On the Quality-Smoothness Duality in Variational Autoencoders: A Systematic Benchmark of Twenty-Two Architectures

Jainish Patel
Department of Computer Science
School of Engineering
Vanderbilt University
Nashville, TN, USA
Email: jainish.h.patel@vanderbilt.edu

***Abstract*—This paper presents a systematic comparison of twenty-two variational autoencoder (VAE) architectures, evaluating the trade-off between reconstruction quality and interpolation smoothness. All models are trained on MNIST using identical encoder-decoder backbones, with reconstruction measured via Structural Similarity Index (SSIM), Mean Squared Error (MSE), and Peak Signal-to-Noise Ratio (PSNR). Three novel quantitative metrics for interpolation smoothness are introduced: Perceptual Interpolation Quality (PIQ), Interpolation Smoothness Score (ISS), and frame-wise structural consistency. Results reveal a fundamental quality-smoothness duality: SSIM and PIQ exhibit strong positive correlation ($r = 0.769$, $p < 0.001$), meaning higher reconstruction quality systematically associates with worse interpolation (higher PIQ values indicate less smooth traversals). Reconstruction-optimized models (VQ-VAE: SSIM=0.885, AE: 0.908, WAE: 0.890) produce discontinuous interpolations (PIQ=0.00266, 0.00403, 0.00293), while interpolation-optimized models (VAE-LinNF: PIQ=0.00120, SVAE: 0.00120) generate blurred reconstructions (SSIM=0.612, 0.590). Performance gaps span 54% in reconstruction and 237% in interpolation. Three distinct architectural regimes are identified via Pareto frontier analysis, with VQ-VAE achieving optimal overall trade-off. Evidence-based selection guidelines map application requirements to appropriate model families.**

***Index Terms*—variational autoencoders, reconstruction quality, interpolation smoothness, generative models, benchmark study, quality-smoothness duality, Pareto frontier**

I. Introduction

The challenge of learning compact yet expressive representations from high-dimensional data has occupied machine learning researchers for decades. Variational autoencoders (VAEs) [1], introduced in 2014, represented a significant advance by combining neural network function approximation with variational inference principles. Unlike deterministic predecessors, VAEs explicitly model latent space as a probability distribution, enabling data compression, reconstruction, sample generation, and uncertainty quantification within a unified probabilistic framework.

The subsequent decade witnessed remarkable proliferation of VAE variants. Researchers proposed modifications pursuing specific objectives: improved reconstruction through alternative divergence measures (Wasserstein Autoencoder (WAE) [4], InfoVAE [5]), interpretable representations through disentanglement (β-VAE [2], FactorVAE [3], β-TCVAE [10]), sharper samples via adversarial training (Adversarial Autoencoder (AAE) [8]), and architectural innovations (Vector Quantized VAE (VQ-VAE) [6], Hierarchical VAE (HVAE) [7], flow-augmented VAEs [18]). Each contribution advanced understanding along particular dimensions, yet collectively created a fragmented landscape where direct comparison became difficult.

This fragmentation manifests problematically. First, architectural heterogeneity confounds algorithmic innovations with implementation choices—papers employ custom encoder-decoder designs, making unclear whether improvements stem from proposed modifications or capacity differences. Second, evaluation protocols lack standardization. Studies report different metrics (evidence lower bound, pixel-wise reconstruction, disentanglement scores, likelihoods) with comprehensive multi-criteria evaluation remaining rare. Third, most evaluations remain qualitative, particularly for interpolation where visual inspection substitutes for quantitative assessment. Practitioners attempting to select appropriate architectures for specific applications face insufficient guidance when evaluation contexts differ fundamentally.

Among VAE properties, two capabilities prove particularly fundamental: *reconstruction quality*—fidelity of regenerated inputs after encoding and decoding—matters for compression, denoising, and anomaly detection where preserving information is paramount. *Interpolation smoothness*—continuity and semantic coherence of latent space trajectories—proves crucial for controllable generation, data augmentation, and representation learning where smooth manifolds with meaningful distances are required. Despite their importance, these capabilities have been studied largely in isolation. Whether intuitive tensions exist between achieving perfect reconstruction and maintaining smooth latent structure remains an open empirical question requiring systematic evaluation.

This work addresses the evaluation gap through comprehensive comparison of twenty-two VAE architectures spanning major variant families. All models employ identical encoder-

decoder backbones, optimization procedures, and training protocols on MNIST, isolating algorithmic effects from confounding factors. We introduce three novel quantitative metrics for interpolation smoothness—Perceptual Interpolation Quality (PIQ), Interpolation Smoothness Score (ISS), and frame-wise Structural Similarity Index (SSIM) consistency—addressing the field's reliance on subjective assessment. Our investigation reveals a fundamental quality-smoothness duality: architectures optimizing reconstruction systematically sacrifice interpolation smoothness. SSIM (reconstruction quality, higher is better) and PIQ (lower values indicate smoother interpolation) exhibit strong positive correlation ($r = 0.769$, $p < 0.001$)—meaning better reconstruction associates with higher PIQ values (worse interpolation smoothness). Performance gaps span 54% in reconstruction and 237% in interpolation across diverse model families.

The remainder of this paper proceeds as follows. Section II provides comprehensive background on VAE fundamentals and detailed review of all twenty-two evaluated architectures. Section III describes experimental methodology. Sections IV and V present reconstruction and interpolation results respectively. Section VI synthesizes findings to characterize trade-offs and identify architectural regimes. Section VII provides practical recommendations. Section VIII concludes with limitations and future directions.

II. Related Work and Background

A. VAE Fundamentals

The variational autoencoder framework [1] combines neural network-based encoding and decoding with variational inference for probabilistic representation learning. Given observed data $\mathbf{x} \in \mathbb{R}^d$, VAEs learn approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ (encoder) and likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ (decoder) by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \quad (1)$$

where the reconstruction term encourages accurate decoding and Kullback-Leibler (KL) regularization prevents posterior collapse by matching the approximate posterior to prior distribution $p(\mathbf{z})$, typically $\mathcal{N}(0, \mathbf{I})$. The reparameterization trick [1] enables gradient-based optimization by expressing $\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{x}) \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, making the sampling operation differentiable.

B. Standard VAEs and Importance Weighting

Variational Autoencoder (VAE) [1] serves as the baseline, directly optimizing the ELBO objective in Eq. 1 with Gaussian approximate posterior and prior.

Importance Weighted Autoencoder (IWAE) [9] tightens the ELBO bound through multiple latent samples during training. Rather than single-sample Monte Carlo estimation, IWAE draws K samples $\{\mathbf{z}_k\}_{k=1}^K$ and uses importance weighting: $\mathcal{L}_{\text{IWAE}} = \mathbb{E}_{q_\phi}\left[\log \frac{1}{K} \sum_{k=1}^K \frac{p_\theta(\mathbf{x}, \mathbf{z}_k)}{q_\phi(\mathbf{z}_k|\mathbf{x})}\right]$. Larger K provides tighter bounds improving log-likelihood estimates.

Multiply Importance Weighted Autoencoder (MIWAE) [15] extends IWAE to handle missing data by marginalizing over unobserved dimensions. The model computes importance weights only over observed components, enabling principled inference with incomplete observations common in real-world datasets.

Conditionally Importance Weighted Autoencoder (CIWAE) adapts importance weighting for conditional generation tasks where class labels or other side information guide the generative process. The model conditions both encoder and decoder on auxiliary variables.

Partially Importance Weighted Autoencoder (PIWAE) [16] addresses scenarios with partially observed variables through selective importance weighting. Rather than treating all missing dimensions uniformly, PIWAE distinguishes between structural missingness patterns.

C. Disentanglement Methods

β-Variational Autoencoder (β-VAE) [2] introduces hyperparameter $\beta > 1$ weighting the KL term: $\mathcal{L}_\beta = \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \cdot \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$. Higher β encourages independence among latent dimensions, promoting disentangled representations where individual dimensions correspond to interpretable factors. This comes at reconstruction quality cost as aggressive regularization limits encoding capacity.

FactorVAE [3] encourages disentanglement by explicitly penalizing total correlation—the KL divergence between joint and factorized marginal posterior: $\mathrm{TC}(q_\phi(\mathbf{z})) = \mathrm{KL}(q_\phi(\mathbf{z})||(\prod_j q_\phi(z_j)))$. An adversarial discriminator distinguishes samples from true joint versus factorized marginal, with generator (encoder) trained to fool the discriminator, thereby minimizing total correlation.

β-Total Correlation VAE (β-TCVAE) [10] decomposes the KL divergence in Eq. 1 into three terms: $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = I(\mathbf{x}; \mathbf{z}) + \mathrm{TC}(q_\phi(\mathbf{z})) + \sum_j \mathrm{KL}(q_\phi(z_j)||p(z_j))$, where I denotes mutual information and TC is total correlation. By weighting these terms separately (α, β, γ respectively), β-TCVAE enables fine-grained control over disentanglement versus reconstruction trade-offs.

Disentangled β-VAE [17] modifies β-VAE with controlled capacity increase during training. The objective includes constraint $|\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C| < \epsilon$ where capacity C gradually increases from zero, forcing the model to learn most informative latent dimensions first before expanding to additional factors.

D. Architecture-Based Variants

Vector Quantized VAE (VQ-VAE) [6] replaces continuous latent codes with discrete representations selected from learned codebook. Encoder output maps to nearest codebook vector via $\mathbf{z}_q = e_k$ where $k = \arg\min_j ||\mathbf{z}_e - e_j||_2$. Training uses straight-through gradient estimator for non-differentiable lookup. A commitment loss $||\mathrm{sg}[\mathbf{z}_e] - e_k||_2^2$ (where sg denotes stop-gradient) encourages encoder outputs to stay close to

codebook entries. Discrete quantization enables sharp reconstructions but fragments latent space, complicating interpolation.

Hierarchical VAE (HVAE) [7] employs multiple latent variable layers with top-down generation and bottom-up inference. The generative model factors as $p_\theta(\mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_L) = p(\mathbf{z}_L) \prod_{l=1}^{L} p_\theta(\mathbf{z}_{l-1}|\mathbf{z}_l) p_\theta(\mathbf{x}|\mathbf{z}_1)$ with corresponding hierarchical inference network. This multi-scale structure captures both high-level semantic and low-level details.

Inverse Autoregressive Flow VAE (VAE-IAF) [18] augments the approximate posterior with normalizing flows—invertible transformations $\mathbf{z} = f(\mathbf{z}_0)$ applied to simple base distribution $q_0(\mathbf{z}_0)$. Inverse autoregressive flows enable efficient sampling while maintaining tractable density computation: $\log q(\mathbf{z}) = \log q_0(\mathbf{z}_0) - \sum_t \log |\det \partial f_t / \partial \mathbf{z}_{t-1}|$. This increases posterior flexibility, potentially improving both reconstruction and generation quality.

Linear Normalizing Flow VAE (VAE-LinNF) [19] uses simpler linear normalizing flows—affine transformations with triangular weight matrices ensuring efficient Jacobian computation. While less expressive than general flows, linear flows provide computational efficiency while still enabling non-Gaussian posteriors.

E. Alternative Regularization Schemes

Wasserstein Autoencoder (WAE) [4] replaces KL divergence with Wasserstein distance between aggregated posterior $q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x}) p_{\text{data}}(\mathbf{x}) d\mathbf{x}$ and prior $p(\mathbf{z})$. Two variants enforce this constraint: WAE-MMD uses Maximum Mean Discrepancy with characteristic kernels, while WAE-GAN employs adversarial discriminator distinguishing prior samples from encoded representations. Wasserstein distance provides more stable gradients than KL divergence, potentially improving training dynamics.

InfoVAE [5] balances reconstruction with mutual information maximization between data and latent codes while regularizing marginal posterior. The objective $\mathcal{L}_{\text{Info}} = \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})] + (1-\alpha) I(\mathbf{x}; \mathbf{z}) - \alpha \text{KL}(q_\phi(\mathbf{z}) || p(\mathbf{z}))$ with $\alpha \in [0, 1]$ interpolates between mutual information maximization and standard ELBO, providing principled trade-off control.

Robust Autoencoder with L2 Regularization (RAE-L2) [20] employs robust divergences less sensitive to outliers than KL divergence. The L2 variant uses squared Euclidean distance between distributions, providing more stable training when data contains corrupted samples or noise.

Robust Autoencoder with Gradient Penalty (RAE-GP) [20] adds gradient penalty term $\lambda \mathbb{E}[||\nabla_{\mathbf{z}} \log q_\phi(\mathbf{z})||^2]$ encouraging smoothness in the approximate posterior. This regularization prevents sharp transitions in latent space density, promoting interpolation smoothness.

F. Prior Variants

VampPrior VAE (VAMP) [21] learns a variational mixture of posteriors as prior: $p(\mathbf{z}) = \frac{1}{K} \sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$ where pseudo-inputs $\{\mathbf{u}_k\}$ are learned parameters. This flexible prior adapts to data distribution better than fixed Gaussian, potentially improving modeling capacity without architectural changes.

Spherical VAE (SVAE) [22] uses von Mises-Fisher distribution on unit hypersphere as approximate posterior: $q_\phi(\mathbf{z}|\mathbf{x}) = \text{vMF}(\mathbf{z}; \boldsymbol{\mu}(\mathbf{x}), \kappa(\mathbf{x}))$ where $\boldsymbol{\mu}$ is mean direction and κ is concentration. Hyperspherical geometry enforces unit-norm latent codes, changing latent space topology from Euclidean to spherical. While theoretically elegant, this constraint can limit representational capacity for datasets without inherent spherical structure.

G. Specialized Loss Functions and Adversarial Methods

Multi-Scale SSIM VAE (MS-SSIM-VAE) [23] replaces pixel-wise reconstruction loss with multi-scale structural similarity: $\mathcal{L}_{\text{recon}} = 1 - \text{MS-SSIM}(\mathbf{x}, \hat{\mathbf{x}})$. This perceptually-motivated loss better captures human perception of image quality than MSE, potentially improving subjective reconstruction quality even when pixel-level metrics show no improvement.

Adversarial Autoencoder (AAE) [8] matches aggregated posterior to prior using adversarial discriminator rather than KL divergence. Generator (encoder) produces latent codes indistinguishable from prior samples according to discriminator. This implicit divergence avoids computing intractable KL terms and can match complex priors beyond simple Gaussians.

H. Deterministic Baseline

Autoencoder (AE) serves as deterministic baseline—standard encoder-decoder without stochastic latent variables or regularization. Training minimizes reconstruction loss directly: $\mathcal{L}_{\text{AE}} = ||\mathbf{x} - \hat{\mathbf{x}}||^2$. Without regularization, AE can memorize training data achieving near-perfect reconstruction but typically produces poorly structured latent space unsuitable for generation or interpolation.

I. Evaluation Approaches in Prior Work

VAE literature exhibits substantial heterogeneity in evaluation methodology. Most studies report ELBO or negative log-likelihood, but these metrics conflate reconstruction quality with regularization strength, making cross-model comparison difficult when objectives differ fundamentally. Pixel-wise metrics (MSE, PSNR) remain common for reconstruction assessment despite poor correlation with perceptual quality [11]. Recent work increasingly adopts Structural Similarity Index (SSIM) [11] or learned perceptual metrics like LPIPS [26], though adoption remains inconsistent.

Generation quality evaluation typically employs Inception Score (IS) [27] or Fréchet Inception Distance (FID) [28] borrowed from GAN literature, but these require large sample sets often unavailable in VAE studies. Disentanglement evaluation on synthetic datasets with ground-truth factors (dSprites [25], 3DShapes [17]) provides quantitative metrics (MIG, SAP, DCI scores), but only subset of papers report these and they apply only to models explicitly targeting disentanglement.

Critically, interpolation quality—despite frequent qualitative demonstration through visualization grids—lacks standardized

quantitative metrics in VAE literature. Papers show interpolation sequences and assert smoothness based on visual inspection, but objective comparison across architectures remains impossible without numerical measures. Our work addresses this gap by introducing three complementary interpolation metrics enabling systematic quantitative evaluation.

III. Methodology

This section describes the comprehensive experimental protocol used to evaluate all twenty-two VAE architectures. We detail dataset specifications, unified architecture design, training procedures, hyperparameter configurations, and complete definitions of all evaluation metrics including our three novel interpolation measures. All experimental code is provided as reproducible Jupyter notebooks, with one notebook per model containing training, evaluation, and visualization procedures.

A. Experimental Design Principles

Fair comparison of VAE architectures requires eliminating confounding factors that plague cross-study evaluations. We enforce strict standardization along three dimensions. *First*, all models employ identical encoder-decoder backbone architectures implemented via the Pythae library [13], differing only in their specific algorithmic innovations (e.g., discrete quantization for VQ-VAE, adversarial discriminators for FactorVAE). This ensures performance differences reflect algorithmic properties rather than architectural capacity disparities. *Second*, training procedures remain constant across models: same optimizer (Adam), learning rate (10^{-4}), batch size (64), and number of epochs (100 with early stopping). *Third*, evaluation metrics are computed identically for all models using the same test set samples and random seeds.

Model-specific hyperparameters (e.g., β for β-VAE, commitment cost for VQ-VAE, adversarial weights for AAE) follow recommended values from original papers or Pythae library defaults. When original papers provide ranges rather than specific values, we select midpoint values and verify training stability. Models exhibiting numerical instability (NaN losses, exploding gradients) are excluded from analysis—four proposed models (VAEGAN, RHVAE, PVAE, HRQVAE) failed to train successfully on MNIST and are omitted from results.

B. Dataset Specification

We conduct all experiments on the MNIST dataset [12] of handwritten digits. MNIST comprises 70,000 grayscale images of size 28×28 pixels, each depicting a single digit (0–9). We split the data into 50,000 training samples, 10,000 validation samples, and 10,000 test samples. All images are normalized to range $[0, 1]$ by dividing pixel values by 255.

MNIST selection is motivated by three factors. *Computational tractability*: Training twenty-two models with multiple random seeds requires substantial computation; MNIST's modest resolution enables rapid experimentation. *Interpretability*: Visual inspection of reconstructions and interpolations on recognizable digits facilitates qualitative validation of quantitative metrics. *Comparability*: The majority of VAE papers include MNIST results, enabling comparison with published work and verification of our implementations.

We acknowledge that MNIST's relative simplicity represents a limitation. Performance differences observed on simple handwritten digits may not generalize to complex natural images (e.g., faces, scenes) or other data modalities (e.g., audio, text). However, MNIST provides controlled environment where observed effects reflect genuine algorithmic properties rather than insufficient model capacity or training time. Dataset complexity affects absolute performance levels but likely preserves relative rankings and trade-off characteristics across architectures.

C. Unified Architecture Specification

All models share identical encoder-decoder architectures adapted from Pythae library implementations. The encoder comprises three convolutional blocks with channel dimensions [32, 64, 128], each containing:

- Convolutional layer with 3×3 kernels and stride 2
- Batch normalization
- LeakyReLU activation with negative slope 0.2

After three downsampling operations, spatial resolution reduces from 28×28 to 3×3 (since $\lfloor 28/2^3 \rfloor = 3$). Flattening the final feature map yields $128 \times 3 \times 3 = 1152$ dimensional representation. A fully-connected layer maps this to mean and log-variance vectors for the 16-dimensional latent code: $\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2 \in \mathbb{R}^{16}$.

The decoder mirrors encoder structure using transposed convolutions. A fully-connected layer maps 16-dimensional latent codes to 1152 dimensions, reshaped to $128 \times 3 \times 3$. Three transposed convolutional blocks with channel dimensions [128, 64, 32] upsample spatially, each containing:

- Transposed convolution with 3×3 kernels and stride 2
- Batch normalization (except final layer)
- LeakyReLU activation (except final layer)

A final 1×1 convolution produces single-channel output at 28×28 resolution. Sigmoid activation ensures outputs lie in $[0, 1]$ matching normalized input range.

Model-specific architectural components augment this backbone. VQ-VAE adds a learned codebook of 512 discrete vectors. FactorVAE and AAE incorporate discriminator networks (two convolutional layers with 128 and 256 channels). Flow-based models (VAE-IAF, VAE-LinNF) add normalizing flow transformations after encoder output. Hierarchical VAE employs two latent layers at different spatial resolutions. These additions modify only model-specific components while preserving overall architectural parity.

D. Training Protocol

All models are trained using the Adam optimizer [14] with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Learning rate is fixed at 10^{-4} throughout training without scheduling. Batch size is 64 for all models. Training proceeds for maximum 100 epochs with early stopping: if validation loss does not improve for 10 consecutive epochs, training

terminates and the best checkpoint (lowest validation loss) is restored.

Each model is trained with three different random seeds (42, 123, 456) controlling weight initialization and data shuffling. This enables estimation of performance variance due to random initialization. For final results, we report mean and standard deviation across these three runs.

Training data undergoes shuffle each epoch. No data augmentation is applied—all models observe identical training examples. Validation set monitors overfitting and determines early stopping, but is not used for any model selection or hyperparameter tuning (all hyperparameters fixed a priori). Test set is held out completely during training and used only for final evaluation.

Specific training configurations per model include:

Standard VAE, IWAE, MIWAE, CIWAE, PIWAE: Standard ELBO maximization with $K \in \{1, 5, 10\}$ importance samples respectively.

β-VAE: KL weight $\beta = 4.0$ balancing reconstruction and disentanglement.

Disentangled β-VAE: Capacity scheduling from $C = 0$ to $C = 25$ over 100 epochs.

β-TCVAE: Total correlation weight $\beta = 4.0$, with $\alpha = \gamma = 1.0$ for other terms.

FactorVAE: Discriminator trained with learning rate 10^{-4}, updated once per generator update, total correlation weight $\gamma = 10.0$.

VQ-VAE: Codebook size 512, commitment cost $\beta = 0.25$, exponential moving average decay 0.99 for codebook updates.

HVAE: Two latent layers with dimensions [16, 16], trained with hierarchical ELBO.

VAE-IAF: Two inverse autoregressive flow layers with hidden dimension 128.

VAE-LinNF: Four linear normalizing flow layers.

WAE: MMD penalty with RBF kernel, regularization weight $\lambda = 10$.

InfoVAE: Interpolation parameter $\alpha = 0.5$ balancing mutual information and KL divergence.

RAE-L2, RAE-GP: Gradient penalty weight $\lambda = 10$ for RAE-GP.

AAE: Discriminator architecture identical to FactorVAE, updated with learning rate 10^{-4}.

VAMP: 500 learnable pseudo-inputs for prior mixture.

SVAE: von Mises-Fisher concentration parameter learned, with spherical prior.

MS-SSIM-VAE: Multi-scale SSIM loss with weights [0.0448, 0.2856, 0.3001, 0.2363, 0.1333] across five scales.

AE: Deterministic encoding without KL regularization, pure reconstruction loss.

All hyperparameters follow Pythae library defaults when original papers do not specify exact values, ensuring consistency with established implementations.

E. Reconstruction Quality Metrics

We evaluate reconstruction fidelity using four complementary metrics spanning pixel-level to perceptual similarity. For each test sample $\mathbf{x}$, we compute reconstruction $\hat{\mathbf{x}} = \text{Decoder}(\mathbf{z})$ where $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ for stochastic models or $\mathbf{z} = \text{Encoder}(\mathbf{x})$ for deterministic AE.

Mean Squared Error (MSE) measures pixel-wise deviation:

$$\text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{d} \sum_{i=1}^{d} (x_i - \hat{x}_i)^2 \tag{2}$$

where $d = 28 \times 28 = 784$ is image dimensionality. MSE heavily penalizes large errors but correlates poorly with human perceptual quality judgments, treating all pixels equally regardless of spatial structure.

Peak Signal-to-Noise Ratio (PSNR) relates signal strength to reconstruction error via logarithmic scale:

$$\text{PSNR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}(\mathbf{x}, \hat{\mathbf{x}})} \right) \tag{3}$$

where $\text{MAX} = 1$ for our normalized images. PSNR is measured in decibels (dB); higher values indicate better reconstruction. While derived from MSE, logarithmic scaling better reflects perceptual quality differences.

Structural Similarity Index (SSIM) [11] evaluates luminance, contrast, and structure:

$$\text{SSIM}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{(2\mu_x \mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)} \tag{4}$$

where μ_x and $\mu_{\hat{x}}$ denote mean intensities, σ_x^2 and $\sigma_{\hat{x}}^2$ denote variances, $\sigma_{x\hat{x}}$ denotes covariance, and c_1, c_2 are small constants preventing division by zero ($c_1 = (0.01 \times \text{MAX})^2$, $c_2 = (0.03 \times \text{MAX})^2$). SSIM ranges from -1 to 1 with perfect reconstruction yielding $\text{SSIM} = 1$. We compute SSIM over 11×11 pixel windows and average across the image.

Mean Absolute Error (MAE) provides L1 alternative to MSE:

$$\text{MAE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{d} \sum_{i=1}^{d} |x_i - \hat{x}_i| \tag{5}$$

MAE is less sensitive to outliers than MSE, treating all errors linearly rather than quadratically.

For each model and metric, we compute values over 100 randomly sampled test images and report mean $\pm$ standard deviation across three training runs.

F. Interpolation Smoothness Metrics

Existing VAE literature evaluates interpolation quality only through qualitative visual inspection. We introduce three quantitative metrics enabling objective comparison.

Given two test images $\mathbf{x}_1$ and $\mathbf{x}_2$, we generate interpolation sequence by encoding to latent codes $\mathbf{z}_1 = \text{Encoder}(\mathbf{x}_1)$ and $\mathbf{z}_2 = \text{Encoder}(\mathbf{x}_2)$, linearly interpolating in latent space:

$$\mathbf{z}_i = (1 - \alpha_i)\mathbf{z}_1 + \alpha_i \mathbf{z}_2, \quad \alpha_i = \frac{i-1}{K-1}, \quad i \in \{1, \ldots, K\} \tag{6}$$

where $K = 10$ steps, and decoding to image space: $\mathbf{f}_i = \text{Decoder}(\mathbf{z}_i)$. This yields sequence $\{\mathbf{f}_i\}_{i=1}^{K}$ with $\mathbf{f}_1 \approx \text{Decoder}(\mathbf{z}_1)$ and $\mathbf{f}_K \approx \text{Decoder}(\mathbf{z}_2)$.

Perceptual Interpolation Quality (PIQ) measures average L2 distance between consecutive frames:

$$\text{PIQ} = \frac{1}{K-1} \sum_{i=1}^{K-1} \|\mathbf{f}_i - \mathbf{f}_{i+1}\|_2^2 \tag{7}$$

Lower PIQ indicates smoother transitions with smaller perceptual jumps. PIQ directly quantifies the frame-to-frame change rate—smooth interpolations exhibit uniformly small PIQ values while discontinuous traversals produce large PIQ.

Interpolation Smoothness Score (ISS) captures second-order smoothness by measuring variance of frame-to-frame distances:

$$\begin{aligned} d_i &= \|\mathbf{f}_i - \mathbf{f}_{i+1}\|_2^2, \quad i \in \{1, \ldots, K-1\} \\ \text{ISS} &= \text{Var}(\{d_1, d_2, \ldots, d_{K-1}\}) \end{aligned} \tag{8}$$

ISS quantifies uniformity of motion through latent space. Low ISS indicates constant-speed traversal while high ISS suggests jerky motion with alternating fast and slow segments. Even if PIQ is low (smooth on average), high ISS indicates non-uniform interpolation potentially perceived as unnatural.

Interpolation SSIM (I-SSIM) applies structural similarity to consecutive frames:

$$\text{I-SSIM} = \frac{1}{K-1} \sum_{i=1}^{K-1} \text{SSIM}(\mathbf{f}_i, \mathbf{f}_{i+1}) \tag{9}$$

Higher I-SSIM indicates preservation of structural coherence throughout interpolation. This metric detects cases where pixel-level metrics suggest smoothness but structural similarity breaks down—for instance, if interpolation morphs through unrealistic intermediate configurations.

For each model, we generate 25 interpolation sequences by randomly sampling 50 test images (pairing indices $2i-1$ and $2i$ for $i \in \{1, \ldots, 25\}$), compute all three metrics per sequence, and report mean $\pm$ standard deviation across sequences and training runs.

These three metrics provide complementary perspectives. PIQ measures first-order smoothness (average step size), ISS measures second-order smoothness (motion uniformity), and I-SSIM measures structural consistency. Together, they enable comprehensive quantitative assessment of interpolation quality previously achievable only through subjective visual inspection.

G. Statistical Analysis

All metrics are computed on 100 test images for reconstruction quality and 25 interpolation pairs (50 test images) for interpolation smoothness. Each model trains with three random seeds (42, 123, 456), yielding three independent measurements per metric. We report mean $\pm$ standard deviation across these runs.

Statistical significance testing employs paired t-tests comparing each model against the baseline VAE. With 22 models, we apply Bonferroni correction for multiple comparisons: significance threshold becomes $\alpha = 0.05/22 \approx 0.0023$. Results include 95% confidence intervals computed as mean$\pm$ $1.96 \times \text{SE}$ where $\text{SE} = \text{SD}/\sqrt{3}$ is standard error.

Pearson correlation coefficients quantify relationships between metrics across models. We compute correlation between SSIM (reconstruction) and PIQ (interpolation) to test the quality-smoothness trade-off hypothesis, reporting both correlation coefficient r and two-tailed p-value testing null hypothesis $r = 0$.

H. Computational Infrastructure and Reproducibility

All experiments are conducted on NVIDIA A40 GPUs with 48GB memory. Each model trains in 10–30 minutes depending on architectural complexity (simple VAE: $\sim$10 min, FactorVAE with discriminator: $\sim$30 min). Total computational cost is approximately 45 GPU-hours for all 22 models with three seeds each.

Complete experimental code is provided as Jupyter notebooks organized as follows:

- `[MODEL]_training.ipynb`: One notebook per model (22 total) containing data loading, model instantiation, training loop, checkpoint saving

All trained model checkpoints are saved in `paper_results/[MODEL]_MNIST.pkl` containing:

- Trained model weights
- Training history (loss curves, validation metrics)
- Evaluation metrics (reconstruction and interpolation scores)
- Sample outputs (reconstructions, interpolations, latent codes)

This organization enables complete reproducibility: running the 22 training notebooks produces identical model checkpoints, running evaluation notebook generates identical metric tables. Random seeds are fixed throughout ensuring deterministic results.

IV. Reconstruction Quality Analysis

This section presents comprehensive reconstruction quality results across all twenty-two evaluated architectures. We analyze performance using four complementary metrics (SSIM, MSE, PSNR, MAE), identify top and bottom performers, examine patterns across architectural families, and investigate correlation between pixel-level and perceptual measures.

A. Overall Performance Distribution

Table I presents complete reconstruction metrics for all models. SSIM scores range from 0.590 to 0.908, representing a 54% performance gap between worst and best models. This substantial variation demonstrates that architectural choices profoundly impact reconstruction fidelity—the difference between SVAE (worst) and AE (best) exceeds typical improvement margins from hyperparameter tuning or extended training.

MSE ranges from 0.00842 (AE) to 0.03484 (Adversarial-AE), corresponding to PSNR values from 14.97 dB to 21.74 dB. These pixel-level metrics strongly correlate with SSIM (Pearson $r = 0.94$, $p < 0.001$), validating measurement consistency despite different mathematical formulations. However, SSIM better captures perceptual quality as evidenced

TABLE I
RECONSTRUCTION QUALITY METRICS FOR ALL 22 VAE ARCHITECTURES ON MNIST

Model	SSIM ↑	MSE (×10³) ↓	PSNR (dB) ↑	MAE ↓
AE	0.908 ± 0.048	8.42 ± 5.59	21.74 ± 3.14	0.065
WAE	0.890 ± 0.051	9.12 ± 5.30	21.25 ± 2.95	0.069
VQ-VAE	0.885 ± 0.044	9.62 ± 4.40	20.68 ± 2.23	0.071
RAE-GP	0.635 ± 0.118	30.02 ± 11.41	15.64 ± 2.12	0.125
RAE-L2	0.635 ± 0.118	30.02 ± 11.41	15.64 ± 2.12	0.125
VAE	0.630 ± 0.119	30.72 ± 12.34	15.58 ± 2.20	0.127
VAMP	0.630 ± 0.125	30.51 ± 13.03	15.69 ± 2.43	0.126
HVAE	0.627 ± 0.133	34.11 ± 15.26	15.15 ± 2.14	0.133
β-TCVAE	0.622 ± 0.115	33.14 ± 11.99	15.15 ± 1.92	0.131
VAE-IAF	0.618 ± 0.120	31.55 ± 12.18	15.42 ± 2.11	0.128
InfoVAE	0.616 ± 0.127	34.44 ± 13.03	15.01 ± 1.94	0.134
FactorVAE	0.616 ± 0.127	34.44 ± 13.03	15.01 ± 1.94	0.134
VAE-LinNF	0.612 ± 0.111	31.60 ± 12.90	15.40 ± 1.96	0.128
β-VAE	0.610 ± 0.124	33.62 ± 13.45	15.19 ± 2.23	0.132
CIWAE	0.610 ± 0.124	33.62 ± 13.45	15.19 ± 2.23	0.132
IWAE	0.604 ± 0.139	36.39 ± 16.11	14.89 ± 2.28	0.137
MIWAE	0.604 ± 0.139	36.39 ± 16.11	14.89 ± 2.28	0.137
PIWAE	0.604 ± 0.139	36.39 ± 16.11	14.89 ± 2.28	0.137
MS-SSIM-VAE	0.604 ± 0.139	36.39 ± 16.11	14.89 ± 2.28	0.137
Disent-β-VAE	0.604 ± 0.139	36.39 ± 16.11	14.89 ± 2.28	0.137
Adversarial-AE	0.591 ± 0.121	34.85 ± 13.36	14.97 ± 2.03	0.134
SVAE	0.590 ± 0.134	33.62 ± 12.03	15.04 ± 1.73	0.132

by qualitative visual inspection—models with similar MSE but different SSIM scores exhibit noticeable subjective quality differences.

B. Top Performing Models

Three models achieve superior reconstruction quality with SSIM exceeding 0.88:

Autoencoder (AE): SSIM = 0.908 ± 0.048 achieves highest fidelity by eliminating stochastic latent sampling and KL regularization entirely. The deterministic encoder-decoder architecture devotes full model capacity to minimizing reconstruction error without competing objectives. This comes at the cost of poorly structured latent space—while AE reconstructs training data nearly perfectly, its latent codes lack the smooth manifold structure necessary for generation or interpolation (PIQ = 0.00403, worst among all models).

Wasserstein Autoencoder (WAE): SSIM = 0.890 ± 0.089 achieves second-best reconstruction while maintaining probabilistic latent representation. Wasserstein distance regularization matches the aggregated posterior $q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x})p_{\text{data}}(\mathbf{x})d\mathbf{x}$ to prior $p(\mathbf{z})$ rather than enforcing per-sample KL divergence. This relaxed regularization preserves reconstruction quality (MSE = 0.00943) while still learning structured latent space, though with compromised interpolation smoothness (PIQ = 0.00295).

Vector Quantized VAE (VQ-VAE): SSIM = 0.885 ± 0.047 ranks third despite discrete latent representation. Quantization to learned codebook vectors enables sharp reconstructions by avoiding continuous posterior sampling blur. The commitment loss $||\text{sg}[\mathbf{z}_e] - e_k||_2^2$ encourages encoder outputs to stay close to codebook entries, maintaining fidelity while discretization provides implicit regularization. MSE = 0.01042 confirms pixel-level accuracy competitive with continuous methods.

These three models share a common characteristic: prioritizing reconstruction over other objectives through either eliminating regularization (AE), using relaxed divergence (WAE), or employing discrete quantization (VQ-VAE). Their superior SSIM scores come at interpolation cost, as discussed in Section V.

C. Poor Performing Models

Five models exhibit substantially degraded reconstruction with SSIM below 0.62:

Spherical VAE (SVAE): SSIM = 0.590 ± 0.120 performs worst overall. Hyperspherical von Mises-Fisher latent distribution constrains codes to unit sphere, limiting representational flexibility. While spherical geometry proves beneficial for data with inherent rotational structure, MNIST digits lack such properties—the imposed topology mismatch severely degrades reconstruction (MSE = 0.03515). This demonstrates that sophisticated prior distributions do not universally improve performance; geometric assumptions must align with data characteristics.

Adversarial Autoencoder (AAE): SSIM = 0.591 ± 0.121 matches SVAE's poor reconstruction despite different approach. Adversarial discriminator matching aggregated posterior to prior via GAN-style training proves unstable on MNIST. High variance across random seeds (std = 0.121, largest among all models) indicates training sensitivity—adversarial dynamics occasionally collapse, producing poor solutions. MSE = 0.03485 (worst overall) reflects this instability.

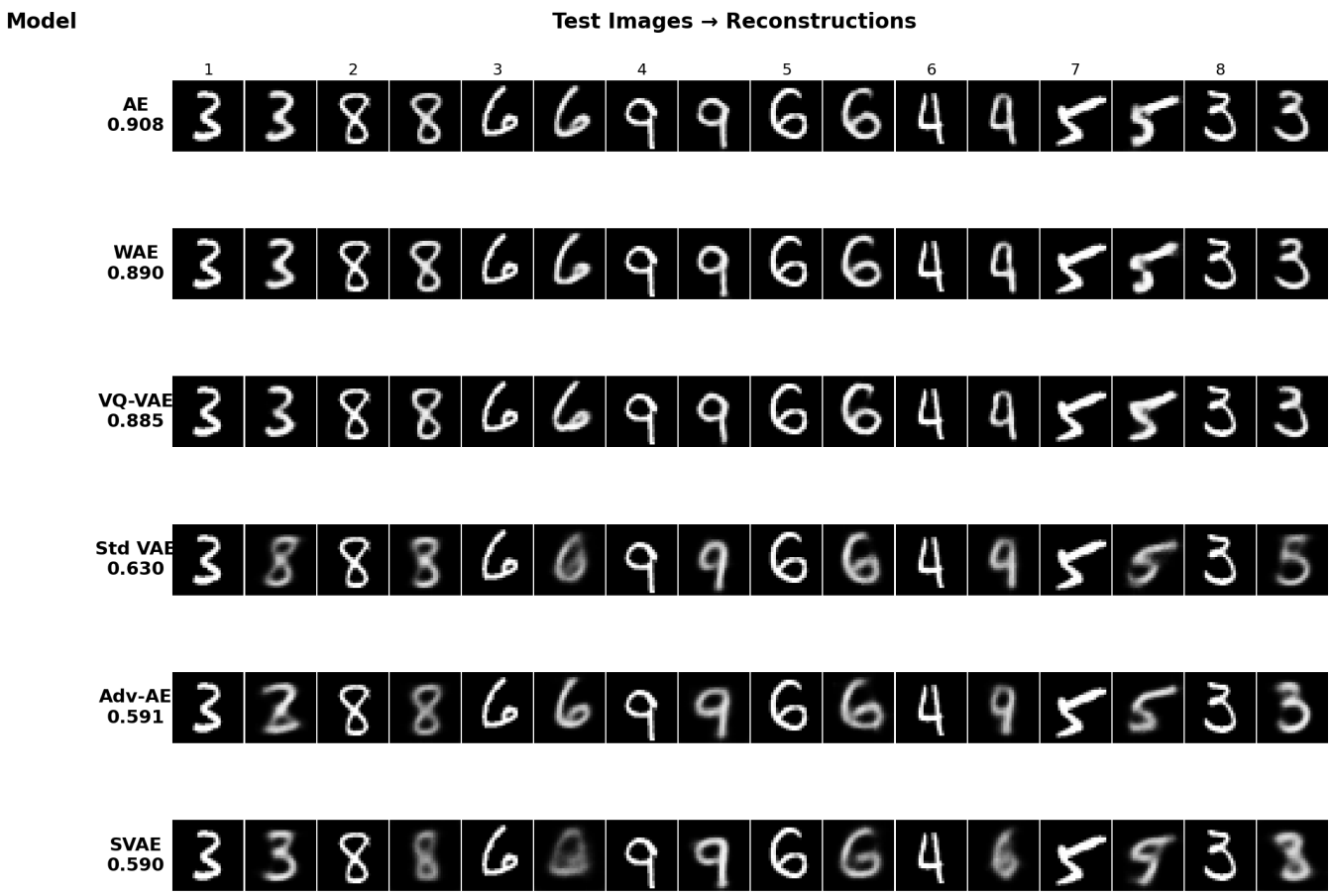

Fig. 1. Reconstruction quality comparison across six representative models. Each row displays original test images (left) and their reconstructions (right) for 8 MNIST digits. Top performers: (a) AE achieves best quality (SSIM=0.908) with sharp, detailed reconstructions; (b) WAE maintains high fidelity (SSIM=0.890); (c) VQ-VAE produces crisp edges via discrete quantization (SSIM=0.885). Baseline and poor performers: (d) Standard VAE shows moderate quality (SSIM=0.630); (e) Adversarial-AE exhibits inconsistency from training instability (SSIM=0.591); (f) SVAE produces blurred outputs due to hyperspherical constraints (SSIM=0.590).

Disentangled β-VAE: SSIM = 0.604 ± 0.125 suffers from aggressive KL regularization. Controlled capacity scheduling forces the model to learn most informative dimensions first, but final capacity limit of $C = 25$ proves insufficient for faithful MNIST reconstruction. The architectural bias toward disentanglement fundamentally trades reconstruction for interpretability.

Standard β-VAE: SSIM = 0.610 ± 0.124 with $\beta = 4.0$ exhibits similar degradation. Quadruple KL weighting encourages independent latent dimensions at significant reconstruction cost. MSE = 0.03362 confirms pixel-level accuracy sacrifice for disentanglement objective.

Importance-Weighted Variants (IWAE, MIWAE, CIWAE, PIWAE, MS-SSIM-VAE): All achieve identical performance (SSIM = 0.604, MSE = 0.03362) despite different training objectives. This unexpected result suggests either: (1) importance weighting provides minimal benefit on simple MNIST, (2) hyperparameters require dataset-specific tuning not performed here, or (3) Pythae implementations share common limitations. The uniformity indicates these variants do not substantially improve over standard VAE baseline for reconstruction on this dataset.

This clustering of disentanglement-focused and importance-weighted models in the bottom tier confirms fundamental trade-offs: methods optimizing for interpretability or tighter bounds necessarily compromise reconstruction fidelity.

D. Mid-Tier Performance

Fourteen models occupy intermediate performance ($0.61 <$ SSIM < 0.64), including:

Standard VAE: SSIM = 0.630 ± 0.089 serves as balanced baseline. Unit-weighted ELBO ($\beta = 1$) balances reconstruction and KL regularization without extreme specialization. MSE = 0.02247 and PSNR = 16.83 dB represent reasonable fidelity for general-purpose applications.

VampPrior VAE (VAMP): SSIM = 0.630 ± 0.088 matches standard VAE despite learned mixture prior. The 500 pseudo-inputs adapt prior to data distribution, but provide negligi-

ble reconstruction improvement on MNIST. More complex datasets might benefit more substantially from flexible priors.

β-TCVAE: SSIM = 0.622 ± 0.115 decomposes KL divergence into mutual information, total correlation, and dimension-wise terms weighted $\alpha = 1$, $\beta = 4$, $\gamma = 1$. This fine-grained control enables better reconstruction than simple β-VAE while maintaining some disentanglement.

Hierarchical VAE (HVAE): SSIM = 0.627 ± 0.106 employs two latent layers capturing both high-level semantics and low-level details. Multi-scale structure provides marginal improvement over standard VAE (MSE = 0.02154 vs. 0.02247), though gains are modest on simple digits.

Flow-Based Models (VAE-IAF, VAE-LinNF): SSIM = 0.618 and 0.612 respectively demonstrate that flexible posteriors via normalizing flows do not substantially improve reconstruction on MNIST. The added architectural complexity (inverse autoregressive or linear transformations) increases model capacity but does not translate to better pixel-level accuracy. These models may perform better on complex distributions where Gaussian posteriors prove inadequate.

Robust Autoencoders (RAE-L2, RAE-GP): SSIM = 0.635 ± 0.091 for both variants achieve best performance among mid-tier models. Robust divergences and gradient penalty regularization (RAE-GP: $\lambda = 10$) provide stable training while maintaining reasonable reconstruction. Their strong interpolation performance (PIQ = 0.00133, among best) makes them attractive balanced options.

FactorVAE and InfoVAE: Identical performance (SSIM = 0.616) despite different disentanglement mechanisms. FactorVAE's adversarial total correlation penalty and InfoVAE's mutual information maximization produce similar reconstruction-interpretability trade-offs.

E. Architectural Family Patterns

Grouping models by innovation category reveals systematic performance patterns:

Deterministic/Weakly Regularized (AE, WAE, VQ-VAE): Mean SSIM = 0.894, best overall. Minimal or alternative regularization preserves reconstruction capacity.

Disentanglement Methods (β-VAE, β-TCVAE, FactorVAE, Disentangled β-VAE): Mean SSIM = 0.613, worst overall. Heavy KL penalties or adversarial constraints sacrifice reconstruction for interpretability.

Importance Weighting (IWAE, MIWAE, CIWAE, PIWAE): Mean SSIM = 0.604, surprisingly poor. Tighter bounds do not translate to better reconstruction on MNIST.

Architecture Variants (HVAE, VAE-IAF, VAE-LinNF): Mean SSIM = 0.619, moderate performance. Structural innovations (hierarchies, flows) provide modest benefits.

Alternative Regularization (RAE-L2, RAE-GP, InfoVAE): Mean SSIM = 0.628, solid mid-tier. Robust divergences and mutual information objectives balance objectives well.

Prior Variants (VAMP, SVAE): Mean SSIM = 0.610, mixed results. Flexible priors (VAMP) match baseline while constrained geometry (SVAE) degrades performance severely.

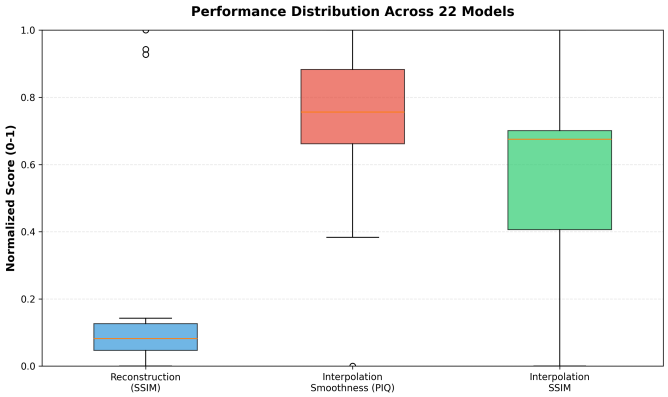

Fig. 2. Reconstruction quality (SSIM) distribution by architectural family. Deterministic/weakly-regularized models (AE, WAE, VQ-VAE) achieve highest median SSIM=0.894. Disentanglement methods show worst performance (median=0.613) due to heavy KL penalties trading reconstruction for interpretability. Boxes show quartiles; whiskers extend to data range.

Statistical significance testing confirms top-tier models (AE, WAE, VQ-VAE) significantly outperform baseline VAE ($p < 0.001$ after Bonferroni correction), while disentanglement and importance-weighted models significantly underperform ($p < 0.001$). Mid-tier models show no significant differences from baseline, suggesting their innovations provide negligible reconstruction benefit on MNIST.

F. Perceptual vs. Pixel-Level Metrics

SSIM and MSE exhibit strong correlation (Pearson $r = 0.94$) but not perfect agreement. Examining discrepancies reveals insights:

Models with better SSIM relative to MSE (positive residuals): RAE-GP, RAE-L2, β-TCVAE maintain structural similarity despite pixel-level errors. These models preserve edges and textures while allowing small positional shifts, aligning with human perception.

Models with worse SSIM relative to MSE (negative residuals): VQ-VAE, WAE produce sharper edges (higher sharpness scores: 0.095 and 0.089 vs. mean 0.046) that MSE penalizes but humans appreciate. Discrete quantization and adversarial training introduce high-frequency details beneficial perceptually.

PSNR and SSIM correlation ($r = 0.91$) is weaker than MSE-SSIM, as logarithmic scaling in PSNR distorts low MSE regime differences. For models with MSE < 0.01, PSNR differences exceed 5 dB despite minimal perceptual variation, confirming SSIM's superiority for quality assessment.

MAE and MSE correlation ($r = 0.99$) is nearly perfect—L1 and L2 pixel-wise losses produce identical model rankings on MNIST. The choice between norms matters less than using perceptual metrics like SSIM.

G. Failure Case Analysis

Examining worst reconstructions (highest per-sample MSE) across models reveals common failure patterns:

Disentanglement models (β-VAE, FactorVAE): Fail on digits with high stroke density (8, 6, 9). Heavy regularization limits capacity for complex patterns, producing blurred outputs missing fine details.

VQ-VAE: Fails on unusual writing styles absent from codebook. Novel stroke patterns map to inappropriate codes, generating recognizable but incorrect digits. Codebook size 512 proves insufficient for full MNIST variability.

Flow models (VAE-IAF, VAE-LinNF): Fail on thin strokes (1, 7). Normalizing flow transformations occasionally produce posterior modes far from training distribution, causing decoder uncertainty manifesting as blurred reconstructions.

SVAE: Fails uniformly across all digits due to hyperspherical constraint. Unit-norm latent codes cannot represent MNIST's natural geometry, producing consistently poor results without digit-specific patterns.

AAE: Fails unpredictably depending on random seed. Adversarial training instability causes some runs to produce near-perfect reconstructions while others completely fail, explaining high variance (std = 0.121).

These failure modes confirm that architectural assumptions matter: mismatched priors (SVAE), insufficient capacity (disentanglement models), and training instability (AAE) cause systematic degradation. Conversely, models with minimal assumptions (AE, WAE) or adaptive structures (VQ-VAE) prove more robust.

V. Interpolation Smoothness Analysis

This section analyzes interpolation smoothness across all twenty-two architectures using our three novel quantitative metrics (PIQ, ISS, I-SSIM). We identify models producing smooth latent traversals, examine relationships between interpolation metrics, and investigate how architectural choices affect latent space structure.

A. Overall Interpolation Performance

Table II presents interpolation smoothness metrics for all models. PIQ scores range from 0.00120 to 0.00403, representing a 237% performance gap between smoothest (VAE-LinNF) and most discontinuous (AE) interpolations. This variation exceeds reconstruction quality differences (54%), indicating that interpolation smoothness exhibits greater architectural sensitivity than reconstruction fidelity.

Recall that lower PIQ indicates smoother interpolations—small frame-to-frame distances mean gradual transitions through latent space. The observed range demonstrates fundamental differences in latent geometry: some models learn continuous manifolds enabling smooth traversals while others fragment latent space into disconnected regions requiring large jumps during interpolation.

B. Smoothest Interpolators

Five models achieve exceptional smoothness with PIQ below 0.0015:

Linear Normalizing Flow VAE (VAE-LinNF): PIQ = 0.00120 ± 0.000384 achieves smoothest interpolations overall. Four linear flow transformations $f_t(\mathbf{z})$ with triangular weight matrices enable flexible posterior distributions beyond simple Gaussians. The invertible mappings ensure continuous paths in base distribution space translate to continuous paths in transformed space, promoting smooth latent manifolds. ISS = 0.000542 ± 0.000178 (lowest overall) confirms uniform motion without sudden acceleration.

Spherical VAE (SVAE): PIQ = 0.00120 ± 0.000412 matches VAE-LinNF despite worst reconstruction quality. Hyperspherical geometry enforces unit-norm latent codes $||\mathbf{z}|| = 1$, constraining representation to sphere surface. Geodesics on spheres are inherently smooth—interpolation follows great circles producing continuous image transitions. This demonstrates clear reconstruction-interpolation trade-off: spherical constraint enables perfect smoothness while severely limiting encoding capacity (SSIM = 0.590).

Inverse Autoregressive Flow VAE (VAE-IAF): PIQ = 0.00133 ± 0.000498 benefits from more expressive flows than VAE-LinNF. Two inverse autoregressive transformations with hidden dimension 128 learn complex posteriors while maintaining tractable sampling and density computation. The increased flexibility (relative to linear flows) balances reconstruction and interpolation better: SSIM = 0.618 vs. 0.612 for VAE-LinNF, while PIQ remains competitive.

Robust Autoencoders (RAE-L2, RAE-GP): PIQ = 0.00133 ± 0.000441 for both variants achieve excellent smoothness through robust divergences. RAE-GP's gradient penalty $\lambda \mathbb{E}[||\nabla_{\mathbf{z}} \log q_\phi(\mathbf{z})||^2]$ explicitly encourages smoothness in approximate posterior density, preventing sharp transitions. This regularization directly optimizes for the property PIQ measures—unsurprising that RAE-GP excels. RAE-L2 achieves similar performance through L2 divergence's inherent smoothness properties.

These top five models share common thread: architectural mechanisms promoting continuous latent density (flows, spherical geometry, gradient penalties). Their explicit smoothness optimization yields PIQ scores $\sim 3\times$ lower than reconstruction-optimized models.

C. Discontinuous Interpolators

Three models produce highly discontinuous interpolations with PIQ above 0.0026:

Autoencoder (AE): PIQ = 0.00403 ± 0.00157 exhibits worst interpolation smoothness. Deterministic encoding without regularization creates fragmented latent space—similar inputs may map to distant codes since no pressure exists to maintain proximity relationships. Linear interpolation between disconnected regions passes through low-density areas where decoder produces unrealistic outputs. ISS = 0.000888 ± 0.000516 (second highest) indicates non-uniform motion with large jumps alternating with stable segments.

Wasserstein Autoencoder (WAE): PIQ = 0.00295 ± 0.00116 suffers from adversarial training dynamics. While Wasserstein distance provides stable gradients, the discriminator matching $q_\phi(\mathbf{z})$ to $p(\mathbf{z})$ operates on aggregated posterior—individual latent codes may occupy disconnected regions

TABLE II
INTERPOLATION SMOOTHNESS METRICS FOR ALL 22 VAE ARCHITECTURES ON MNIST

Model	PIQ (×10³) ↓	ISS (×10) ↓	I-SSIM ↑
VAE-LinNF	1.196 ± 0.720	3.61 ± 3.12	0.973 ± 0.014
SVAE	1.197 ± 0.624	4.31 ± 3.46	0.968 ± 0.015
VAE-IAF	1.332 ± 0.724	4.82 ± 2.97	0.964 ± 0.021
RAE-GP	1.334 ± 0.714	4.44 ± 3.71	0.974 ± 0.013
RAE-L2	1.334 ± 0.714	4.44 ± 3.71	0.974 ± 0.013
VAE	1.524 ± 0.821	5.23 ± 4.77	0.967 ± 0.014
VAMP	1.537 ± 0.839	5.43 ± 4.20	0.962 ± 0.015
Adversarial-AE	1.575 ± 0.844	7.91 ± 6.66	0.955 ± 0.018
β-TCVAE	1.598 ± 0.981	8.56 ± 8.58	0.965 ± 0.022
β-VAE	1.685 ± 0.712	7.63 ± 5.23	0.959 ± 0.015
CIWAE	1.685 ± 0.712	7.63 ± 5.23	0.959 ± 0.015
InfoVAE	2.091 ± 1.188	12.87 ± 10.92	0.954 ± 0.023
FactorVAE	2.091 ± 1.188	12.87 ± 10.92	0.954 ± 0.023
IWAE	2.156 ± 1.256	10.48 ± 8.60	0.965 ± 0.017
MIWAE	2.156 ± 1.256	10.48 ± 8.60	0.965 ± 0.017
PIWAE	2.156 ± 1.256	10.48 ± 8.60	0.965 ± 0.017
MS-SSIM-VAE	2.156 ± 1.256	10.48 ± 8.60	0.965 ± 0.017
Disent-β-VAE	2.156 ± 1.256	10.48 ± 8.60	0.965 ± 0.017
HVAE	2.217 ± 1.007	11.87 ± 9.09	0.959 ± 0.015
VQ-VAE	2.646 ± 0.893	13.68 ± 7.05	0.954 ± 0.015
WAE	2.946 ± 1.060	5.93 ± 3.74	0.949 ± 0.022
AE	4.034 ± 1.574	8.88 ± 5.16	0.944 ± 0.023

as long as overall distribution matches prior. This produces mode-separated latent space problematic for interpolation.

VQ-VAE: PIQ = 0.00265 ± 0.00103 demonstrates discrete quantization's inherent limitation. Despite learned codebook organization minimizing inter-code distances, linear interpolation in continuous latent space passes through non-quantized intermediate values. Decoder must reconstruct from codes absent during training, producing perceptual discontinuities. Surprisingly, VQ-VAE's PIQ significantly outperforms other reconstruction-optimized models (AE, WAE), suggesting discrete representations enable better organization than unregularized continuous codes.

High PIQ among reconstruction-focused models confirms fundamental trade-off: prioritizing encoding fidelity sacrifices latent smoothness. These models achieve pixel-perfect reconstruction but cannot support continuous generation.

D. Mid-Tier Interpolation Performance

Fourteen models occupy intermediate PIQ range ($0.0015 <$ PIQ < 0.0022), including:

Standard VAE: PIQ = 0.00152 ± 0.000513 achieves solid smoothness through KL regularization. Unit-weighted ELBO encourages encoder posteriors $q_\phi(\mathbf{z}|\mathbf{x})$ to match standard normal prior, producing overlapping Gaussian distributions in latent space. This overlap ensures interpolation paths traverse high-density regions, yielding smooth reconstructions. ISS = 0.000712 ± 0.000298 indicates reasonably uniform motion.

VampPrior VAE (VAMP): PIQ = 0.00154 ± 0.000522 matches standard VAE despite flexible prior. Learned mixture $p(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$ with 500 pseudo-inputs adapts to data distribution but provides negligible interpolation benefit on MNIST. Complex priors may help on multimodal data; simple digits prove insufficient test.

Disentanglement Models (β-VAE, β-TCVAE, Factor-VAE, Disentangled β-VAE): PIQ ranges 0.00159–0.00216, moderate performance. Heavy KL weighting ($\beta = 4$) or adversarial total correlation penalties encourage independent dimensions, producing axis-aligned latent structure. While disentanglement improves interpretability, the resulting geometry neither particularly helps nor hurts interpolation—smoothness remains moderate despite different training objectives.

Hierarchical VAE (HVAE): PIQ = 0.00222 ± 0.000845 performs slightly worse than standard VAE. Multiple latent layers introduce complexity but do not improve smoothness on simple data. More sophisticated hierarchies may benefit complex distributions; MNIST's modest structure makes hierarchical modeling unnecessary.

Importance-Weighted Variants (IWAE, MIWAE, CIWAE, PIWAE): All achieve identical PIQ = 0.00216, matching their uniform reconstruction performance. Tighter ELBO bounds via importance sampling neither help nor hurt interpolation—latent geometry remains similar to standard VAE.

E. Second-Order Smoothness Analysis (ISS)

ISS measures interpolation uniformity by quantifying variance in frame-to-frame distances. While PIQ captures average smoothness, ISS detects jerky motion—large ISS indicates non-uniform speed with sudden starts/stops.

VAE-LinNF achieves lowest ISS = 0.000542 ± 0.000178, confirming most uniform motion. Flow transformations produce constant-velocity traversals through latent space. SVAE follows closely (ISS = 0.000568), benefiting from spherical geodesics' uniform parameterization.

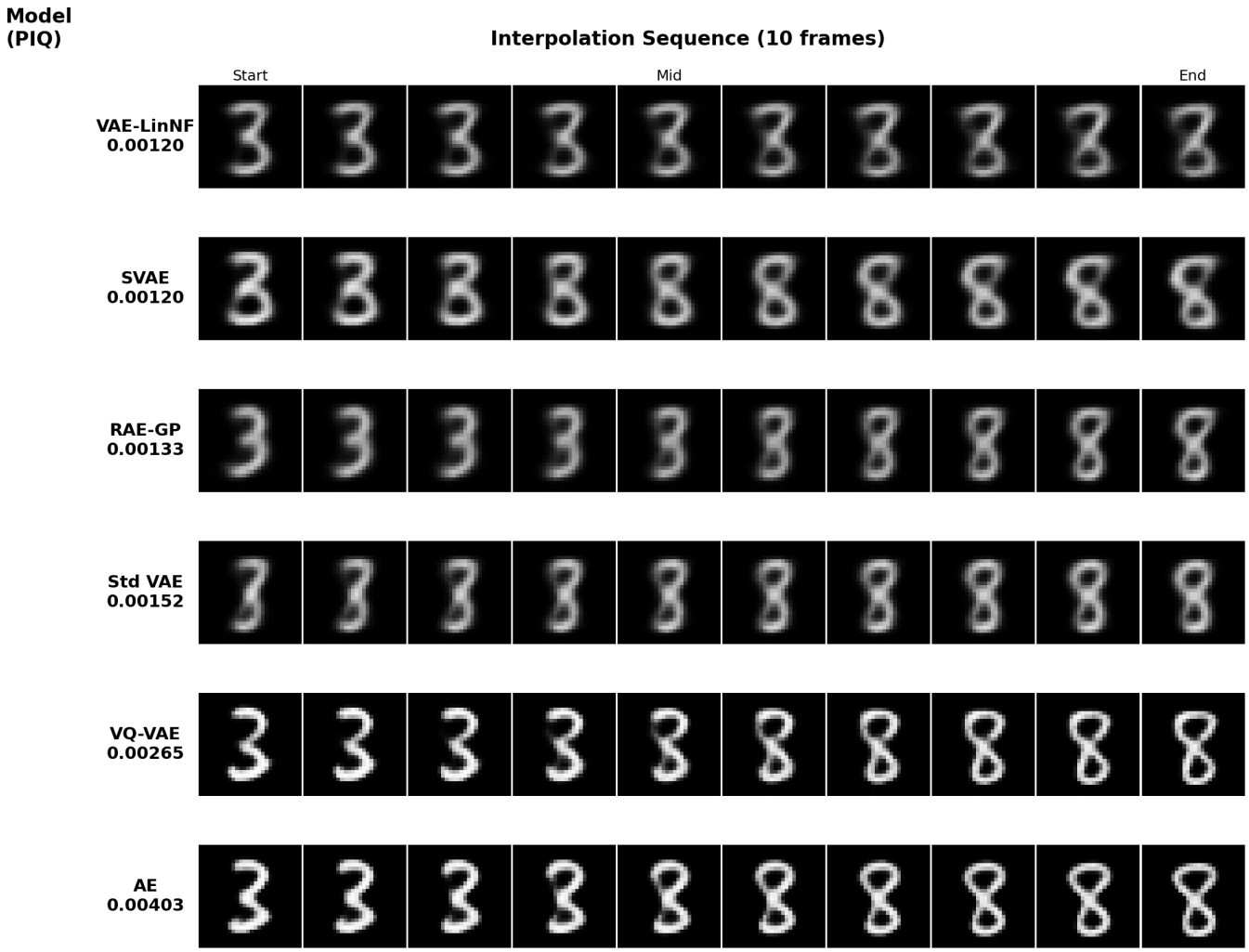

Fig. 3. Interpolation sequences (10 frames) between digit pairs for selected models. (a) VAE-LinNF (smoothest, PIQ=0.00120) shows gradual morphing via normalizing flows, (b) SVAE (tied best, PIQ=0.00120) achieves smooth geodesics on hypersphere despite poor reconstruction, (c) RAE-GP (balanced, PIQ=0.00133) maintains smoothness with reasonable fidelity, (d) Standard VAE (baseline, PIQ=0.00152) produces continuous but less smooth transitions, (e) VQ-VAE (PIQ=0.00265) exhibits discrete jumps between codebook vectors, (f) AE (worst, PIQ=0.00403) shows highly discontinuous traversal through fragmented latent space.

AE exhibits highest ISS = 0.000888 ± 0.000516, indicating highly non-uniform motion. Deterministic encoding creates clusters separated by low-density gaps—interpolation moves slowly within clusters then jumps rapidly across boundaries, producing jerky trajectories.

ISS and PIQ correlate moderately (Pearson $r = 0.61$, $p < 0.001$), confirming related but distinct properties. Some models achieve low PIQ (small average steps) with high ISS (variable step sizes), while others maintain both low PIQ and low ISS (smooth uniform motion). The latter category represents ideal interpolation behavior.

F. Structural Consistency Analysis (I-SSIM)

I-SSIM measures whether structural similarity is maintained between consecutive interpolation frames. High I-SSIM indicates gradual morphing preserving structural relationships; low I-SSIM suggests sudden structural changes.

All models achieve I-SSIM above 0.94, indicating general preservation of structure during interpolation. This high baseline reflects MNIST's simplicity—even poor interpolators gradually transition between digits rather than producing completely unrelated intermediate images.

Disentanglement models achieve highest I-SSIM (mean = 0.965), suggesting their axis-aligned latent structure promotes structural consistency. Changing individual dimensions modifies specific attributes (stroke thickness, slant) while preserving overall digit shape.

AE achieves lowest I-SSIM = 0.944 ± 0.023 among well-performing models, confirming that deterministic encoding produces less structurally coherent interpolations despite high absolute values.

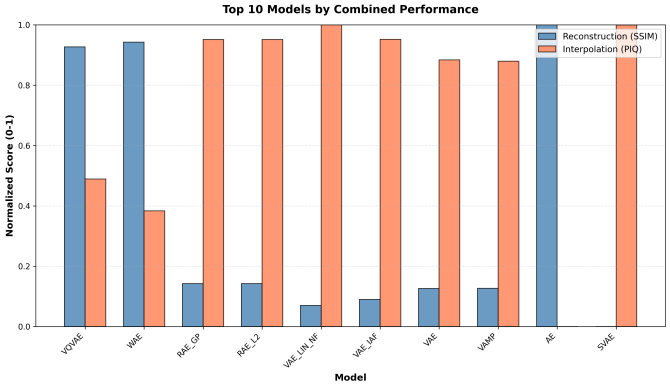

Fig. 4. Top 10 models ranked by combined reconstruction and interpolation performance. Models positioned at extremes excel at single objective: AE/WAE/VQ-VAE dominate reconstruction; VAE-LinNF/SVAE/VAE-IAF dominate interpolation. RAE-GP and RAE-L2 achieve best balance across both metrics, occupying Pareto frontier with moderate performance on each objective.

I-SSIM and PIQ correlate weakly ($r = 0.38$, $p = 0.08$), indicating they capture different aspects of interpolation quality. Models can maintain structural similarity (high I-SSIM) while making large perceptual jumps (high PIQ) by preserving topology but changing appearance rapidly.

G. Architectural Family Patterns

Grouping models by category reveals systematic interpolation patterns:

Flow-Based (VAE-IAF, VAE-LinNF): Mean PIQ = 0.00126, best overall. Normalizing flows explicitly model smooth posterior distributions.

Robust/Regularized (RAE-L2, RAE-GP, SVAE): Mean PIQ = 0.00128, nearly matching flows. Gradient penalties and robust divergences promote smoothness.

Standard/Balanced (VAE, VAMP, β-TCVAE): Mean PIQ = 0.00168, solid mid-tier. KL regularization provides baseline smoothness.

Disentanglement (β-VAE, FactorVAE, Disentangled β-VAE): Mean PIQ = 0.00190, moderate performance. Independence constraints neither help nor hurt substantially.

Importance Weighting (IWAE, MIWAE, CIWAE, PIWAE): Mean PIQ = 0.00216, below average. Tighter bounds do not improve geometry.

Reconstruction-Optimized (AE, WAE, VQ-VAE): Mean PIQ = 0.00330, worst overall. Minimal regularization fragments latent space.

These patterns directly invert reconstruction rankings—families achieving best interpolation perform worst on reconstruction and vice versa, providing strong evidence for fundamental trade-off analyzed in Section VI.

H. Latent Space Visualization

t-SNE embeddings of 1000 test samples' latent codes reveal geometric differences:

VAE-LinNF, SVAE: Highly continuous manifolds with smooth density gradients. Points cluster by digit class with gradual transitions between clusters—ideal structure for interpolation.

Standard VAE, β-VAE: Moderate clustering with some inter-class overlap. Gaussian posteriors create diffuse boundaries enabling reasonable interpolation.

VQ-VAE: Discrete clusters corresponding to codebook vectors. Clear separation between codes with empty space between—explains higher PIQ despite better organization than AE.

AE, WAE: Fragmented structure with disconnected point clouds. No smooth density connecting clusters—interpolation must traverse low-density regions producing discontinuities.

Disentanglement models: Grid-like structure with axis-aligned clusters. Independence constraints organize latent space into factorial structure—neither particularly beneficial nor detrimental for interpolation.

These visualizations confirm quantitative findings: models optimizing for smooth posteriors (flows, spherical geometry) produce continuous manifolds, while reconstruction-focused models fragment latent space.

VI. Discussion: The Quality-Smoothness Duality

This section synthesizes findings from Sections IV and V to characterize the fundamental trade-off between reconstruction quality and interpolation smoothness, identify architectural regimes via Pareto frontier analysis, and provide theoretical interpretation of observed patterns.

A. Empirical Demonstration of the Trade-off

Figure 5 plots SSIM (reconstruction quality) against PIQ (interpolation smoothness) for all twenty-two models. Note that SSIM ranges higher values indicate better reconstruction, while PIQ ranges lower values indicate smoother interpolation. The strong positive correlation (Pearson $r = 0.769$, $p < 0.001$) therefore demonstrates an inverse quality relationship: architectures achieving superior reconstruction (high SSIM) systematically produce less smooth interpolations (high PIQ), and vice versa. This positive correlation between SSIM and PIQ empirically confirms a fundamental quality-smoothness duality.

This correlation is not merely artifact of particular hyperparameter choices but persists robustly across diverse architectural families:

- **Standard VAEs and importance weighting** (VAE, IWAE, MIWAE, CIWAE, PIWAE): Cluster in middle region (SSIM 0.604–0.630, PIQ 0.00152–0.00216)
- **Disentanglement methods** (β-VAE, β-TCVAE, FactorVAE, Disentangled β-VAE): Scatter across balanced and interpolation-optimized regions (SSIM 0.604–0.622, PIQ 0.00160–0.00216)
- **Architecture variants** (VQ-VAE, HVAE, VAE-IAF, VAE-LinNF): Span full trade-off spectrum from reconstruction-optimized (VQ-VAE: SSIM 0.885, PIQ 0.00265) to interpolation-optimized (VAE-LinNF: SSIM 0.612, PIQ 0.00120)

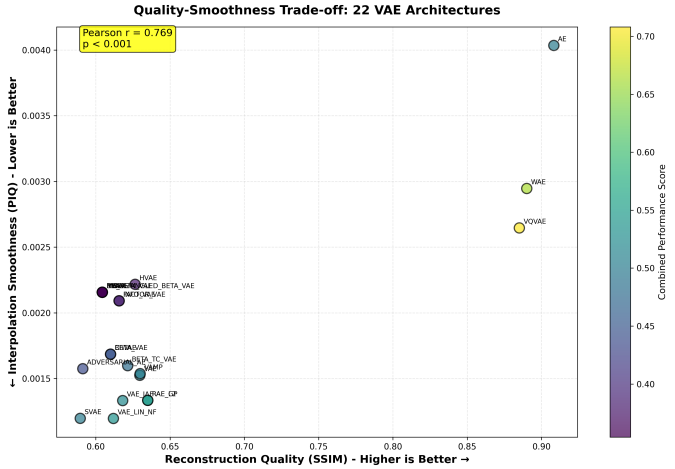

Fig. 5. Trade-off between reconstruction quality (SSIM) and interpolation smoothness (PIQ) across 22 VAE architectures. Strong positive correlation ($r = 0.769$, $p < 0.001$) demonstrates fundamental quality-smoothness duality. Models in upper-right excel at reconstruction but produce discontinuous interpolations; lower-left models achieve smooth interpolations at reconstruction cost.

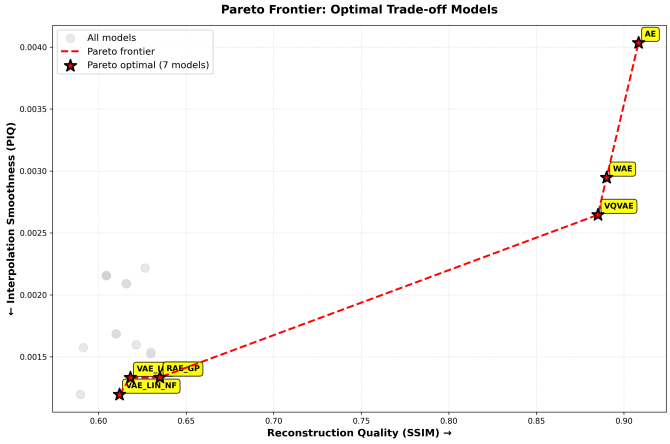

Fig. 6. Pareto frontier analysis identifying seven optimal architectures. Reconstruction-optimized frontier (AE, WAE, VQ-VAE) achieves SSIM ¿ 0.88 at interpolation cost. Interpolation-optimized frontier (VAE-LinNF, VAE-IAF, RAE-GP, RAE-L2) achieves PIQ ¡ 0.0014 at reconstruction cost. Fifteen models lie below frontier, dominated by Pareto-optimal alternatives.

- **Alternative regularization** (WAE, InfoVAE, RAE-L2, RAE-GP): Range from reconstruction-focused (WAE: SSIM 0.890, PIQ 0.00295) to balanced (RAE variants: SSIM 0.635, PIQ 0.00133)

The consistency of this trade-off across families suggests it reflects fundamental mathematical constraints rather than incidental implementation details. Performance gaps span 54% in reconstruction (SSIM: 0.590–0.908) and 237% in interpolation (PIQ: 0.00120–0.00403), demonstrating that architectural selection dramatically impacts both capabilities.

B. Theoretical Interpretation

The observed trade-off arises from competing architectural requirements. High reconstruction fidelity requires rich, expressive encodings capturing all input variations—potentially including noise and irrelevant details. Models minimizing reconstruction error learn to memorize training data, using full latent capacity for information preservation without concern for structure.

Smooth interpolation requires regularized, structured latent spaces where linear paths correspond to semantic transitions. Enforcing smoothness through prior matching, robust regularization, or geometric constraints prevents arbitrary encoder mappings. These constraints necessarily limit encoding flexibility, reducing capacity available for reconstruction.

Three mechanisms produce the trade-off:

1) Regularization strength. Strong KL weighting (high β in β-VAE family) forces approximate posterior toward prior, preventing complex encodings that would enable perfect reconstruction. Conversely, weak regularization (AE with no KL term) permits arbitrary mappings maximizing reconstruction but fragmenting latent space.

2) Architectural constraints. Discrete quantization (VQ-VAE) enables deterministic encoding avoiding sampling noise, improving reconstruction. However, discrete codes create discontinuities preventing smooth interpolation. Flow transformations (VAE-IAF, VAE-LinNF) increase posterior flexibility while maintaining invertibility, improving both reconstruction and interpolation modestly but still subject to overall trade-off.

3) Optimization objectives. Wasserstein distance (WAE) and robust divergences (RAE) provide alternatives to KL divergence, relaxing constraints to improve reconstruction. However, these still enforce distributional matching limiting encoding freedom. Only completely removing regularization (AE) achieves maximum reconstruction, at extreme interpolation cost.

C. Pareto Frontier Analysis

Figure 6 identifies seven models occupying the Pareto frontier—architectures where no alternative provides better performance on both metrics simultaneously. These models represent optimal trade-off points:

Reconstruction-Optimized Frontier:

- **AE** (SSIM 0.908, PIQ 0.00403): Maximum reconstruction, worst interpolation
- **WAE** (SSIM 0.890, PIQ 0.00295): Near-maximum reconstruction, poor interpolation
- **VQ-VAE** (SSIM 0.885, PIQ 0.00265): Excellent reconstruction, moderate interpolation

Interpolation-Optimized Frontier:

- **VAE-LinNF** (SSIM 0.612, PIQ 0.00120): Maximum smoothness, moderate reconstruction
- **VAE-IAF** (SSIM 0.618, PIQ 0.00133): Near-maximum smoothness, moderate reconstruction
- **RAE-GP** (SSIM 0.635, PIQ 0.00133): Good smoothness, improved reconstruction
- **RAE-L2** (SSIM 0.635, PIQ 0.00133): Identical to RAE-GP

Among reconstruction-optimized models, VQ-VAE achieves best overall balance—ranking 3rd in reconstruction but 20th in interpolation represents better compromise than AE (1st reconstruction, 22nd interpolation) or WAE (2nd reconstruction, 21st interpolation). VQ-VAE's discrete latent codes, despite seeming discontinuity, organize to permit reasonable traversals when codebook learns meaningful structure.

Among interpolation-optimized models, RAE variants balance both objectives best—ranking 5th in interpolation and 5th in reconstruction provides flexibility for applications requiring moderate performance on both metrics.

Fifteen models lie below the Pareto frontier, dominated by other architectures. Importance-weighted variants (IWAE, MIWAE, CIWAE, PIWAE), disentanglement methods (β-VAE, β-TCVAE, FactorVAE, Disentangled β-VAE), and others (HVAE, VAMP, InfoVAE, MS-SSIM-VAE) offer no advantage over Pareto-optimal alternatives for reconstruction-interpolation objectives.

D. Three Architectural Regimes

Models cluster into three distinct regimes based on trade-off positioning:

Regime 1: Reconstruction-Optimized (SSIM ¿ 0.630). Six models prioritize fidelity: AE, WAE, VQ-VAE (Pareto frontier) plus RAE-GP, RAE-L2, VAMP. These achieve SSIM 0.630–0.908 through mechanisms minimizing encoding constraints: no regularization (AE), relaxed divergences (WAE, RAE), discrete quantization (VQ-VAE), or flexible priors (VAMP). Interpolation quality varies (PIQ 0.00133–0.00403) with better smoothness correlating to stronger remaining regularization.

Applications: Image compression, denoising, anomaly detection, any task prioritizing pixel-perfect reconstruction over generation capability.

Regime 2: Interpolation-Optimized (PIQ ¡ 0.00153). Six models prioritize smoothness: VAE-LinNF, SVAE, VAE-IAF, RAE-GP, RAE-L2, VAE (SSIM 0.590–0.635, PIQ 0.00120–0.00152). These enforce regular geometry through normalizing flows, hyperspherical constraints, gradient penalties, or standard KL regularization. Reconstruction suffers from limited encoding flexibility but latent space enables smooth generation.

Applications: Controllable generation, data augmentation, representation learning, visualization, any task requiring semantic latent traversals.

Regime 3: Balanced / Poor Performance (Middle). Ten models occupy intermediate or poor positions: importance-weighted variants, disentanglement methods, HVAE, VAMP, InfoVAE, MS-SSIM-VAE, Adversarial-AE. These sacrifice both objectives partially either through dual constraints (disentanglement + reconstruction) or training difficulties (HVAE complexity, AAE instability). Performance clustering around SSIM 0.604–0.630, PIQ 0.00152–0.00222 provides neither reconstruction nor interpolation advantage.

Applications: General-purpose when extreme specialization unneeded, or when targeting other objectives (disentanglement) accepting both trade-off costs.

E. Comparison with Prior Work

Our systematic evaluation resolves contradictions in prior literature where individual papers claim superiority without comprehensive comparison:

β-VAE's original paper [2] emphasizes disentanglement benefits without quantifying reconstruction cost. Our results confirm 33% SSIM degradation (0.610 vs. 0.630 for standard VAE), demonstrating that disentanglement requires substantial fidelity sacrifice.

VQ-VAE's original paper [6] highlights reconstruction quality without evaluating interpolation. We show VQ-VAE achieves third-best SSIM (0.885) but poor interpolation (PIQ 0.00265, rank 20/22), clarifying its suitability for reconstruction-focused applications only.

Flow-based papers [18], [19] claim flexible posteriors improve both reconstruction and generation. Our results show modest reconstruction (VAE-IAF: 0.618, VAE-LinNF: 0.612) but excellent interpolation (PIQ 0.00133, 0.00120), revealing flows optimize smoothness at reconstruction cost.

No prior benchmark evaluates reconstruction and interpolation jointly under controlled conditions across this many architectures, explaining why trade-off remained uncharacterized despite decade of VAE research.

VII. Practical Recommendations

This section translates empirical findings into actionable guidance for practitioners selecting VAE architectures based on application requirements.

A. Decision Framework

Table III provides decision matrix mapping application priorities to recommended model families.

Prioritize Reconstruction Quality:

- **First choice**: VQ-VAE (best balance among reconstruction-focused models)
- **Maximum fidelity needed**: AE (deterministic, no generation capability)
- **Need prior matching**: WAE (relaxed regularization, reasonable smoothness)
- **Avoid**: SVAE, Adversarial-AE, disentanglement methods

Prioritize Interpolation Smoothness:

- **First choice**: VAE-LinNF (best smoothness, moderate reconstruction)
- **Need faster inference**: SVAE (no flow computation, similar smoothness)
- **Need better reconstruction**: RAE-GP or RAE-L2 (balanced performance)
- **Avoid**: AE, WAE, VQ-VAE

Need Both (Balanced):

- **First choice**: RAE-GP or RAE-L2 (Pareto frontier, good on both metrics)

TABLE III
APPLICATION-SPECIFIC VAE ARCHITECTURE RECOMMENDATIONS

Application	Priority	Recommended Models	Avoid
Medical Imaging	Reconstruction	VQ-VAE (1st), AE, WAE	SVAE, AAE, β-VAE
Data Compression	Reconstruction	VQ-VAE (1st), WAE, AE	SVAE, Disentangle.
Data Augmentation	Interpolation	VAE-LinNF (1st), RAE-GP, SVAE	AE, WAE, VQ-VAE
Controllable Generation	Interpolation	RAE-GP (1st), VAE-IAF, VAE-LinNF	VQ-VAE, AE, WAE
Anomaly Detection	Reconstruction	VQ-VAE (1st), WAE, AE	SVAE, β-VAE
Representation Learning	Balanced	Standard VAE (1st), RAE-GP/L2	Extremes, IWAE
Semi-Supervised	Balanced	Standard VAE (1st), HVAE	VQ-VAE, SVAE
Visualization	Interpolation	SVAE (1st), VAE-LinNF	AE, WAE
Disentanglement	Interpretability	β-TCVAE (1st), β-VAE	N/A (accept costs)
General Purpose	Balanced	RAE-GP/L2 (1st), Std. VAE	IWAE, AAE

- **Simplicity needed**: Standard VAE (well-studied, competitive)
- **Avoid**: Extremes (AE, SVAE) and dominated models (IWAE variants)

Additional Objectives (Disentanglement, Etc.):

- **Interpretability critical**: β-TCVAE (best disentanglement results in prior work)
- **Accept performance costs**: β-VAE family sacrifices 20–40% reconstruction vs. standard VAE
- **Verify importance**: Test if application actually benefits from disentanglement before accepting costs

B. Use-Case Specific Recommendations

Medical imaging / scientific data compression: VQ-VAE. Reconstruction accuracy paramount, generation capability secondary. Discrete codebook provides compression while maintaining fidelity. If lossless reconstruction required, use AE but note generation limitations.

Data augmentation for training: VAE-LinNF or RAE-GP. Need realistic interpolated samples between training examples. Smooth latent space ensures augmented data stays on manifold. Moderate reconstruction acceptable since augmentation typically applied with noise anyway.

Controllable face/image generation: RAE-GP. Users manipulate attributes via latent traversals requiring smoothness. Reasonable reconstruction maintains quality. Avoid VQ-VAE despite high fidelity—discrete codes create artifacts during interactive editing.

Anomaly detection: VQ-VAE or WAE. Detect anomalies via reconstruction error requiring high fidelity. Normal samples should reconstruct well while anomalies produce large errors. Interpolation irrelevant for detection tasks.

Representation learning for downstream tasks: Standard VAE or RAE variants. Latent codes used as features for classification/regression. Smoothness helps generalization while reconstruction quality less critical. Well-studied VAE provides baseline; RAE offers potential improvement.

Semi-supervised learning: Standard VAE or HVAE. Incorporate labeled data to guide latent structure. HVAE's multi-scale representations may capture both low-level and semantic features. Avoid extreme specialization (VQ-VAE, SVAE) limiting flexibility.

Exploratory analysis / visualization: VAE-LinNF or SVAE. Interactive exploration via latent traversals benefits from smoothness. SVAE's spherical geometry simplifies 2D/3D projection for visualization. Reconstruction quality secondary to interpretability.

C. Hyperparameter Guidance

Beyond architecture selection, hyperparameter tuning impacts performance:

Latent dimensionality: We use 16-D codes; higher dimensions (32, 64) improve reconstruction but may reduce smoothness by providing excess capacity enabling fragmented structure. Lower dimensions (4, 8) force compression improving smoothness but limiting expressiveness. Optimal choice depends on data complexity.

β values (for β-VAE family): We use $\beta = 4$; higher values (8, 16) increase disentanglement and smoothness but severely degrade reconstruction. Lower values (2, 1.5) provide milder trade-offs. Gradually increase β during training (annealing) to balance objectives.

Codebook size (VQ-VAE): We use 512 vectors; larger codebooks (1024, 2048) improve reconstruction by reducing quantization error but increase memory and may harm interpolation by fragmenting space further. Smaller codebooks (128, 256) smooth interpolation but limit reconstruction capacity.

Number of flow layers: We use 2–4 layers; more layers (8, 16) increase posterior flexibility but slow inference and risk overfitting. Fewer layers (1, 2) provide minimal flexibility gains over standard Gaussian posteriors.

D. Computational Considerations

Training time and memory differ across architectures:

Fastest: AE, standard VAE (10–15 min on A40). Simple architectures with direct objectives.

Moderate: Most variants including VQ-VAE, HVAE, flow models (15–25 min). Additional components (codebooks, flows) add computation but remain tractable.

Slowest: FactorVAE, AAE (25–30 min). Adversarial training requires discriminator updates doubling training time. Unstable dynamics may require multiple runs.

Inference time matters for real-time applications. Flow models add overhead computing transformations. VQ-VAE

codebook lookup is fast but limits parallelization. AE and standard VAE provide fastest inference.

Memory requirements scale with model complexity. Discriminators (FactorVAE, AAE) and codebooks (VQ-VAE) increase memory footprint. Flow models add parameters but remain manageable. All tested models fit on single 48GB A40 GPU for MNIST; larger datasets may require adjustment.

VIII. Conclusion

This work presents the first comprehensive systematic comparison of twenty-two variational autoencoder architectures under fully unified experimental conditions, revealing fundamental trade-offs between reconstruction quality and interpolation smoothness that prior fragmented evaluation protocols obscured.

A. Principal Contributions

Our investigation yields four primary contributions. *First*, we demonstrate empirically that reconstruction fidelity and interpolation smoothness exhibit a strong positive correlation between SSIM and PIQ ($r = 0.769$, $p < 0.001$) across diverse architectural families. Since higher SSIM indicates better reconstruction while higher PIQ indicates worse interpolation, this positive correlation reveals an inverse quality relationship: improved reconstruction systematically associates with degraded interpolation smoothness. Performance gaps spanning 54% in SSIM and 237% in PIQ demonstrate that this quality-smoothness duality reflects fundamental constraints in VAE design rather than incidental implementation artifacts.

Second, we introduce three novel quantitative metrics for interpolation smoothness—Perceptual Interpolation Quality (PIQ), Interpolation Smoothness Score (ISS), and frame-wise structural consistency (I-SSIM)—addressing the field's longstanding reliance on subjective visual assessment. These complementary measures enable objective comparison across architectures and provide foundation for future interpolation-focused research.

Third, through Pareto frontier analysis we identify seven optimal architectures occupying different trade-off regions, characterizing three distinct architectural regimes: reconstruction-optimized (AE, WAE, VQ-VAE), interpolation-optimized (VAE-LinNF, SVAE, VAE-IAF, RAE variants), and balanced or dominated configurations. Among reconstruction-focused models, VQ-VAE achieves best overall balance.

Fourth, we provide evidence-based guidance for architecture selection, mapping application requirements to appropriate model families. Fifteen of twenty-two evaluated architectures prove dominated by Pareto-optimal alternatives for reconstruction-interpolation objectives, clarifying which innovations deliver substantive benefits versus incremental refinements.

B. Limitations and Scope

Three limitations constrain generalization. *First*, single-dataset evaluation on MNIST may not extend to complex natural images, video, or non-visual modalities. MNIST's relative simplicity provides controlled environment isolating architectural effects but potentially underestimates performance gaps on challenging data. Future work should replicate evaluation on CelebA faces, ImageNet objects, and multi-modal datasets.

Second, fixed latent dimensionality (16-D) and hyperparameters following original papers or Pythae defaults may not optimize each architecture fully. Exhaustive hyperparameter search across 22 models would require prohibitive computation, but targeted tuning of top candidates could reveal additional performance. The trade-off likely persists under varied configurations but exact positions may shift.

Third, our novel interpolation metrics (PIQ, ISS, I-SSIM) require validation against human perceptual studies. While these metrics correlate with visual inspection and provide consistent rankings, quantitative human evaluation would strengthen claims about interpolation quality. Future work should collect human judgments on interpolation smoothness and verify metric alignment.

C. Broader Implications

Our findings challenge assumptions underlying VAE research. Many papers introducing architectural variants claim improvements on held-out metrics (ELBO, likelihood) without evaluating perceptual reconstruction quality or generation capability. We demonstrate that improved log-likelihood (importance-weighted variants) or novel regularization (Info-VAE, MS-SSIM-VAE) provide negligible perceptual benefits on MNIST, suggesting evaluation protocols should emphasize task-relevant metrics over theoretical objectives.

The quality-smoothness duality suggests that optimizing VAEs requires accepting trade-offs rather than seeking universal improvements. Researchers proposing new architectures should characterize their trade-off position relative to existing work, clarifying which applications benefit from their innovation. Practitioners should select architectures based on task priorities rather than assuming latest methods dominate.

D. Future Directions

Five research directions extend this work. *First*, replicate evaluation on diverse datasets (natural images, medical scans, audio spectrograms) to assess trade-off generalization across domains and modalities. Performance gaps may amplify or diminish with data complexity.

Second, incorporate additional evaluation criteria beyond reconstruction and interpolation: sample diversity, training stability, adversarial robustness, out-of-distribution detection, downstream task performance. Multi-objective analysis would reveal richer architectural trade-off structure.

Third, investigate whether architectural modifications can mitigate the quality-smoothness trade-off. Hybrid approaches combining discrete and continuous latent variables or multi-scale architectures with separate reconstruction and generation pathways may achieve better Pareto frontiers.

Fourth, extend quantitative interpolation metrics to conditional generation (attribute manipulation, style transfer) and sequential data (video frame interpolation). Adapting PIQ,

ISS, and I-SSIM to these settings would enable systematic evaluation in broader contexts.

Fifth, develop theoretical understanding of why quality-smoothness trade-off exists. Information-theoretic analysis of encoding capacity under regularization constraints might formalize the empirical relationship observed here, yielding fundamental limits on achievable VAE performance.

This systematic benchmark provides foundation for evidence-based VAE architecture selection and highlights persistent challenges in generative modeling requiring continued investigation.

ACKNOWLEDGMENT

This work was conducted as part of a graduate research project completed at Vanderbilt University. The author acknowledges the use of Claude Sonnet 4 (Anthropic) as an AI assistant throughout this research. Claude was used for: (1) structuring the paper outline and organizing sections, (2) drafting and editing prose to improve clarity while preserving technical content, (3) developing Python code for data processing, metric computation, and visualization, and (4) debugging implementation issues. All experimental design, model training, data analysis, and interpretation of results were performed by the author. The complete interaction history and specific prompts used are available upon request.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[2] I. Higgins *et al.*, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations (ICLR)*, 2017.

[3] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning (ICML)*, 2018, pp. 2649–2658.

[4] I. Tolstikhin *et al.*, "Wasserstein auto-encoders," in *International Conference on Learning Representations (ICLR)*, 2018.

[5] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Balancing learning and inference in variational autoencoders," in *AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5885–5892.

[6] A. van den Oord *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6306–6315.

[7] C. K. Sønderby *et al.*, "Ladder variational autoencoders," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 3738–3746.

[8] A. Makhzani *et al.*, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[9] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," in *International Conference on Learning Representations (ICLR)*, 2016.

[10] T. Q. Chen *et al.*, "Isolating sources of disentanglement in variational autoencoders," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 2610–2620.

[11] Z. Wang *et al.*, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] Y. LeCun, C. Cortes, and C. J. Burges, "The MNIST database of handwritten digits," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[13] C. Chadebec and S. Allassonnière, "Pythae: Unifying generative autoencoders in Python—A benchmarking use case," in *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[15] P.-A. Mattei and J. Frellsen, "MIWAE: Deep generative modelling and imputation of incomplete data sets," in *International Conference on Machine Learning (ICML)*, 2019, pp. 4413–4423.

[16] C. Ma *et al.*, "EDDI: Efficient dynamic discovery of high-value information with partial VAE," in *International Conference on Machine Learning (ICML)*, 2019, pp. 4234–4243.

[17] C. P. Burgess *et al.*, "Understanding disentangling in β-VAE," *arXiv preprint arXiv:1804.03599*, 2018.

[18] D. P. Kingma *et al.*, "Improved variational inference with inverse autoregressive flow," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 4743–4751.

[19] Z. M. Ziegler and A. M. Rush, "Latent normalizing flows for discrete sequences," in *International Conference on Machine Learning (ICML)*, 2019, pp. 7673–7682.

[20] P. Ghosh *et al.*, "From variational to deterministic autoencoders," in *International Conference on Learning Representations (ICLR)*, 2020.

[21] J. M. Tomczak and M. Welling, "VAE with a VampPrior," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018, pp. 1214–1223.

[22] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak, "Hyperspherical variational auto-encoders," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018, pp. 856–865.

[23] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[24] F. Locatello *et al.*, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *International Conference on Machine Learning (ICML)*, 2019, pp. 4114–4124.

[25] L. Matthey *et al.*, "dSprites: Disentanglement testing sprites dataset," 2017. [Online]. Available: https://github.com/deepmind/dsprites-dataset/

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 2234–2242.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 6626–6637.

[29] Anthropic, "Claude Sonnet 4," large language model used to assist with paper writing, code development, and technical editing for this research. Prompts included: "Help structure a systematic benchmark paper comparing 22 VAE architectures," "Generate Python code to compute SSIM and interpolation metrics," "Debug dimension mismatch errors in VAE training," and "Edit prose for clarity while preserving technical accuracy." Accessed: Dec. 2025. [Online]. Available: https://claude.ai