. Early layers learn universal low-level features that remain consistent across visual domains, explaining why freezing these layers during fine-tuning works effectively. Later layers become increasingly task-specific, justifying why the final classifier always requires retraining for new domains.

Practical deployment considerations favor pretrained weights with head-only fine-tuning as the optimal trade-off between computational efficiency and performance. This approach minimizes training time, reduces hardware requirements, and achieves competitive accuracy by leveraging the universal feature representations learned during ImageNet pretraining.

3. Task 2: Vision Transformer Analysis

3.1. Methodology

I utilized the pre-trained google/vit-base-patch16-224 model from HuggingFace for attention analysis and robustness testing. Attention weight extraction focused on the CLS token's attention to patch tokens from the final transformer layer, configured with output attentions=True. The 196-dimensional attention vector (14×14 patches for 224×224 images with 16×16 patches) was reshaped to spatial dimensions and visualized as heatmap overlays using alpha blending.

Robustness experiments involved masking 0.3-0.7 of patches during inference using two strategies: random masking with uniform sampling and structured center masking targeting the middle 7×7 patch region. Center masking identified patches by distance from image center, collecting the nearest mask fraction patches for equivalent comparison to random masking. Linear probe evaluation compared CLS token embeddings versus mean-pooled patch embeddings using logistic regression on downstream classification tasks.

3.2. Results

Attention visualization successfully demonstrated ViT's interpretability advantages. The model correctly focused on object boundaries in test images, with attention maps clearly outlining the dog's figure in my experiments. This behavior confirms that ViT bases predictions on semantically relevant image regions rather than spurious correlations.

Masking robustness revealed differential vulnerability patterns. Random masking showed minimal accuracy degradation even with 30 percent missing patches, demonstrating the model's ability to interpolate missing information from neighboring regions. Conversely, structured center masking caused significant performance drops when object-centric regions were removed, indicating heavy dependence on semantically important patches.

Pooling strategy comparison validated CLS token superiority for supervised pretrained models. The CLS token probe consistently outperformed mean pooling approaches, reflect-

ing its specialized role in aggregating global information during supervised pretraining. This specialization makes CLS tokens particularly effective for classification tasks compared to generic patch embedding averaging.

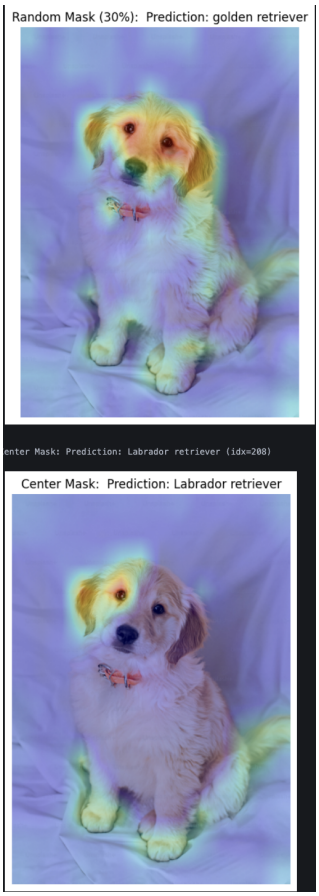

Figure 4. ViT masking comparison

3.3. Discussion

ViTs provide native interpretability advantages over CNNs through built-in attention mechanisms. While CNNs require post-hoc visualization tools like Grad-CAM that compute gradients with respect to feature maps, ViT attention weights offer direct insight into model focus areas without additional computation. Different attention heads exhibit specialization patterns - some focus locally on textures and edges while others attend globally to entire objects or contextual relationships, something I explored in another assignment.

The robustness patterns reflect ViT's architectural strengths and limitations. Random masking tolerance stems from global self-attention mechanisms that enable remaining patches to communicate across the entire spatial grid, allowing interpolation around missing regions. The training regime also includes dropout-like stochasticity that prepares the model for missing information. However, structured masking of semantically critical regions (typically image centers where objects reside) causes major performance degradation because it removes concentrated object information that cannot be reconstructed from peripheral context alone.

Attention occasionally exhibits peculiar behaviors worth noting for practitioners. Background regions with similar colors or textures to target objects sometimes attract spurious attention, suggesting the model relies on low-level visual similarities alongside semantic understanding. This behavior highlights the importance of diverse training data and careful evaluation when deploying ViTs in critical applications.

4. Task 3: GAN Training Dynamics

4.1. Methodology

I implemented an MLP GAN architecture for MNIST with generator layers [100→256→512→784] using ReLU activations and tanh output, and discriminator layers [784→256→256→1] with LeakyReLU(0.2) and sigmoid output. Training employed separate Adam optimizers with learning rate 2e-4, Beta=0.5, Beta=0.999, and MSE loss. Images were normalized to [-1,1] range to match the tanh output activation.

I deliberately induced three training pathologies through controlled perturbations. Gradient vanishing was triggered by increasing discriminator learning rate 10× to 2e-3 for the first 5 epochs, creating an overpowered discriminator. Mode collapse was induced by increasing generator learning rate 10× while maintaining discriminator at 2e-4, allowing the generator to exploit discriminator weaknesses. Discriminator overfitting was created by restricting training data to 1000 MNIST samples, forcing the discriminator to memorize limited examples.

Each pathology was followed by targeted mitigation strategies. Gradient vanishing was addressed through label smoothing (real labels = 0.9) and non-saturating generator loss. Mode collapse was mitigated using balanced training ratios (2:1 discriminator:generator steps). Discriminator overfitting was countered with dropout regularization (p=0.3) in hidden layers.

4.2. Results

Gradient vanishing manifested as expected when the discriminator overwhelmed the generator. Strong discriminator performance led to D loss approaching zero while G loss remained high and flat, indicating the generator received no meaningful learning signal. D(x) approached 1.0 for real samples while D(G(z)) approached 0.0 for generated samples, creating vanishing gradients for the generator. La-

bel smoothing and non-saturating loss successfully restored generator learning by maintaining non-zero gradient signals.

Mode collapse occurred predictably under unbalanced training regimes. High generator learning rates caused identical outputs across different noise inputs, with all generated samples resembling the same digit type. This behavior reflects a local equilibrium where the generator exploits a specific weakness in the discriminator rather than learning diverse data distributions. Balanced training ratios (training discriminator twice per generator step) successfully restored output diversity.

Discriminator overfitting emerged clearly on the limited 1000-sample dataset. The discriminator achieved 100 percent training accuracy but failed to generalize to unseen real samples, assigning low probabilities to new real images while maintaining high confidence on memorized training examples. This memorization behavior provided unhelpful gradients to the generator, preventing meaningful learning. Dropout regularization effectively prevented premature discriminator convergence.

Final GAN performance produced recognizable digit-like shapes comparable to reference implementations. My smaller batch size (32 vs 100) and label smoothing approach resulted in slightly softer outputs but more stable training dynamics throughout the experiment.

Figure 5. Vanilla GAN loss

4.3. Discussion

GAN training requires delicate equilibrium between generator and discriminator components, with each pathology representing different failure modes of this balance. Gradient vanishing occurs when the discriminator becomes too confident, providing no learning signal to the generator through saturated sigmoid outputs. This problem particularly affects the original minimax formulation where log(1-D(G(z))) approaches zero gradient as D(G(z)) approaches zero.

Mode collapse represents a different equilibrium failure

Figure 6. Vanilla GAN result visualization

where the generator discovers a specific output that reliably fools the discriminator and stops exploring other modes of the data distribution. This behavior is mathematically stable but undesirable, as it fails to capture the full diversity of the target distribution. The phenomenon explains why GAN evaluation requires diversity metrics alongside quality assessments.

Discriminator overfitting demonstrates how finite training sets can cause memorization rather than genuine pattern learning. An overfit discriminator provides poor teaching signals to the generator because it makes decisions based on memorized examples rather than generalizable features. This insight explains why larger datasets generally improve GAN performance and why regularization techniques prove essential for stable training.

These controlled experiments reveal that simple interventions can effectively address common GAN pathologies. Label smoothing, balanced training schedules, and regularization represent practical tools for practitioners facing unstable GAN training in real applications.

5. Task 4: VAE Posterior Collapse

5.1. Methodology

I trained a VAE on FashionMNIST using a convolutional encoder-decoder architecture with 3 convolutional blocks and latent dimension 64. The encoder outputs mean and log-variance vectors for reparameterization trick implementation: z = mean + std dev noise where noise N(0,I). The decoder symmetrically reconstructs 28×28 images from latent codes.

The loss function combined MSE reconstruction loss and KL divergence. Training used Adam optimizer with learning rate 1e-3 for 50 epochs, monitoring ELBO components separately to detect posterior collapse.

To address posterior collapse, I implemented -VAE with KL

annealing using cosine scheduling: (t) = 0.1 + 0.9 × (1 - cos(t/T))/2 where T=20 epochs. This approach gradually increases the KL weight from 0.1 to 1.0, allowing the encoder to develop meaningful representations before full regularization. I also experimented with alternative priors including Laplacian and exponential distributions for comparison.

5.2. Results

Initial training exhibited classic posterior collapse symptoms despite decreasing overall loss. Generated samples appeared as uniform blobs with minimal recognizable structure, indicating the decoder learned to produce average outputs independent of latent inputs. High reconstruction loss paired with very low KL divergence confirmed that encoder outputs matched the prior too closely, effectively ignoring input information.

The collapse mechanism became clear through ELBO component analysis. Early training phases showed KL divergence dominating the loss function, forcing encoder outputs toward the standard normal prior before meaningful representations could develop. This premature regularization caused the decoder to learn input-independent generation, treating all latent codes as equivalent.

-VAE with KL annealing provided significant but incomplete improvement. The mitigation strategy successfully improved latent space structure and interpolation capabilities, with smoother transitions between encoded test samples demonstrating better representation learning. However, generated samples remained somewhat blurry compared to other generative models, indicating the fundamental reconstruction-regularization trade-off in VAEs still limits generation quality.

Alternative prior experiments showed degraded results compared to Gaussian distributions. Laplacian priors produced less smooth latent interpolations with tendency to focus on single clothing types rather than diverse generation. This behavior suggests that Gaussian priors align better with the continuous optimization landscape of neural network training.

5.3. Discussion

Posterior collapse represents a fundamental challenge in VAE training where the balance between reconstruction accuracy and latent regularization fails. The phenomenon occurs when KL regularization overwhelms the reconstruction objective, causing encoders to learn trivial mappings that ignore input information. This failure mode explains why vanilla VAEs often produce blurry samples compared to GANs or other generative approaches.

The Beta-VAE annealing approach addresses collapse by implementing curriculum learning for the regularization

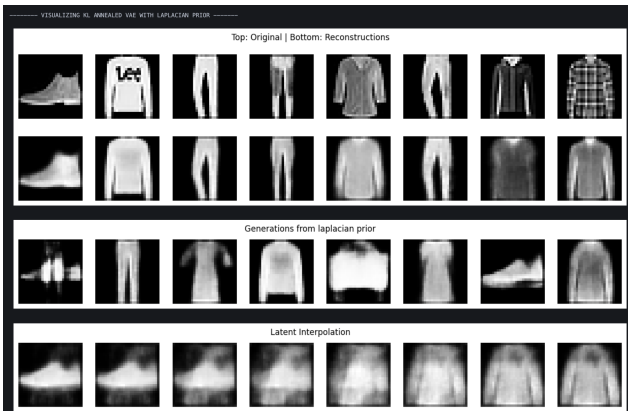

Figure 7. VAE visualization using Laplacian prior

term. By starting with reduced KL weight (=0.1), the encoder can develop meaningful representations that capture input variations before full regularization is applied. This scheduling allows the reconstruction term to establish useful encoder-decoder relationships that persist even under stronger regularization.

However, the persistent blur in VAE generations reveals inherent limitations of the reconstruction-based training objective. Unlike GANs that learn through adversarial competition, VAEs optimize pixel-wise reconstruction which tends to produce averaged outputs when uncertainty exists. This mathematical property explains why VAEs excel at interpolation and representation learning but lag behind GANs in sample sharpness.

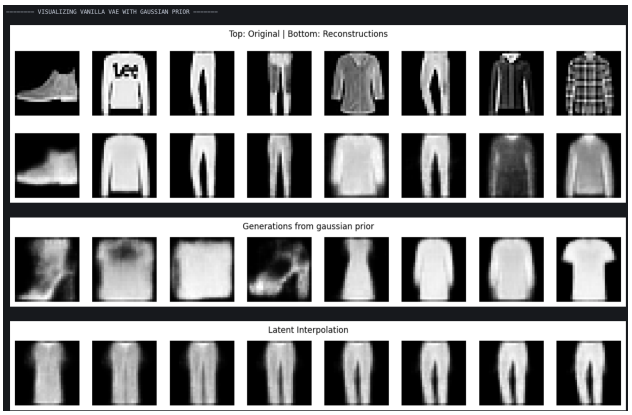

Figure 8. VAE visualization using Gaussian prior

The superior performance of Gaussian priors likely stems from their mathematical properties aligning with continuous optimization and the central limit theorem. Neural network activations often approximate Gaussian distributions, making Gaussian priors a natural choice for latent

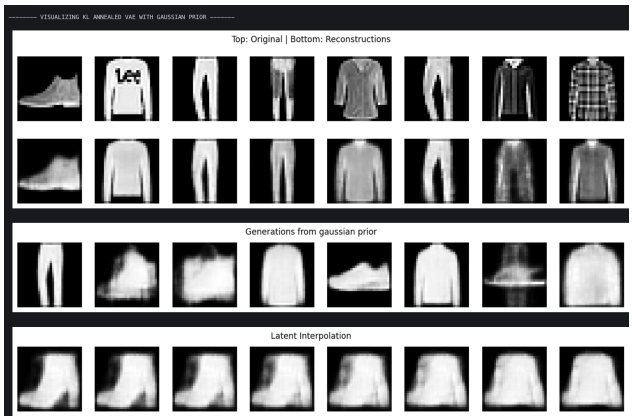

Figure 9. VAE visualization using Gaussian prior and KL annealing

space regularization compared to more exotic distributions.

6. Task 5: CLIP Modality Gap

6.1. Methodology

I evaluated CLIP's ViT-B/32 model on STL-10 using three prompting strategies across 1000 test samples. Plain prompts used direct class names ("airplane", "bird"), template prompts followed standard format ("a photo of a class"), and descriptive prompts employed enhanced descriptions ("a centered, high quality, beautiful clear photo of a class"). Text prompts were tokenized and encoded to 512-dimensional embeddings alongside image embeddings from the vision encoder.

Zero-shot classification computed cosine similarities between image embeddings and all text prompt embeddings, with predictions based on highest similarity scores. Modality gap analysis extracted embeddings for STL-10 samples and corresponding text labels, projecting them to 2D using t-SNE for visualization with different colors representing image versus text modalities.

Procrustes alignment implemented orthogonal transformation using scipy.linalg.orthogonalprocrustes to find rotation matrix R minimizing ——XR - Y—— where X represents image embeddings and Y represents corresponding text embeddings. The learned transformation was applied to test embeddings for re-evaluation of classification accuracy.

6.2. Results

Zero-shot performance revealed interesting prompting dynamics across the three strategies. Plain labels achieved 96.10 percent accuracy, template prompts ("a photo of a") improved to 96.60 percent, while descriptive prompts ("a centered, high quality, beautiful clear photo of a") decreased to 94.90 percent. This pattern suggests that moderate prompt engineering helps but excessive description may confuse the model.

Modality gap visualization confirmed clear separation between image and text embedding clusters in the shared representation space. Different modalities occupied distinct regions despite sharing semantic information, with the gap persisting even for semantically aligned image-text pairs. This separation indicates that CLIP's contrastive training creates structured but segregated embedding spaces.

Procrustes alignment achieved minimal improvement, with classification accuracy increasing only 0.4 percent after orthogonal transformation. The small gain suggests that CLIP's contrastive training already achieves reasonable cross-modal correspondence despite apparent visual separation in embedding space. The transformation successfully reduced visual separation in t-SNE plots but provided limited practical benefits.

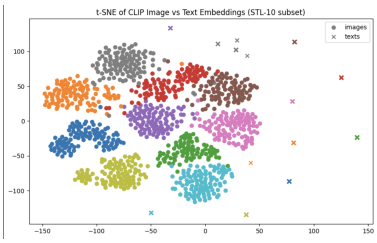

Figure 10. visualization before procustes alignments of shared latent space

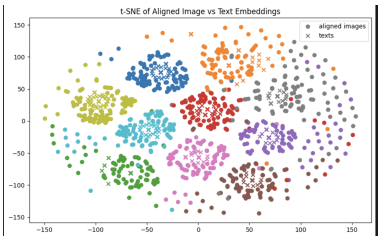

Figure 11. visualization after procustes alignments of shared latent space

6.3. Discussion

The modality gap phenomenon reveals fundamental properties of contrastive learning in multimodal systems. CLIP's training objective encourages matching between corresponding image-text pairs while pushing apart non-corresponding pairs, but doesn't explicitly enforce overlapping distributions for different modalities. This approach creates structured embedding spaces where semantic relationships are preserved through relative distances rather than absolute positioning.

The prompting results provide practical insights for CLIP deployment. Template prompts ("a photo of a") improve performance by providing consistent formatting that matches CLIP's training distribution, where images were often captioned with similar phrase structures. However, overly descriptive prompts decrease performance by introducing specificity that may not align with the more general captions seen during training. This finding suggests that effective prompt engineering requires balancing informativeness with distributional alignment.

The minimal improvement from Procrustes alignment indicates that CLIP's modality gap doesn't significantly hinder performance in practice. The contrastive learning framework enables effective cross-modal understanding through structured similarity computation rather than requiring perfect distributional overlap. This insight suggests that apparent embedding separation in visualization doesn't necessarily indicate poor alignment - the relative structure within the space carries the semantic information needed for successful zero-shot transfer.

These findings have important implications for multimodal model deployment. Practitioners should focus on prompt formatting that matches training distributions rather than adding excessive descriptive detail. Additionally, the persistent modality gap shouldn't be viewed as a fundamental limitation but rather as evidence of structured representation learning that preserves cross-modal relationships through geometric arrangement rather than distributional overlap.

7. Conclusion

This empirical analysis provides practical insights into four fundamental deep learning architectures through controlled experimentation. The findings demonstrate critical architectural properties that influence real-world deployment success.

ResNet experiments confirm the necessity of residual connections for deep network training, with skip connections providing essential gradient flow that enables stable optimization. Transfer learning effectiveness validates using pre-trained features with head-only fine-tuning as the optimal efficiency-performance trade-off for most computer vision applications.

Vision Transformer analysis reveals native interpretability advantages through built-in attention mechanisms, though attention occasionally focuses on spurious visual similarities. The robustness patterns show ViTs handle random missing information well but struggle when semantically critical regions are corrupted.

GAN training dynamics illustrate how delicate equilibrium requirements make these models prone to specific failure modes. Understanding gradient vanishing, mode collapse, and discriminator overfitting enables practitioners to apply targeted interventions like label smoothing, balanced training, and regularization for stable training.

VAE posterior collapse analysis demonstrates the reconstruction-regularization trade-off fundamental to probabilistic generative models. While Beta-VAE annealing can improve latent representations, the inherent averaging tendency of reconstruction-based objectives limits sample sharpness compared to adversarial approaches.

CLIP modality gap investigation shows that apparent embedding separation doesn't prevent effective cross-modal understanding when semantic relationships are preserved through geometric structure. Prompt engineering should focus on distributional alignment rather than excessive descriptive detail.

These insights emphasize the importance of understanding architectural behavior beyond theoretical properties. Practical deployment success requires awareness of training dynamics, failure modes, and mitigation strategies specific to each architecture family.