LISA: Latent Inference on the Sphere for A-posteriori sampling

Anonymous authors
Paper under double-blind review

Abstract

Deterministic *autoencoders* (AEs) train stably and reconstruct sharply, but unconditional generation is typically delegated to an ex-post latent density fit (e.g., MVG/GMM)[1] followed by decoding sampled codes (Ghosh et al., 2019b). We argue that this pipeline is brittle because reconstruction training does not define a well-conditioned sampling interface in latent space. We support this with two diagnostics: **(i)** deterministic decoders can be strongly sensitive to latent radius, so decoding can be poorly conditioned along an otherwise unconstrained radial degree of freedom, an effect we quantify via a controlled radial-scaling test in latent space. This implies that any ex-post sampler that perturbs radius can provoke large output changes in simple Euclidean AEs. Secondly, the **(ii)** learned latent representations can exhibit strong directional concentration, revealed by a spiky directional second-moment spectrum (low effective rank / large leading eigenvalue), making simple density models poorly calibrated for sampling. Motivated by these failure modes, we propose *Latent Inference on the Sphere for A-posteriori sampling* (LISA), a deterministic autoencoder that projects codes onto the hypersphere to remove latent scale and adds a repulsive hyperspherical kernel energy regularizer to promote directional coverage. Combined with projected MVG/GMM samplers, LISA yields a simple plug-in prior for unconditional generation. Across standard image benchmarks and structured generation tasks, this approach achieves lower *Frechet inception distance* (FID) among a broad suite of autoencoder baselines, without stochastic encoders, adversarial training, or learned priors, highlighting latent-geometry conditioning as an effective design principle for deterministic generative modelling. We also introduce *Grammar-LISA* for structured data (arithmetic expressions), demonstrating that the approach extends beyond image data.

1 Introduction

Deep generative models provide a powerful framework for modelling complex data distributions, learning useful representations, and synthesizing new samples. Popular approaches include *variational autoencoders* (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014), *generative adversarial networks* (GANs) (Goodfellow et al., 2014), autoregressive models (Van Den Oord et al., 2016; Salimans et al., 2017), and more recent diffusion-based methods (Ho et al., 2020). Although GANs and autoregressive models often achieve state-of-the-art image quality, they can be difficult to train or expensive to sample from (Larochelle & Murray, 2011; Germain et al., 2015). On the other hand, VAEs reformulate the task of learning representations for high-dimensional data as a variational inference problem. This involves optimizing an *evidence lower bound* (ELBO) objective that balances the quality of encoded samples using a stochastic encoder-decoder pair while encouraging the latent space to adhere to a fixed prior distribution. VAEs cast generative modelling as variational inference, maximizing an ELBO that trades off reconstruction quality against a KL regulariser enforcing a simple prior over latents.

A recurring practical gap is the lack of unconditional generation with deterministic autoencoders: unlike VAEs, deterministic AEs do not learn an explicit prior over the latent space. A standard workaround, popularized in deterministic/regularized AE frameworks, is ex-post density estimation: first train the autoencoder

[1] Multivariate Gaussian (MVG), Gaussian Mixture Model (GMM).

for reconstruction (plus regularization), then fit a density model (e.g., MVG/GMM) to the aggregated codes and sample through the decoder (Ghosh et al., 2019b).

However, ex-post latent sampling is not a benign plug-in step. It is geometry-sensitive and often ill-conditioned for deterministic AEs trained primarily for reconstruction. The core reason is that reconstruction training constrains only the encoder-decoder composition and does not uniquely determine the latent geometry that an ex-post sampler interacts with. In particular, deterministic AEs admit latent reparameterizations (e.g., global scaling and anisotropic linear transforms) that can preserve reconstructions after a corresponding change of decoder, while substantially altering the aggregated code distribution that ex-post density fitting attempts to approximate. As a result, the common recipe *fit a Gaussian/GMM in latent space and decode samples* is not intrinsically well-posed under a pure reconstruction objective: the fitted model and the behavior of decoded samples depend on latent parameterization choices that are unconstrained by training. In practice, this manifests as a conditioning problem for generation: simple Euclidean samplers can produce latent draws that are atypical with respect to the geometry the decoder was effectively trained on, pushing decoding away from its support and degrading unconditional sample quality. This phenomenon is already implicit in deterministic AE lines that rely on ex-post density estimation for generation, where sample quality depends critically on the induced latent structure and its compatibility with the chosen sampler (Ghosh et al., 2019b). We formalize this source of ill-posedness in the next section.

Contributions. We make the following contributions:

- We formalize how reconstruction training in deterministic AEs underdetermines the geometry required for ex-post sampling, making it ill-conditioned. (**Section** 2).
- We identify two key failure modes in ex-post sampling; **(i)** decoder radial sensitivity (**Section** 2.2), and **(ii)** latent directional concentration (**Section** 3.4), and quantify them via simple diagnostics in Fig. 1 and Table 1 respectively.
- We propose **LISA** (**Section** 3), which enforces radial invariance by decoding from $\mathbb{S}^{d-1}$ and prevents directional collapse using the uniformity regularizer from Wang & Isola (2020).
- We employ a projected Euclidean proposal (MVG/GMM) in $\mathbb{R}^d$, formally characterizing the resulting spherical distribution as a pushforward measure with an explicit density formula (**Section** 3.4).
- We show that LISA makes simple samplers competitive: it improves unconditional generation across benchmarks within the deterministic regime, without needing stochastic encoders or adversarial training (**Section** 5). We also propose *grammar-LISA* in **Section** 7 which is specifically tailored for sequence data.
- Theoretically, we show that (**Section** 6):
  - **Wasserstein's AE interpretation** (**Theorem** 1): We reframe the hyperspherical uniformity regularizer as a kernel *maximum mean discrepancy* (MMD) objective, formally linking this geometric shaping to *Wasserstein autoencoder* (WAE) (Tolstikhin et al., 2017) principles.
  - **Mismatch-controlled generation bound** (**Theorem** 2): We derive a Wasserstein-type inequality bounding expected generation error by the divergence between the projected ex-post sampler and the encoder's aggregate distribution, theoretically justifying the use of regularization to minimize sampling mismatch.

2 A Minimal Statement of Ill-posedness

Let $\mathcal{X}$ denote the data space (e.g., images) and let p_{data} be the data distribution over $\mathcal{X}$. We consider a deterministic autoencoder parameterized by an encoder $g_\phi : \mathcal{X} \to \mathbb{R}^d$ (with parameters ϕ) and a decoder $f_\theta : \mathbb{R}^d \to \mathcal{X}$ (with parameters θ), where d is the latent dimensionality. We denote the function composition by $\circ$.

We begin by formalizing the source of ill-posedness: The reconstruction objective constrains only the composition $f_\theta \circ g_\phi$ and is invariant to latent reparameterizations that can be absorbed by a corresponding change of decoder. As a result, the encoder-induced code distribution, and hence the behavior of any ex-post sampling and decoding pipeline, is not uniquely determined by reconstruction training. Hence, the decoding is ill-conditioned under ex-post density sampling.

Given a nonnegative reconstruction loss $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ (mean-squared error), the autoencoder is trained by minimizing the expected reconstruction error

$$\min_{\phi,\theta} \mathbb{E}_{x\sim p_{\text{data}}}\Big[\ell\big(x,\ f_\theta(g_\phi(x))\big)\Big]. \tag{1}$$

For a sample $x \sim p_{\text{data}}$, we denote its latent code by $z = g_\phi(x) \in \mathbb{R}^d$ and its reconstruction by $\widehat{x} = f_\theta(z) = f_\theta(g_\phi(x)) \in \mathcal{X}$. We, write the aggregated latent distribution induced by the encoder as the pushforward

$$q_\phi \ := \ (g_\phi)_\# p_{\text{data}}, \qquad \text{i.e.,} \qquad q_\phi(B) = \Pr_{x\sim p_{\text{data}}}\big[g_\phi(x) \in B\big] \ \ \forall B \subseteq \mathbb{R}^d. \tag{2}$$

A key issue is that Eq. 1 does not uniquely identify q_ϕ. Indeed, for any bijection $T : \mathbb{R}^d \to \mathbb{R}^d$ with inverse T^{-1}, define the reparameterized autoencoder

$$g_{\phi,T} \ := \ T \circ g_\phi, \qquad f_{\theta,T} \ := \ f_\theta \circ T^{-1}. \tag{3}$$

Then $f_{\theta,T}(g_{\phi,T}(x)) = f_\theta(g_\phi(x))$ for all x, so the reconstructions (and hence the objective value) are unchanged, while the aggregated latent distribution transforms as

$$q_{\phi,T} \ := \ (g_{\phi,T})_\# p_{\text{data}} \ = \ T_\# \, q_\phi. \tag{4}$$

Ex-post density estimation subsequently fits a sampler p_ψ (e.g., MVG/GMM) to the current aggregated latent distribution, for example by

$$\psi^\star(\phi) \in \arg\min_{\psi} \ D_{f/MMD}(q_\phi \,\|\, p_\psi) \quad \text{(e.g., some } f\text{-divergence or MMD on latent samples).} \tag{5}$$

Because q_ϕ itself is not identified by Eq. 1 (it can be replaced by $T_\# q_\phi$ without affecting reconstructions), any ex-post fit $p_{\psi^\star(\phi)}$ that targets q_ϕ is not uniquely determined by the reconstruction objective either. Consequently, the distribution of latents produced by the overall generation pipeline, *sample from an ex-post model and decode*, depends on latent parameterization choices that are unconstrained by training. Operationally, a simple Euclidean fit (MVG/GMM) can assign nontrivial mass to regions that are atypical under the encoder-induced codes (in the geometry consumed by the decoder), forcing the decoder to operate away from its effective training support and degrading unconditional sample quality.

Conclusion. Under a pure reconstruction objective, the latent parameterization is not uniquely identified: different encoders can yield identical reconstructions while inducing different aggregated code geometries. Since ex-post generation decodes samples from a fixed sampler family (e.g., MVG/GMM), the resulting sampling–decoding pipeline can be ill-conditioned unless the model additionally constrains the latent interface used at test time.

2.1 Our viewpoint: The latent is a sampling interface

In deterministic autoencoders, unconditional generation is typically performed by sampling a latent code and decoding it. This makes the latent space not only a representation used for reconstruction, but also the *sampling interface* through which ex-post samplers (e.g., MVG/GMM) produce codes that are subsequently decoded at test time.

The non-identifiability result in **Section** 2 implies that the reconstruction objective in Eq. 1 does not uniquely determine the latent geometry that the ex-post sampling–decoding pipeline relies on. Concretely, there exist latent reparameterizations that preserve reconstructions after a corresponding change of decoder, while transforming the aggregated latent distribution q_ϕ (and hence any induced distribution used at sampling

time) that ex-post fitting targets. A particularly important special case is global scaling: for any $c > 0$, letting $T_c(z) = cz$ yields a reparameterized encoder $g_{\phi,c} = T_c \circ g_\phi$ and a compensated decoder $f_{\theta,c} = f_\theta \circ T_c^{-1}$ that yield identical reconstructions under the reparameterized model, but induce a rescaled latent distribution $q_{\phi,c} = (T_c)_\# q_\phi$. Therefore, the plug-in recipe, *fit a Euclidean density in latent space and decode its samples* is not intrinsically well-posed under Eq. 1 unless the model constrains the latent interface consumed during decoding.

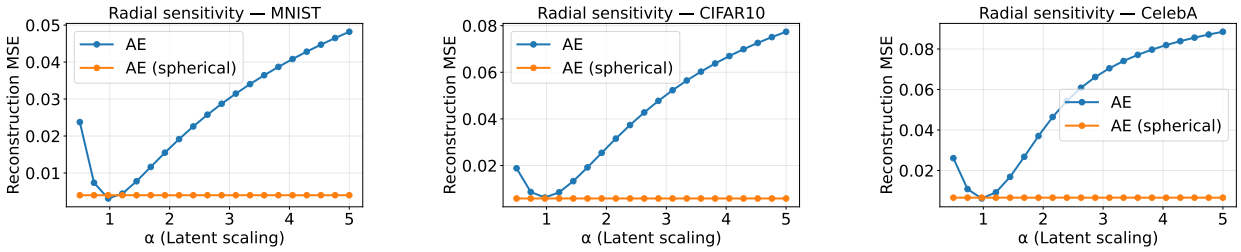
2.2 Fixing the latent interface by removing radius

Our approach does not resolve non-identifiability under arbitrary latent reparameterizations T. Instead, it targets a practically dominant subset: by projecting to $\mathbb{S}^{d-1}$ we remove the global radial (scale) degree of freedom $T_c(z) = cz, c > 0$. Given an encoded code $z = g_\phi(x) \in \mathbb{R}^d \setminus \{0\}$, we define its normalized direction by radial projection

$$\tilde{z} \;=\; \Pi(z) \;:=\; \frac{z}{\|z\|_2} \in \mathbb{S}^{d-1}, \qquad z \neq \mathbf{0}. \tag{6}$$

The resulting distribution over directions is the pushforward $\tilde{q}_\phi := \Pi_\# q_\phi$. This choice has two immediate consequences that are directly relevant for ex-post sampling: **(i)** it yields bounded support (all latents lie on $\mathbb{S}^{d-1}$), and **(ii)** it endows the latent interface with a canonical reference geometry, since the maximum-entropy distribution on $\mathbb{S}^{d-1}$ is the uniform surface measure.

Why directional geometry matches decoder behavior? Once latents are normalized, variation in latent radius is eliminated, and proximity becomes primarily angular. This matches a practical property of deterministic decoders: they can be highly sensitive to changes in latent radius, so decoding can be poorly conditioned along an otherwise unconstrained radial direction. We quantify this effect with our *decoder radial sensitivity* diagnostic in Fig. 1: scaling a latent code z by $\alpha > 0$ can substantially change reconstruction error under a standard AE decoder $f_\theta(\alpha z)$, whereas decoding after projection $f_\theta(\Pi(\alpha z))$ is invariant to α. Thus, normalization is not cosmetic; it explicitly removes an unconstrained radial degree of freedom that otherwise makes the ex-post sampling-decoding pipeline ill-conditioned.



(a) MNIST: conditioning to α (b) CIFAR10: conditioning to α (c) CelebA: conditioning to α

Figure 1: We probe decoder conditioning by scaling an encoded latent z to αz and evaluating reconstruction MSE. Euclidean AEs are highly radius-conditioned, while projecting latents to the hypersphere (AE (spherical), decoding $\Pi(\alpha z)$) yields near-invariance to α.

2.3 Shaping the spherical latent: Repulsion for directional coverage

Projecting codes onto the hypersphere fixes the latent interface by removing radius, but it does not by itself ensure that the induced directional distribution $\tilde{q}_\phi$ is well covered. In particular, the reconstruction objective in Eq. 1 can still drive many datapoints to occupy a small number of directions on $\mathbb{S}^{d-1}$, producing a concentrated $\tilde{q}_\phi$. Such directional concentration is undesirable for ex-post sampling: a simple fitted sampler will either **(i)** place mass in directions where the decoder has little support or **(ii)** overfit a few modes, both of which can degrade unconditional generation.

A hyperspherical repulsive energy. To discourage directional concentration while retaining a fully deterministic autoencoder, we propose to add a repulsive kernel energy over normalized codes. For any

$x, y \sim p_{data}$, we define the kernel

$$k(x,y) \;:=\; \exp\Big(-t\,\big\|\Pi(g_\phi(x)) - \Pi(g_\phi(y))\big\|_2^2\Big), \qquad t>0, \tag{7}$$

which depends only on the angular proximity of unit vectors (equivalently, on their dot product). We then use the uniformity functional introduced by Wang & Isola (2020) in the alignment–uniformity analysis of hyperspherical representations in self-supervised learning:

$$\mathcal{L}_{\text{unif}}(\phi) \;:=\; \log \mathbb{E}_{x,y \overset{\text{i.i.d}}{\sim} p_{data}}\Big[k(x,y)\Big]. \tag{8}$$

Minimizing $\mathcal{L}_{\text{unif}}$ penalizes large average kernel similarity among normalized codes, i.e., it discourages many points from collapsing into a small set of directions. Intuitively, this acts as a repulsive potential on $\mathbb{S}^{d-1}$ and encourages broader directional coverage of $\tilde{q}_\phi$.

3 LISA: Latent Inference on the Sphere for A-posteriori sampling

Goal. We study deterministic autoencoders for unconditional generation via ex-post latent sampling. Our central design choice is to treat the latent space as a *sampling interface*: we fix the decoder interface by removing latent scale via hyperspherical projection and regularize directional coverage, making lightweight ex-post samplers (MVG/GMM), when combined with projection, behave more robustly at test time.

3.1 Setup and notation

Let $\mathcal{X}$ denote the data space (e.g., images) and p_{data} a distribution over $\mathcal{X}$. We consider a deterministic encoder–decoder pair

$$g_\phi : \mathcal{X} \to \mathbb{R}^d, \qquad f_\theta : \mathbb{S}^{d-1} \to \mathcal{X},$$

with parameters ϕ and θ, where d is the latent dimensionality and $\mathbb{S}^{d-1} := \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ is the unit hypersphere. For $x \sim p_{\text{data}}$, define the (unnormalized) code $z := g_\phi(x) \in \mathbb{R}^d$ and its normalized direction

$$\tilde{z} \;:=\; \Pi(z) \;:=\; \frac{z}{\|z\|_2} \in \mathbb{S}^{d-1}, \qquad z \neq \mathbf{0}. \tag{9}$$

We ignore the measure-zero event $z = \mathbf{0}$; in implementation one can add $\varepsilon > 0$ in the denominator. We use a nonnegative reconstruction loss $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ (MSE loss). Let the aggregated latent distribution and its induced directional distribution be the pushforwards

$$q_\phi := (g_\phi)_\# p_{\text{data}} \quad \text{on } \mathbb{R}^d, \qquad \tilde{q}_\phi := \Pi_\# q_\phi \quad \text{on } \mathbb{S}^{d-1}. \tag{10}$$

3.2 Model definition

LISA is a deterministic autoencoder that **(i)** decodes from normalized directions $\tilde{z} \in \mathbb{S}^{d-1}$ and **(ii)** adds a hyperspherical repulsion term to discourage directional concentration. For an input $x \in \mathcal{X}$, LISA computes

$$z = g_\phi(x) \in \mathbb{R}^d, \qquad \tilde{z} = \Pi(z) \in \mathbb{S}^{d-1}, \qquad \hat{x} = f_\theta(\tilde{z}) \in \mathcal{X}.$$

3.3 Training objective: Reconstruction on the sphere + Directional repulsion

Reconstruction on the sampling interface. We train the decoder on normalized latents so that it is explicitly calibrated to the interface used at sampling time:

$$\mathcal{L}_{\text{rec}}(\phi,\theta) := \mathbb{E}_{x\sim p_{\text{data}}}\Big[\ell\big(x,\; f_\theta(\Pi(g_\phi(x)))\big)\Big]. \tag{11}$$

Directional repulsion on $\mathbb{S}^{d-1}$ (Wang & Isola, 2020). Normalization alone does not guarantee that $\tilde{q}_\phi$ covers $\mathbb{S}^{d-1}$: reconstruction pressure can still concentrate mass in a small set of directions. To discourage such

concentration, we add a repulsive kernel energy over normalized codes. Given a minibatch $x_i, \ldots, x_{|B|} \sim p_{data}$ and $\tilde{z}_i = \Pi(g_\phi(x_i))$, define the hyperspherical kernel

$$k(\tilde{z}_i, \tilde{z}_j) \;:=\; \exp\Big(-t\,\|\tilde{z}_i - \tilde{z}_j\|_2^2\Big), \qquad t > 0, \tag{12}$$

which depends only on the angular proximity of unit vectors (since $\|\tilde{z}_i - \tilde{z}_j\|_2^2 = 2 - 2\langle \tilde{z}_i, \tilde{z}_j\rangle$). We use the uniformity functional

$$\mathcal{L}_{\text{unif}}(\phi) := \log\left(\frac{1}{|B|(|B|-1)} \sum_{i \neq j} k(\tilde{z}_i, \tilde{z}_j)\right), \tag{13}$$

which penalizes large average kernel similarity among features and thereby discourages directional concentration.

Final objective. LISA minimizes a deterministic reconstruction–geometry tradeoff:

$$\min_{\phi,\theta}\; \mathcal{L}_{\text{rec}}(\phi,\theta) \;+\; \lambda\,\mathcal{L}_{\text{unif}}(\phi), \qquad \lambda \geq 0. \tag{14}$$

To verify that $\mathcal{L}_{\text{unif}}$ indeed reduces directional concentration at the distribution level (beyond a minibatch effect), we measure the anisotropy of the induced directional code distribution $\tilde{q}_\phi$ via the spectrum of its second-moment matrix. Concretely, given normalized codes $\{\tilde{z}_i\}_{i=1}^n$ from the evaluation split, we form $M := \frac{1}{n}\sum_{i=1}^n \tilde{z}_i \tilde{z}_i^\top$ and report summary statistics of its eigenvalues: the leading eigenvalue $\lambda_{\max}$, the condition number $\lambda_{\max}/\lambda_{\min}$, and the effective rank. Table 1 shows that LISA consistently yields a flatter spectrum (smaller $\lambda_{\max}$ and condition number, larger effective rank) than a simple AE with hyperspherical latent space, indicating reduced directional concentration and improved coverage on $\mathbb{S}^{d-1}$.

Importantly, we do *not* claim that $\tilde{q}_\phi$ becomes perfectly uniform; rather, λ controls how strongly we discourage extreme directional collapse while allowing reconstruction to dominate when necessary.

3.4 Sampling with lightweight Euclidean MVG/GMM + projection

Test-time pipeline. LISA performs unconditional generation by combining a lightweight Euclidean density model with the same spherical interface used during training. Concretely, we first collect unnormalized encoder codes

$$z_i := g_\phi(x_i) \in \mathbb{R}^d \quad \text{for } x_i \sim p_{\text{data}},$$

fit a Euclidean sampler p_ψ (MVG or GMM) to $\{z_i\}_{i=1}^N$ (where N is the cardinality of the dataset) in $\mathbb{R}^d$, draw samples $y \sim p_\psi$, and then decode only after radial projection:

$$y \sim p_\psi \subset \mathbb{R}^d, \qquad \tilde{y} := \Pi(y) \in \mathbb{S}^{d-1}, \qquad \hat{x} := f_\theta(\tilde{y}). \tag{15}$$

This is deliberately aligned with training, where the decoder is also trained exclusively on projected latents (Eq. 11): the decoder never receives an unnormalized vector.

What is (and is not) being modeled? A Euclidean MVG/GMM p_ψ is a density with respect to Lebesgue measure on $\mathbb{R}^d$ and is *not* a density on the manifold $\mathbb{S}^{d-1}$. Accordingly, we do *not* interpret the fitted MVG/GMM as a density on the sphere, nor do we claim any spherical maximum-likelihood optimality. Instead, we use p_ψ as a proposal distribution in $\mathbb{R}^d$ whose samples are subsequently mapped onto the sphere by Π. The resulting directional sampling distribution seen by the decoder is the pushforward measure

$$\tilde{p}_\psi \;:=\; \Pi_\# p_\psi \qquad \text{on } \mathbb{S}^{d-1}, \tag{16}$$

i.e., $\tilde{y} \sim \tilde{p}_\psi$ when $\tilde{y} = \Pi(y)$ and $y \sim p_\psi$. The pushforward $\tilde{p}_\psi$ is always a valid probability measure on $\mathbb{S}^{d-1}$ whenever p_ψ assigns zero mass to the origin (which holds for MVG/GMM under standard assumptions). The following lemma makes this precise and gives an explicit density formula.

Lemma 1 (Projected sampler induces a density on the sphere (Mardia & Jupp, 2009))**.** *Let p be a probability density on $\mathbb{R}^d$ (w.r.t. Lebesgue measure) such that $p(\{0\}) = 0$. Define $\tilde{p} := \Pi_{\#}p$ as the distribution of $\tilde{y} = y/\|y\|_2$ on $\mathbb{S}^{d-1}$. Then $\tilde{p}$ admits a density w.r.t. the surface measure σ on $\mathbb{S}^{d-1}$ given by*

$$\frac{d\tilde{p}}{d\sigma}(u) \;=\; \int_0^\infty p(ru)\, r^{d-1}\, dr, \qquad u \in \mathbb{S}^{d-1}, \tag{17}$$

and $\int_{\mathbb{S}^{d-1}} \frac{d\tilde{p}}{d\sigma}(u)\, d\sigma(u) = 1$.

Lemma 1 provides the exact sense in which sample in $\mathbb{R}^d$ then project defines a proper directional distribution. In particular, if $p_\psi = \mathcal{N}(\mu, \Sigma)$ is a single Gaussian, then $\tilde{p}_\psi$ is a projected normal distribution on $\mathbb{S}^{d-1}$; if p_ψ is a Gaussian mixture, then $\tilde{p}_\psi$ is a mixture of projected normals.

What LISA solves exactly ? This sampling choice should be read as a conditioning strategy, not as a claim of fully identified latent geometry. **Section** 2 shows that a pure reconstruction objective leaves broad latent reparameterizations unconstrained; consequently, ex-post generation can be brittle because the sampler–decoder pipeline depends on latent geometry that reconstruction alone does not pin down. LISA targets two practically dominant sources of brittleness:

1. **Radial gauge (scale) sensitivity at the decoder input.** Because LISA decodes only $\tilde{y} = \Pi(y)$, the decoder is invariant to radial rescaling of Euclidean samples:

   $$f_\theta(\Pi(\alpha y)) = f_\theta(\Pi(y)) \qquad \forall \alpha > 0,\ y \neq \mathbf{0}.$$

   Thus, any radial variability produced by a Euclidean MVG/GMM proposal is discarded before decoding. This directly addresses the dominant failure mode revealed by the decoder radial-scaling diagnostic (Fig. 1): in standard Euclidean AEs, $f_\theta(\alpha z)$ can vary sharply with α, whereas LISA's decoder interface eliminates this dependence by construction.

2. **Directional concentration.** Projection alone does not guarantee that the directional code distribution $\tilde{q}_\phi$ covers $\mathbb{S}^{d-1}$: reconstruction can still concentrate mass in a small set of directions, which can make the induced directional sampler $\tilde{p}_\psi = \Pi_{\#}p_\psi$ poorly calibrated. The repulsive uniformity regularizer from Wang & Isola (2020) (Eq. 13) explicitly discourages such concentration, and we verify this effect via directional spectrum diagnostics (Table 1).

(a) CelebA ($d = 64$)

Model	$\lambda_{\max}\downarrow$	$\frac{\lambda_{\max}}{\lambda_{\min}}\downarrow$	$\rho\uparrow$
AE	0.17	32.34	21.07
LISA	**0.08**	**13.77**	**34.96**

(b) CIFAR-10 ($d = 128$)

Model	$\lambda_{\max}\downarrow$	$\frac{\lambda_{\max}}{\lambda_{\min}}\downarrow$	$\rho\uparrow$
AE	0.10	35.18	49.21
LISA	**0.05**	**16.82**	**78.24**

(c) MNIST ($d = 16$)

Model	$\lambda_{\max}\downarrow$	$\frac{\lambda_{\max}}{\lambda_{\min}}\downarrow$	$\rho\uparrow$
AE	0.17	8.35	9.47
LISA	**0.16**	**7.62**	**10.83**

Table 1: Directional concentration diagnostics (smaller $\lambda_{\max}$ and condition number $\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)$ are better; larger effective rank (ρ) is better.

Remark. We do *not* claim that fitting a Euclidean MVG/GMM yields a maximum-likelihood estimator for a spherical likelihood. Rather, our claim is: once the decoder-side interface is fixed (projection) and the induced directional distribution is shaped (repulsion), a lightweight Euclidean MVG/GMM proposal combined with projection provides a simple, reproducible, and effective ex-post sampler for unconditional generation in deterministic autoencoders.

To our knowledge, LISA is the first method to systematically formalize and address the geometric mismatch between post-hoc samplers and the learned latent space in deterministic autoencoders, without requiring adversarial training, likelihood modelling, or multi-stage pipelines.

4 Related Work

Autoencoders, VAEs, and the Latent Space Dilemma. Autoencoders learn efficient encodings of data but do not inherently provide a way to sample new data from the latent space. A vanilla AE can reconstruct training examples well, yet if one samples a random latent vector and decodes it, the result is often nonsensical because the encoder's latent distribution is uncontrolled. The advent of Variational Autoencoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014) addressed this by introducing a prior on the latent space (typically an isotropic Gaussian) and training the encoder as an approximate posterior. This approach enables generative sampling (by drawing from the prior), but it comes at a cost: VAEs often produce blurry outputs due to the latent noise injected during training and the constraints of the ELBO objective. In practice, the VAE's objective can lead to the posterior collapse or latent variable under-utilization problem, where the decoder ignores the latent code, especially when using powerful decoders (Zhao et al., 2017; Van Den Oord et al., 2017). As a result, while VAEs ensure a sampleable latent space, they may fail to learn a truly meaningful latent representation or sharp generative results in complex image domains. Some remedies, such as restricting the capacity of the conditional distribution (Chen et al., 2016), have been proposed, but they require careful manual tuning and are often tailored to specific tasks.

Enhancing Latent Utilization in VAEs. A number of researchers have proposed modifications to make VAEs use their latent dimensions more effectively. InfoVAE (Zhao et al., 2019) is one such approach that adds a term to maximize the mutual information between the latent variables and the observations. By doing so, InfoVAE mitigates the VAE's tendency to ignore the latent code, leading to richer encodings and improved generative quality. Similarly, the InfoMax-VAE (Rezaabad & Vishwanath, 2020) explicitly optimizes a mutual information objective on top of the VAE framework. This method was shown to learn more meaningful representations and even outperform earlier variants like InfoVAE and β-VAE (Higgins et al., 2017) in terms of representation quality and downstream generation. (Notably, β-VAE introduced a weighted KL-divergence to restrict information capacity, which aids disentanglement at the expense of reconstruction fidelity. InfoMax-VAE instead aims to boost information without sacrificing as much fidelity.) This information-focused VAEs implicitly tackle the same issue as our work, encouraging latent codes to carry useful data variation, albeit by modifying the training loss rather than post-processing the latents.

Aligning the Latent Distribution with the Prior. Numerous studies have explored diagnosing the VAE framework by analyzing its objective function components (Hoffman & Johnson, 2016; Zhao et al., 2017; Alemi et al., 2018) and suggesting enhancements to overcome optimization challenges (Rezende et al., 2014; Dai & Wipf, 2019). Deterministic denoising (Vincent et al., 2008) and *contractive autoencoders* (CAEs) (Rifai et al., 2011) have been investigated for their ability to capture smooth data manifolds. Various heuristic methods have attempted to imbue these models with generative abilities, including through MCMC schemes (Rifai et al., 2012; Bengio et al., 2013). However, these methods present challenges related to convergence diagnosis, require extensive tuning (Cowles & Carlin, 1996), and have not been effectively scaled beyond MNIST, which led to their replacement by VAEs.

Addressing the mismatch between the aggregated posterior and the prior often involves making the prior more expressive (Kingma & Welling, 2013; Bauer & Mnih, 2019), which alters the VAE objective and demands significant additional computational resources. Using another VAE to estimate the latent space of the original VAE (Dai & Wipf, 2019) reintroduces many optimization challenges, and it is much more resource-intensive compared to simply fitting a straightforward $q_\psi(z)$ post-training. Another line of work focuses on ensuring the latent space of a VAE/AE matches the chosen prior distribution, thereby avoiding regions of the prior that decode to invalid samples (often referred to as the prior hole problem) (Aneja et al., 2021). *Adversarial autoencoders* (AAEs) (Makhzani et al., 2015) add a discriminator to the deterministic encoder-decoder pair, producing sharper samples but at the cost of increased computational complexity and potential instability due to adversarial training. *Wasserstein autoencoders* (WAEs) (Tolstikhin et al., 2017) generalize AAEs by framing autoencoding as an *optimal transport* (OT) problem. Both stochastic and deterministic models can be trained by minimizing a relaxed OT cost function, using either an adversarial loss term or the maximum mean discrepancy score between prior and $q_\phi(z)$ as a regularizer instead of KL-divergence. The most effective VAE architectures for images and audio to date are variations of the *vector quantized variational autoencoders* (VQ-VAE) (Van Den Oord et al., 2017; Razavi et al., 2019). Despite

the name, VQ-VAEs are neither stochastic nor variational; rather, they are deterministic autoencoders. However, VQ-VAEs require complex discrete autoregressive density estimators and a training loss that is non-differentiable due to the quantization of the latent variable.

Two Stage and Post-hoc Latent modelling. Rather than baking a complex prior into a single training stage, an elegant strategy is to learn the latent distribution in a second stage. The Two-Stage VAE proposed by Dai & Wipf (2019) is a seminal example. This two-step approach was shown to dramatically improve sample fidelity. The success of this method underscores that learning an accurate latent prior (even post hoc) can close the gap in generative performance without abandoning the VAE framework.

Regularized Autoencoders (RAE) Ghosh et al. (2019b) drop the VAE prior during training: they learn a *deterministic* AE and use explicit regularizers (e.g., on latent covariance or the decoder Jacobian) to keep the representation well-behaved. After training, they fit an ex-post density (MVG/GMM) to the empirical latent codes and sample from this surrogate prior. Despite its simplicity, this two-step approach can yield strong non-adversarial generation quality: with suitable regularization and a fitted GMM prior, RAEs substantially improved FID on MNIST, CIFAR-10, and CelebA, narrowing the gap between deterministic AEs and VAEs. These results show that a latent space can be made sampleable post hoc by modelling the aggregate posterior, without likelihood-based training. We follow the same principle, but shape the latent distribution using a different mechanism.

Inspired by ex-post density estimation from Ghosh et al. (2019b), a growing body of work has shifted toward deterministic autoencoders for generation, driven by their training stability, lack of variational noise, and sharper reconstructions. To facilitate generation, methods such as *implicit rank minimizing autoencoder* (IRMAE) (Jing et al., 2020), GMM-*deterministic autoencoder* (GMM-DAE) (Saseendran et al., 2021), and *low rank autoencoder* (LoRAE) (Mazumder et al., 2024) employ regularization to structure the latent space. Specifically, GMM-DAE enforces compatibility with multimodal distributions in the latent space, while IRMAE and LoRAE induce a rank-deficient latent geometry. These approaches improve sample quality without generative likelihoods, but they leave the latent sampling process ill-specified. In particular, the common ex-post sampling strategy, fit a Euclidean density (e.g., GMM) to encoder latents and decodes from it, can yield drastically different results depending on latent geometry, a fact rarely acknowledged in prior literature. Further, Chadebec & Allassonnière (2022) proposed a geometric approach to sampling in VAEs by leveraging the Riemannian structure induced by the encoder-decoder (for benchmarking, we call this method *geometric*-VAE (Geo-VAE)). Instead of using a fixed Gaussian prior, they sample uniformly with respect to the latent manifold's volume, leading to significantly improved generation. Their results show that careful post-hoc sampling, without altering model parameters, can close much of the performance gap between basic VAEs and more complex generative models.

Autoencoders with Spherical Latent Spaces. Previous works like $\mathcal{S}$-VAE (Davidson et al., 2018) project latent codes onto a hypersphere to leverage uniform priors, but they do so during training and are primarily designed for stochastic encoders. These approaches often struggle to scale to high-dimensional settings or generalize beyond the imposed geometry. Another work which closely aligns with LISA is *spherical autoencoder* (SAE) (Zhao et al., 2020), which normalizes codes onto a hypersphere to obtain a non-Euclidean latent representation, but they do not explicitly **(i)** diagnose the sampler/decoder failure mode driven by radial perturbations, or **(ii)** add an explicit mechanism to prevent directional collapse that makes simple samplers poorly calibrated. None address the core identifiability issue introduced by reconstruction losses: many latent distributions yield identical reconstructions but induce different aggregate statistics, making post-hoc sampling ill-posed. Prior work sidesteps this by assuming a fixed prior (e.g., standard Gaussian), but this assumption is broken in deterministic settings.

LISA: Normalizing latent geometry for well-posed sampling. LISA builds upon the practical success of post-hoc density modelling in deterministic autoencoders but addresses a crucial and overlooked challenge: the ill-posedness of ex-post latent sampling due to unconstrained latent geometry. In contrast, LISA retains a standard deterministic AE training objective and introduces a post-training normalization step to explicitly enforce directional isotropy of latent codes. By projecting latent vectors onto the sphere and aligning the sampling distribution with this induced geometry, LISA removes an unconstrained radial degree of freedom that renders typical Gaussian or GMM-based samplers fragile.

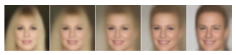
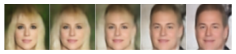
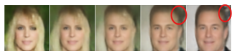
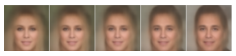
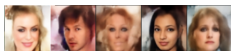
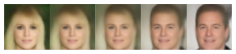
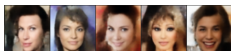
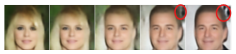
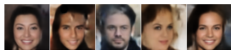
	Interpolation	Random Samples	Reconstruction
Original			
VAE			
AE			
AE (Sphere)			
WAE			
RAE-L_2			
RAE-GP			
IRMAE			
Geo-VAE			
LoRAE			
LISA			

Table 2: Qualitative evaluation of sample quality for VAE, AE, famous deterministic autoencoders (WAE and RAE), and LISA on CelebA reveals that LISA stands out by generating cleaner samples and interpolating smoothly in the latent space. LISA also reconstructs sharper images despite incorporating a regularization term. More qualitative results on different datasets are given in Appendix C. A quantitative evaluation of reconstruction FID and interpolation FID is given in Appendix E.

5 Experiments

LISA is motivated by a practical failure mode in deterministic AEs used for unconditional generation: generation is performed by fitting an ex-post Euclidean density (e.g., MVG/GMM) in latent space and decoding samples, but the reconstruction objective does not itself enforce a sampling-compatible latent interface. Our experiments, therefore, ask: *does explicitly fixing and shaping the latent interface make simple post-hoc samplers reliable enough to yield competitive unconditional generation?*

We trained LISA on the MNIST (Deng, 2012), CIFAR-10 (Krizhevsky et al., 2009), and CelebA (Liu et al., 2015) datasets and compared it with a wide range of generative and deterministic autoencoders. The generative family includes VAE (Kingma & Welling, 2013), β-VAE (Higgins et al., 2017), Info-VAE (Zhao et al., 2019), $\mathcal{S}$-VAE (Davidson et al., 2018), VAE-Vamp Prior (Tomczak & Welling, 2018), VQ-VAE (Van Den Oord et al., 2017; Razavi et al., 2019), HVAE (Caterini et al., 2018), VAE-LinNF (Rezende & Mohamed, 2015), VAE-IAF (Kingma et al., 2016), CV-VAE (Ballé et al., 2016; Ghosh et al., 2019a) and 2s-VAE (Dai & Wipf, 2019). The deterministic family includes a simple AE, WAE (Tolstikhin et al., 2017), RAE (Ghosh et al., 2019b), IRMAE (Jing et al., 2020), GMM-DAE (Saseendran et al., 2021), Geo-VAE (Chadebec & Allassonnière, 2022), and LoRAE (Mazumder et al., 2024). We set the latent space dimensions to 16, 128, and 64 for MNIST, CIFAR-10, and CelebA, respectively, for LISA[2], trained it for 100 epochs for each dataset.

[2]Details about model architecture, hyperparameters, and FID evaluations are deferred to **Appendix** A. Sensitivity analysis of hyperparameters λ and t is also reported in **Appendix** D

Post-hoc sampler used at inference. At inference time we fit a Euclidean density to unnormalized encoder training set outputs, then *project only the sampled latents* before decoding:

$$x \sim p_{\text{data}} \xrightarrow{g_\phi} z = g_\phi(x) \in \mathbb{R}^d, \quad \text{fit } p_\psi \text{ (MVG or GMM) on } \{z_i\}_{i=1}^N, \quad y \sim p_\psi, \quad \hat{x} = f_\theta(\Pi(y)). \tag{18}$$

This is mathematically well-defined: p_ψ is a proposal on $\mathbb{R}^d$ and Π pushes it forward to a valid distribution on $\mathbb{S}^{d-1}$. Crucially, the decoder always receives unit-norm inputs both in training and in sampling, so *radial mismatch cannot occur at decode time*. This is the specific ill-conditioning LISA targets: instability caused by unconstrained radius in ex-post sampling.

For unconditional generation we evaluate the *Frechet inception distance* (FID) (Heusel et al., 2017; Parmar et al., 2022) score (lower is better), using the standard feature-based protocol used in prior RAE, IRMAE, and LoRAE. We fit the ex-post MVG/GMM prior p_ψ using latent codes $\{g_\phi(x)\}_{x \in \mathcal{D}_{\text{train}}}$ from the training split only. The reported FIDs are computed against the test split, but the fitted prior does not use test examples. We also report representation quality via downstream classification accuracy from latent embeddings in Table 4, as a sanity check that shaping the sampling interface does not collapse useful structure.

Note: We do *not* position LISA as competing with diffusion models or heavy learned priors. Our goal is to study how far one can go with a simple deterministic encoder/decoder plus a lightweight ex-post sampler.

Discussion: Table 2 shows reconstructions, random samples, and latent interpolations. Random samples from LISA exhibit coherent global structure and fewer broken decodes than vanilla AEs, consistent with the claim that generation improves when the decoder is only queried on the interface it was trained on (unit-norm directions), rather than off-support radii. Interpolations remain smooth, indicating that the representation maintains continuity under the spherical interface. For quantitative results, we follow prior work Ghosh et al. (2019b); Jing et al. (2020); Chadebec & Allassonnière (2022); Mazumder et al. (2024), we adopt the same evaluation protocol and criteria to ensure a fair and directly comparable assessment across models. These findings are supported by the data showcased in Table 3, where our model achieves the best FID score among the 17 benchmarked autoencoders (standing second for one category), highlighting the effectiveness of our proposed model.

	MNIST		CIFAR-10		CelebA	
	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM
VAE	34.75	42.30	281.42	283.86	61.69	65.14
β-VAE	19.34	12.65	113.18	90.3	53.89	50.65
VAE-Vamp Prior	55.76	–	141.12	–	60.16	–
VQ-VAE	<u>18.21</u>	–	<u>87.61</u>	–	49.18	–
InfoVAE	36.71	29.44	251.59	238.43	66.65	63.90
$\mathcal{S}$-VAE	53.89	–	–	–	–	–
HVAE	26.96	27.76	274.16	261.14	59.48	51.48
VAE-LinNF	29.31	28.46	240.32	247.09	56.54	53.36
VAE-IAF	27.53	27.00	236.08	235.43	55.43	53.61
CV-VAE	33.79	17.87	94.75	86.64	48.87	49.30
2s-VAE	18.81	–	109.77	–	49.70	–
AE	29.79	16.78	91.67	85.65	69.43	62.35
AE (Sphere)	33.77	15.74	95.57	89.67	<u>47.71</u>	42.21
WAE	22.86	12.22	111.44	<u>84.77</u>	50.59	43.12
RAE-GP	24.61	13.01	92.90	84.97	48.45	**38.70**
RAE-L_2	25.92	12.45	92.03	84.79	51.84	45.10
IRAME	26.58	22.31	–	–	58.98	48.96
GMM-DAE	–	12.82	–	–	–	44.79
Geo-VAE	–	**8.53**	–	93.53	–	48.71
LoRAE	19.50	11.69	–	–	56.29	46.43
LISA **(Ours)**	**17.97**	<u>10.32</u>	**89.01**	**83.16**	**45.30**	<u>39.36</u>

Table 3: Evaluation of all models by FID (lower is better, best models in **bold**, second best in <u>underline</u>). We evaluate each model by $\mathcal{N}$: random samples generated according to the prior distribution $p(z)$ (isotropic Gaussian for VAE, WAE, β-VAE, InfoVAE, another VAE for 2s-VAE) or by fitting an MVG (for the remaining models); GMM: random samples generated by fitting a mixture of 10 Gaussians in the latent space. LISA is competitive with or outperforms previous models throughout the evaluation. Experiments on high-resolution datasets like ImageNet-1k and LSUN-Church are given in Appendix F.

	$\mathcal{S}$-VAE	RAE-L_2	RAE-GP	WAE-MMD	IRMAE	Geo-VAE	LoRAE	VAE	AE	AE (Sphere)	LISA
MNIST	86.97	92.35	90.55	88.21	98.10	93.22	**98.60**	90.41	91.48	95.91	98.38
CIFAR-10	–	54.02	54.05	53.84	–	38.21	–	51.34	53.41	41.70	**55.83**

Table 4: Classification accuracy (↑) for MNIST and CIFAR-10 using different autoencoders.

6 Theoretical Analysis

In this section, we establish two complementary results that together justify LISA's modelling choices. First, we give a geometry-aware mismatch bound showing that, when the decoder consumes unit-norm latents, the expected sampling-time risk under the projected ex-post sampler is controlled by a Wasserstein mismatch between the sampler's directional distribution and the encoder's directional aggregate. This result directly motivates treating the hypersphere as a sampling interface and shaping the induced distribution of directions. Second, we connect LISA's repulsive uniformity regularizer to a WAE-style MMD penalty toward the uniform prior on the sphere. Because the target prior is the uniform (Haar) measure, the MMD expansion collapses into a pure repulsive energy plus a constant, explaining why LISA can use a simple pairwise repulsion term while still aligning with the WAE perspective. All proofs are deferred to the **Appendix** B.

6.1 A geometry-aware bound for ex-post sampling on the spherical interface

The *encoder-induced directional aggregate* is the pushforward measure

$$\tilde{q}_\phi := (\Pi \circ g_\phi)_\# \, p_{\text{data}} \quad \text{on } \mathbb{S}^{d-1}. \tag{19}$$

At test time we fit a Euclidean density p_ψ in $\mathbb{R}^d$ (e.g., MVG/GMM) to *train* codes $\{z_i = g_\phi(x_i)\}$, sample $Y \sim p_\psi$, and decode only after projection, $\hat{x} = f_\theta(\Pi(Y))$.

Although p_ψ lives in Euclidean space, the *actual* latent distribution seen by the decoder is the *projected sampler* on the sphere:

$$\tilde{p}_\psi := \Pi_\# p_\psi, \qquad \text{i.e.,} \qquad U = \Pi(Y) \sim \tilde{p}_\psi \;\text{ when }\; Y \sim p_\psi. \tag{20}$$

Therefore, the sampling problem for LISA is intrinsically a comparison of *directional* measures $\tilde{p}_\psi$ and $\tilde{q}_\phi$ on $\mathbb{S}^{d-1}$, regardless of how p_ψ is fit in $\mathbb{R}^d$.

Metric structure on the sphere. We use the geodesic distance $d_{\mathbb{S}}(u, v) = \arccos(\langle u, v\rangle) \in [0, \pi]$ and the induced 1-Wasserstein distance

$$W_1^{\mathbb{S}}(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} d_{\mathbb{S}}(u, v)\, d\gamma(u, v), \tag{21}$$

where $\Gamma(\mu, \nu)$ denotes couplings with marginals μ and ν.

A sampling-time risk functional. Let $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ be the reconstruction loss used for training (e.g., MSE loss). We define a *sampling-time risk* for any measure μ on $\mathbb{S}^{d-1}$ by

$$\mathcal{R}(\mu) := \mathbb{E}_{U\sim\mu}\big[r(U)\big], \qquad r(u) := \mathbb{E}_{x\sim p_{\text{data}}}\big[\ell\big(x, f_\theta(u)\big)\big]. \tag{22}$$

This is *not* FID; it is a clean analytic proxy that isolates the dependence of sampling quality on **(i)** the decoder geometry on $\mathbb{S}^{d-1}$ and **(ii)** the mismatch between $\tilde{p}_\psi$ and $\tilde{q}_\phi$.

Assumption 1 (Interface Lipschitzness)**.** *The decoder $f_\theta : \mathbb{S}^{d-1} \to \mathcal{X}$ is L_f-Lipschitz from $(\mathbb{S}^{d-1}, d_{\mathbb{S}})$ to $(\mathcal{X}, \|\cdot\|_2)$, i.e $\|f_\theta(u) - f_\theta(v)\|_2 \leq L_f\, d_{\mathbb{S}}(u, v) \quad \forall u, v \in \mathbb{S}^{d-1}$.*

Assumption 2 (Loss Lipschitzness in the reconstruction argument)**.** *For every fixed $x \in \mathcal{X}$, the map $\hat{x} \mapsto \ell(x, \hat{x})$ is L_ℓ-Lipschitz, i.e $|\ell(x, \hat{x}) - \ell(x, \hat{x}')| \leq L_\ell \|\hat{x} - \hat{x}'\|_2 \quad \forall x, \hat{x}, \hat{x}' \in \mathcal{X}$.*

Theorem 1 (Mismatch controls sampling-time risk on the spherical interface)**.** *Under Assumptions 1–2, the function $r : \mathbb{S}^{d-1} \to \mathbb{R}$ defined in Eq. 22 is $(L_\ell L_f)$-Lipschitz w.r.t. $d_{\mathbb{S}}$. Consequently, for any probability measures μ, ν on $\mathbb{S}^{d-1}$,*

$$\big|\mathcal{R}(\mu) - \mathcal{R}(\nu)\big| \;\le\; L_\ell L_f \, W_1^{\mathbb{S}}(\mu, \nu). \tag{23}$$

In particular, for the LISA directional aggregate $\tilde{q}_\phi$ and the projected ex-post sampler $\tilde{p}_\psi$,

$$\mathcal{R}(\tilde{p}_\psi) \;\le\; \mathcal{R}(\tilde{q}_\phi) \;+\; L_\ell L_f \, W_1^{\mathbb{S}}(\tilde{p}_\psi, \tilde{q}_\phi). \tag{24}$$

Theorem 1 isolates a single, sampling-relevant quantity: once the decoder consumes directions on $\mathbb{S}^{d-1}$, the only way the ex-post sampler affects $\mathcal{R}(\tilde{p}_\psi)$ (under Lipschitz regularity) is through the mismatch $W_1^{\mathbb{S}}(\tilde{p}_\psi, \tilde{q}_\phi)$. This provides a precise target aligned with LISA's design: **(i)** projection fixes the radial gauge at the decoder input (sampling-induced radius variation is removed before decoding), and **(ii)** the repulsion regularizer shapes $\tilde{q}_\phi$ to be less directionally concentrated, making it a more stable target for lightweight projected Euclidean families (e.g., projected-normal mixtures induced by MVG/GMM). We emphasize that the theorem is stated for the proxy risk Eq. 22, not for FID; the role of the theory is to formalize *why* LISA focuses on directional mismatch on the sampling interface.

6.2 Uniformity regularization as an MMD-to-uniform penalty: The WAE connection

Kernel on the sphere and MMD. Let U denote the uniform (Haar) probability measure on $\mathbb{S}^{d-1}$. Consider the Gaussian (RBF) kernel restricted to the sphere:

$$k(u, v) := \exp\Big(-\frac{1}{2\sigma^2}\|u - v\|_2^2 \Big), \qquad u, v \in \mathbb{S}^{d-1},\ \sigma > 0. \tag{25}$$

Since $\|u - v\|_2^2 = 2 - 2\langle u, v\rangle$ on the sphere, this kernel depends only on the angular proximity. For a probability measure Q on $\mathbb{S}^{d-1}$, define $\mathrm{MMD}_k^2(U, Q)$ in the standard way.

Theorem 2 (MMD to the uniform sphere reduces to a $Q \times Q$ kernel energy)**.** *Let Q be any probability measure on $\mathbb{S}^{d-1}$. Then*

$$\mathrm{MMD}_k^2(U, Q) = \mathbb{E}_{z, z' \sim Q}\big[k(Z, Z')\big] - C_{d,\sigma}, \tag{26}$$

with $C_{d,\sigma} = \exp(-\sigma^{-2})\,\Gamma\!\left(\frac{d}{2}\right)(2\sigma^2)^{\frac{d-2}{2}}\, I_{\frac{d-2}{2}}\!\big(\sigma^{-2}\big)$, where $I_\nu(\cdot)$ is the modified Bessel function of the first kind. Furthermore, for i.i.d. samples $z_1, \ldots, z_N \sim Q$, the unbiased estimator satisfies

$$\widehat{\mathrm{MMD}}_k^2(U, Q) = \frac{1}{N(N-1)} \sum_{i \ne j} k(z_i, z_j) - C_{d,\sigma}, \qquad \mathbb{E}\big[\widehat{\mathrm{MMD}}_k^2(U, Q)\big] = \mathrm{MMD}_k^2(U, Q). \tag{27}$$

Implication for LISA (explicit WAE-style interpretation). Recall that LISA induces the directional aggregate $\tilde{q}_\phi$ in Eq. 19. If we choose $t = \frac{1}{2\sigma^2}$ in the LISA kernel (Eq. 12), then the empirical uniformity term

$$\mathcal{L}_{\mathrm{unif}}(\phi) = \log\Big(\frac{1}{N(N-1)} \sum_{i \ne j} k(\tilde{z}_i, \tilde{z}_j) \Big)$$

is a monotone transform of $\mathbb{E}_{\tilde{q}_\phi \times \tilde{q}_\phi}[k]$. By Theorem 2, minimizing $\mathbb{E}_{\tilde{q}_\phi \times \tilde{q}_\phi}[k]$ is equivalent (same minimizers over ϕ) to minimizing $\mathrm{MMD}_k^2(U, \tilde{q}_\phi)$, since they differ only by the additive constant $C_{d,\sigma}$. Therefore, LISA can be viewed as a deterministic *WAE-style* objective on the spherical interface: it couples reconstruction on $\mathbb{S}^{d-1}$ with an MMD penalty that pulls the induced directional aggregate toward the uniform prior on the sphere, but with a simplification specific to the uniform prior: the $U \times U$ and $U \times Q$ terms are constants and disappear, leaving a purely repulsive $Q \times Q$ energy. This interpretation clarifies why the regularizer directly targets directional coverage rather than Euclidean moment matching in $\mathbb{R}^d$.

7 Grammar-LISA: Modelling Structured Inputs

Given 50,000 univariate expressions generated from a formal grammar (Kusner et al., 2017), the goal is to find an expression that best fits target data by minimizing $\log(1 + \text{MSE})$, where MSE is computed between the expression outputs and target points (targets follow (Kusner et al., 2017)).

We follow the architecture and experimental protocol of Kusner et al. (2017) and add our proposed training losses. Baselines include *Grammar VAE* (GVAE) (Kusner et al., 2017), *Charecter VAE* (CVAE) (Gómez-Bombarelli et al., 2018), *Grammar constant variance VAE* (GCVVAE) (Ghosh et al., 2019b), *Grammar RAE* (GRAE) (Ghosh et al., 2019b), and *Grammar GMM-DAE* (G-GMM-DAE) (Saseendran et al., 2021). We consider the validity of the new samples generated by the models. A well-structured latent space should yield valid samples following the defined grammar/rules of the used dataset. Our model achieves better validation and average scores as shown in Table 5. All results are averaged over 5 *Bayesian optimization* (BO) trials.

Problem	Method	Frac. Valid (↑)	Avg. Score (↓)
Expression	GVAE	0.99 ± 0.01	3.47 ± 0.24
	CVAE	0.86 ± 0.06	4.75 ± 0.25
	GRAE	**1.00 ± 0.00**	3.22 ± 0.03
	G-GMM-DAE	**1.00 ± 0.00**	3.32 ± 0.04
	G-LISA	**1.00 ± 0.00**	**3.16 ± 0.21**

Table 5: % of valid samples of equations generated by different methods and their average mean score.

8 Conclusion

In this paper, we address the issue of representation collapse in VAEs. Our findings reveal that the conventional objective of VAEs is inadequate for learning general and useful representations. We also observe that complex generative networks tend to inhibit the model from acquiring constructive representations. To tackle these challenges, we propose a novel information-based, simple, and deterministic autoencoder that maximizes the information retained in the latent representations from the observations. Through extensive experiments, we demonstrate that our model outperforms or matches the performance of other state-of-the-art autoencoders on both image datasets and challenging structured datasets.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International conference on machine learning*, pp. 159–168. PMLR, 2018.

Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. {NCP}-{vae}: Variational autoencoders with noise contrastive priors, 2021. URL https://openreview.net/forum?id=c1xAGI3nYST.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 66–75. PMLR, 2019.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. *Advances in neural information processing systems*, 26, 2013.

Anthony L Caterini, Arnaud Doucet, and Dino Sejdinovic. Hamiltonian variational auto-encoder. *Advances in Neural Information Processing Systems*, 31, 2018.

Clément Chadebec and Stéphanie Allassonnière. A geometric perspective on variational autoencoders. *Advances in neural information processing systems*, 35:19618–19630, 2022.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American statistical Association*, 91(434):883–904, 1996.

Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pp. 881–889. PMLR, 2015.

Partha Ghosh, Arpan Losalka, and Michael J Black. Resisting adversarial attacks using gaussian mixture variational autoencoders. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 541–548, 2019a.

Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019b.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.

Li Jing, Jure Zbontar, et al. Implicit rank-minimizing autoencoder. *Advances in Neural Information Processing Systems*, 33:14736–14746, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International conference on machine learning*, pp. 1945–1954. PMLR, 2017.

Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 29–37. JMLR Workshop and Conference Proceedings, 2011.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

Alokendu Mazumder, Tirthajit Baruah, Bhartendu Kumar, Rishab Sharma, Vishwajeet Pattanaik, and Punit Rathore. Learning low-rank latent spaces with simple deterministic autoencoder: Theoretical and empirical insights. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2851–2860, 2024.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Ali Lotfi Rezaabad and Sriram Vishwanath. Learning representations by maximizing mutual information in variational autoencoders. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2729–2734. IEEE, 2020.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on international conference on machine learning*, pp. 833–840, 2011.

Salah Rifai, Yoshua Bengio, Yann Dauphin, and Pascal Vincent. A generative process for sampling contractive auto-encoders. *arXiv preprint arXiv:1206.6434*, 2012.

Surojit Saha, Sarang Joshi, and Ross Whitaker. Ard-vae: A statistical formulation to find the relevant latent dimensions of variational autoencoders. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 889–898. IEEE, 2025.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

Amrutha Saseendran, Kathrin Skubch, Stefan Falkner, and Margret Keuper. Shape your space: A gaussian mixture regularization approach to deterministic autoencoders. *Advances in Neural Information Processing Systems*, 34:7319–7332, 2021.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Jakub Tomczak and Max Welling. Vae with a vampprior. In *International conference on artificial intelligence and statistics*, pp. 1214–1223. PMLR, 2018.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Deli Zhao, Jiapeng Zhu, and Bo Zhang. Latent variables on spheres for sampling and inference, 2020. URL https://openreview.net/forum?id=rJx2slSKDS.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 5885–5892, 2019.

A Model Architecture and Hyperparameter Details

Datasets	MNIST	CIFAR-10	CelebA
Encoder	$x \in \mathbb{R}^{32\times32\times1}$ $\to Conv_{32} \to ReLU$ $\to Conv_{64} \to ReLU$ $\to Conv_{128} \to ReLU$ $\to Conv_{256} \to ReLU$ Flatten 1024 $\to FC_{128} \to z \in \mathbb{R}^{16}$	$x \in \mathbb{R}^{64\times64\times3}$ $\to Conv_{128} \to BN + ReLU$ $\to Conv_{256} \to BN + ReLU$ $\to Conv_{512} \to BN + ReLU$ $\to Conv_{1024} \to BN + ReLU$ Flatten 16,384 $\to FC_{4096} \to z \in \mathbb{R}^{128}$	$x \in \mathbb{R}^{64\times64\times3}$ $\to Conv_{128} \to BN + ReLU$ $\to Conv_{256} \to BN + ReLU$ $\to Conv_{512} \to BN + ReLU$ $\to Conv_{1024} \to BN + ReLU$ Flatten 16,384 $\to FC_{4096} \to z \in \mathbb{R}^{64}$
Decoder	$z \in \mathbb{R}^{16}$ FC_{8096} Reshape to $8 \times 8 \times 128$ $\to ConvT_{64} \to ReLU$ $\to ConvT_{32} \to ReLU$ $\to ConvT_{3} \to Tanh$ $\hat{x} \in \mathbb{R}^{32\times32\times1}$	$z \in \mathbb{R}^{128}$ $FC_{65536} \to ReLU$ Reshape to $8 \times 8 \times 1024$ $\to ConvT_{1024} \to BN + ReLU$ $\to ConvT_{512} \to BN + ReLU$ $\to ConvT_{256} \to BN + ReLU$ $\to ConvT_{3} \to Sigmoid$ $\hat{x} \in \mathbb{R}^{64\times64\times3}$	$z \in \mathbb{R}^{64}$ $FC_{65536} \to ReLU$ Reshape to $8 \times 8 \times 1024$ $\to ConvT_{1024} \to BN + ReLU$ $\to ConvT_{512} \to BN + ReLU$ $\to ConvT_{256} \to BN + ReLU$ $\to ConvT_{128} \to BN + ReLU$ $\to ConvT_{3} \to Sigmoid$ $\hat{x} \in \mathbb{R}^{64\times64\times3}$

Table 6: Architecture of encoder and decoder for MNIST, Cifar and CelebA dataset. In our model (EIMAE), the latent vectors are normalized before feeding into the decoder.

Dataset	MNIST	CIFAR-10	CelebA
Batch Size	100	100	100
Epochs	100	100	100
Training Examples	60,000	50,000	16,2079
Test Examples	10,000	10,000	20,000
Learning Rate	5×10^{-3}	5×10^{-3}	5×10^{-3}
t	5×10^{-3}	5×10^{-3}	5×10^{-3}
λ	5×10^{-4}	10^{-2}	10^{-2}

Table 7: The hyperparameters for each experiment are elaborated in the following table. The determination of the number of epochs was guided by the aim of attaining a stage of converged reconstruction error.

B Proofs

B.1 Proof of Theorem 1

Theorem 3 (Mismatch controls sampling-time risk on the spherical interface)**.** *Under Assumptions 1–2, the function $r : \mathbb{S}^{d-1} \to \mathbb{R}$ defined in Eq. 22 is $(L_\ell L_f)$-Lipschitz w.r.t. $d_{\mathbb{S}}$. Consequently, for any probability measures μ, ν on $\mathbb{S}^{d-1}$,*

$$\left|\mathcal{R}(\mu) - \mathcal{R}(\nu)\right| \;\le\; L_\ell L_f \, W_1^{\mathbb{S}}(\mu, \nu). \tag{28}$$

In particular, for the LISA directional aggregate $\tilde{q}_\phi$ and the projected ex-post sampler $\tilde{p}_\psi$,

$$\mathcal{R}(\tilde{p}_\psi) \;\le\; \mathcal{R}(\tilde{q}_\phi) \;+\; L_\ell L_f \, W_1^{\mathbb{S}}(\tilde{p}_\psi, \tilde{q}_\phi). \tag{29}$$

Proof. **Step 1: r is Lipschitz on $(\mathbb{S}^{d-1}, d_{\mathbb{S}})$.** Fix $u, v \in \mathbb{S}^{d-1}$. By Assumption 2, for each realization of $X \sim p_{\text{data}}$,

$$\big|\ell(X, f_\theta(u)) - \ell(X, f_\theta(v))\big| \le L_\ell \, \|f_\theta(u) - f_\theta(v)\|_2.$$

Taking expectation over X and using linearity of expectation gives

$$|r(u) - r(v)| = \Big|\mathbb{E}_X\big[\ell(X, f_\theta(u)) - \ell(X, f_\theta(v))\big]\Big| \le \mathbb{E}_X\big|\ell(X, f_\theta(u)) - \ell(X, f_\theta(v))\big| \le L_\ell \, \|f_\theta(u) - f_\theta(v)\|_2.$$

Now apply Assumption 1 to obtain

$$|r(u) - r(v)| \le L_\ell L_f \, d_{\mathbb{S}}(u, v),$$

so r is $(L_\ell L_f)$-Lipschitz on the sphere.

Step 2: Kantorovich–Rubinstein duality on the sphere. By Kantorovich–Rubinstein duality for $W_1^{\mathbb{S}}$,

$$W_1^{\mathbb{S}}(\mu, \nu) = \sup_{\varphi:\, \mathrm{Lip}(\varphi) \le 1} \Big(\mathbb{E}_{U\sim\mu}[\varphi(U)] - \mathbb{E}_{U\sim\nu}[\varphi(U)]\Big),$$

where Lipschitzness is w.r.t. $d_{\mathbb{S}}$. Apply this with $\varphi = r/(L_\ell L_f)$, whose Lipschitz constant is at most 1 by Step 1:

$$\mathcal{R}(\mu) - \mathcal{R}(\nu) = \mathbb{E}_\mu[r] - \mathbb{E}_\nu[r] \le L_\ell L_f \, W_1^{\mathbb{S}}(\mu, \nu).$$

Swapping μ and ν yields the absolute-value form equation 23. Finally, set $\mu = \tilde{p}_\psi$ and $\nu = \tilde{q}_\phi$ to get equation 24. □

B.2 Proof of Theorem 2

Before proving **Theorem** 2, we first introduce the following definitions and lemmas.

Let $\mathbb{S}^{d-1} := \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ and let U denote the uniform (Haar) probability measure on $\mathbb{S}^{d-1}$. We consider the Gaussian (RBF) kernel restricted to the sphere,

$$k(u, v) \;:=\; \exp\Big(-\tfrac{1}{2\sigma^2}\|u - v\|_2^2\Big), \qquad u, v \in \mathbb{S}^{d-1},\ \sigma > 0. \tag{30}$$

Note that for $u, v \in \mathbb{S}^{d-1}$, $\|u - v\|_2^2 = 2 - 2\langle u, v\rangle$, hence

$$k(u, v) \;=\; \exp(-\sigma^{-2}) \, \exp(\langle u, v\rangle/\sigma^2). \tag{31}$$

Lemma 2 (Inner product of two independent uniform directions)**.** *Let $x, z \overset{i.i.d.}{\sim} U$ on $\mathbb{S}^{d-1}$ and set $T = \langle x, z\rangle \in [-1, 1]$. Then T has density*

$$p_d(t) = \frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d-1)/2)}(1 - t^2)^{\frac{d-3}{2}}, \qquad t \in [-1, 1].$$

Proof. Since the joint distribution of x and z is invariant under rotations, the distribution of the scalar $\langle x, z\rangle$ depends only on the angle between x and z. (In a rigorous argument one does not fix z, but by rotational invariance one may show that the marginal distribution of T is identical to that obtained by first rotating the coordinate system so that $z = (1, 0, \ldots, 0)$; the final result is independent of this choice.)

In that coordinate system we have

$$T = \langle x, z\rangle \;= x_1,$$

where $x = (x_1, x_2, \ldots, x_l)$ is uniformly distributed on $\mathbb{S}^{d-1}$. It is a standard fact given by Mardia & Jupp (2009) that the marginal density of the first coordinate x_1 is given by

$$p(x_1) = C_l(1 - x_1^2)^{\frac{l-3}{2}}, \quad x_1 \in [-1, 1],$$

where the normalization constant C_l is determined by

$$\int_{-1}^{1} C_l(1-x_1^2)^{\frac{l-3}{2}}\,dx_1 = 1.$$

Using the standard Beta integral, one can show that

$$C_l = \frac{\Gamma\left(\frac{l}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{l-1}{2}\right)}.$$

Thus, the density of u is

$$p_d(t) = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)}\,(1-t^2)^{\frac{d-3}{2}}, \quad t \in [-1,1].$$

This completes the proof. □

Lemma 3 (Uniform cross-term is constant)**.** *For any fixed $v \in \mathbb{S}^{d-1}$,*

$$\mathbb{E}_{w\sim U}\big[k(w,v)\big] \;=\; \mathbb{E}_{w,w'\sim U}\big[k(w,w')\big] \;=:\; C_{d,\sigma}. \tag{32}$$

Moreover, $C_{d,\sigma}$ admits the closed form

$$C_{d,\sigma} \;=\; \exp(-\sigma^{-2})\,\Gamma\!\left(\frac{d}{2}\right)(2\sigma^2)^{\frac{d-2}{2}}\,I_{\frac{d-2}{2}}\!\left(\sigma^{-2}\right), \tag{33}$$

where $I_\nu(\cdot)$ is the modified Bessel function of the first kind.

Proof. The equality in Eq. 32 follows from rotational invariance of U: for fixed v, the law of $\langle w, v\rangle$ with $w \sim U$ is the same as the law of $\langle w, w'\rangle$ with $(w, w') \sim U \times U$, hence the expectations of any function of the inner product coincide.

For the closed form, use Eq. 31 and Lemma 2:

$$\mathbb{E}_{w,w'\sim U}[k(w,w')] = \exp(-\sigma^{-2})\int_{-1}^{1}\exp(t/\sigma^2)\,p_d(t)\,dt.$$

Applying the standard integral identity

$$\int_{-1}^{1} e^{\alpha t}(1-t^2)^{\nu-\frac{1}{2}}dt = \sqrt{\pi}\,\Gamma\!\left(\nu+\frac{1}{2}\right)\left(\frac{2}{\alpha}\right)^{\nu} I_\nu(\alpha) \qquad (\alpha>0,\ \nu>-\tfrac{1}{2}),$$

with $\nu = \frac{d-2}{2}$ and $\alpha = \sigma^{-2}$, and multiplying by the normalization constant in $p_d(t)$ yields equation 33. □

Theorem 4 (MMD to the uniform sphere reduces to a $Q \times Q$ kernel energy)**.** *Let Q be any probability measure on $\mathbb{S}^{d-1}$ and let $\mathrm{MMD}_k^2(U,Q)$ denote the squared maximum mean discrepancy with kernel k. Then*

$$\mathrm{MMD}_k^2(U,Q) \;=\; \mathbb{E}_{z,z'\sim Q}\big[k(z,z')\big] \;-\; C_{d,\sigma}, \tag{34}$$

where $C_{d,\sigma}$ is the constant in Lemma 3. Furthermore, given i.i.d. samples $z_1,\ldots,z_N \sim Q$, the unbiased estimator satisfies

$$\widehat{\mathrm{MMD}}_k^2(U,Q) \;:=\; \frac{1}{N(N-1)}\sum_{i\neq j} k(z_i,z_j) \;-\; C_{d,\sigma}, \qquad \mathbb{E}\left[\widehat{\mathrm{MMD}}_k^2(U,Q)\right] = \mathrm{MMD}_k^2(U,Q). \tag{35}$$

Proof. By the standard MMD expansion (Gretton et al., 2012),

$$\mathrm{MMD}_k^2(U, Q) = \mathbb{E}_{U\times U}[k] - 2\mathbb{E}_{U\times Q}[k] + \mathbb{E}_{Q\times Q}[k].$$

By Lemma 3, $\mathbb{E}_{U\times U}[k] = \mathbb{E}_{U\times Q}[k] = C_{d,\sigma}$, which gives equation 34. Unbiasedness of equation 35 follows because the U-statistic $\frac{1}{N(N-1)}\sum_{i\neq j} k(z_i, z_j)$ is an unbiased estimator of $\mathbb{E}_{Q\times Q}[k]$. □

C More Qualitative Results

(a) VAE, $\mathcal{N}(\mathbf{0}, \mathbb{I})$ (b) AE (Sphere), GMM (c) AE (Sphere), MVG

(d) AE, GMM (e) AE, MVG (f) WAE, GMM

(g) WAE, $\mathcal{N}(\mathbf{0}, \mathbb{I})$ (h) RAE-L_2, GMM (i) RAE-L_2, MVG

(j) RAE-GP, GMM (k) RAE-GP, MVG (l) IRMAE, GMM

(m) IRMAE, MVG (n) LORAE, GMM (o) LORAE, MVG

(p) Geo-VAE, MVG (q) Geo-VAE, GMM (r) Geo-VAE, Manifold Aware Sampler

(s) LISA, GMM (t) LISA, MVG

Figure 2: Qualitative comparison of generated CelebA samples across models and latent priors.

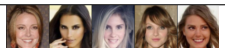
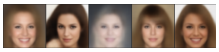
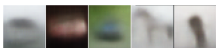
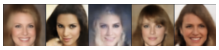
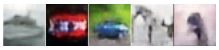
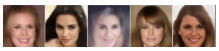
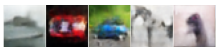
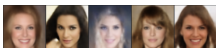
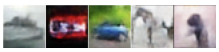
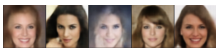
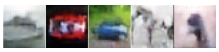
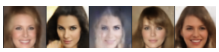
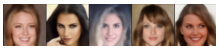
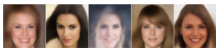
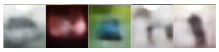
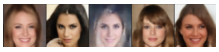
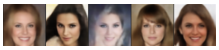
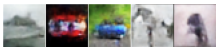
Model	CelebA	MNIST	CIFAR-10
Original			
VAE			
AE			
AE (Sphere)			
WAE			
RAE-L_2			
RAE-GP			
IRMAE			–
Geo-VAE			
LORAE			–
LISA			

Table 8: A qualitative comparison of the reconstruction quality achieved by different autoencoder models across the CelebA, MNIST, and CIFAR-10 datasets.

D Effect of tuning hyperparameters on FID score

In this section, we investigate three crucial hyperparameters: **(i)** penalty parameter (λ), **(ii)** encoder output dimension (latent dimension), and **(iii)** the precision parameter of Gaussian kernel (t). We seek insights into their impact on LISA's generative capacity. All experiments are conducted on the CelebA dataset, where two parameters are fixed while the third is varied. The fixed parameters are set to their default values as specified in **Table** 7.

D.1 Effect of penalty parameter λ on FID score

Elevating the penalty term results in a more pronounced regularization effect, whereas reducing it lessens this impact. We investigate the effect of the regularization penalty parameter on FID scores.

λ	1	0.5	0.1	5×10^{-2}	10^{-2}	5×10^{-3}	10^{-3}	5×10^{-4}
MVG	47.08	43.7	46.93	49.55	45.30	47.79	54.72	55.47
GMM	41.84	38.48	41.54	44.57	39.36	42.75	48.66	49.89

Table 9: Generative performance of LISA on CelebA dataset across different values of λ, where $t = 5 \times 10^{-3}$ and $l = 64$.

D.2 Effect of latent dimension on FID score

Latent Dimension (l)	16	32	64	128	256
MVG	70.55	53.03	45.30	50.19	53.52
GMM	65.23	48.16	39.36	44.73	47.08

Table 10: Generative performance of LISA on CelebA dataset across different values of l, where $t = 5 \times 10^{-3}$ and $\lambda = 10^{-2}$.

D.3 Effect of precision parameter (t) on FID score

t	0.1	5×10^{-2}	10^{-2}	5×10^{-3}	10^{-3}	5×10^{-4}	10^{-4}
MVG	46.89	49.07	49.38	45.30	46.58	46.92	48.54
GMM	41.67	46.55	44.78	39.36	41.11	41.70	43.01

Table 11: Generative performance of LISA on CelebA dataset across different values of t, where $\lambda = 10^{-2}$ and $l = 64$.

E Quantitative Experiments on Reconstruction and Interpolation

Table 12: Comparison of reconstruction FID (rFID) and interpolation FID (iFID) across CelebA, MNIST, and CIFAR-10. Lower values are better. The best results are shown in bold and the second-best results are underlined.

Model	**rFID** (↓)			**iFID** (↓)		
	CelebA	**MNIST**	**CIFAR-10**	**CelebA**	**MNIST**	**CIFAR-10**
AE	31.42	<u>12.54</u>	56.90	41.13	13.70	<u>71.41</u>
AE-Sphere	35.66	14.34	57.69	<u>40.21</u>	17.55	72.34
VAE	59.32	43.55	274.94	74.62	50.96	293.60
WAE	31.61	**12.39**	56.96	41.29	15.71	72.36
RAE-L2	38.46	12.65	<u>56.71</u>	45.42	13.37	71.97
RAE-GP	**30.89**	12.69	57.54	40.73	13.70	72.82
InfoVAE	61.19	30.64	226.31	72.29	34.05	264.26
HVAE	45.63	28.94	245.91	60.02	36.32	277.07
LoRAE	37.17	13.70	–	41.56	13.21	–
IRMAE	106.61	52.33	–	110.40	57.20	–
GMM-VAE	300.44	74.17	–	300.40	76.93	–
GEO-VAE	38.32	12.63	168.34	42.47	<u>12.93</u>	165.97
LISA	<u>31.38</u>	12.69	**51.77**	**37.23**	**12.44**	**63.63**

Discussion: Table 12 compares reconstruction and interpolation quality across CelebA, MNIST, and CIFAR-10. Unlike conventional autoencoders that devote the entire latent capacity to minimizing reconstruction error, LISA explicitly trades reconstruction quality for learning a well-structured spherical latent space

through the proposed uniformity regularizer. As a result, LISA is not optimized solely for reconstruction fidelity. Nevertheless, its reconstruction performance remains highly competitive, achieving the best result on CIFAR-10 and remaining very close to the strongest competing methods on CelebA and MNIST. This demonstrates that enforcing a generation-friendly latent geometry incurs only a fraction of the reconstruction cost.

The benefit of this trade-off is most evident in interpolation quality. LISA consistently achieves the lowest interpolation FID across all three datasets, substantially outperforming both deterministic and variational autoencoder baselines. Since interpolation traverses intermediate regions of the latent space, these results indicate that LISA learns a considerably smoother and more semantically coherent latent manifold. The spherical latent regularization prevents latent representations from collapsing into irregular high-density regions, ensuring that intermediate latent codes remain on the learned data manifold and decode into realistic images. Consequently, LISA not only supports high-quality sampling but also produces significantly more meaningful latent interpolations, highlighting that the proposed regularization simultaneously improves latent continuity and generative robustness while preserving competitive reconstruction performance.

F Experiments on High Resolution Images

In this section, we report results on high-resolution datasets, namely ImageNet-1K and LSUN-Church. To strengthen the benchmark, we additionally include *Automatic Relevancy Determination in Variational Autoencoders* (ARD-VAE) (Saha et al., 2025) as a recent competing method. All models reported in Table 13 are trained directly on the original image resolutions, i.e., 224×224 for ImageNet-1K and 256×256 for LSUN-Church.

Table 13: Comparison on high-resolution datasets using multivariate Gaussian sampling (MVG), reconstruction FID (rFID), Gaussian mixture sampling (GMM), and interpolation FID (iFID). Lower values are better. The best results are shown in bold and the second-best results are underlined. Gain/Loss is computed relative to the strongest competing baseline.

Model	**ImageNet**				**LSUN-Church**			
	MVG (↓)	**rFID** (↓)	**GMM** (↓)	**iFID** (↓)	**MVG** (↓)	**rFID** (↓)	**GMM** (↓)	**iFID** (↓)
AE	182.29	134.66	185.84	157.71	283.00	225.10	277.44	244.04
RAE-GP	186.78	**131.74**	185.52	<u>152.97</u>	276.83	219.24	<u>264.31</u>	245.73
RAE-L2	181.30	134.87	184.86	155.12	<u>268.70</u>	**211.11**	265.52	<u>234.89</u>
AE-Sphere	190.41	135.32	190.75	161.85	274.29	226.09	273.29	236.73
VAE	361.22	365.95	370.64	364.62	446.52	546.86	478.10	424.47
WAE	203.84	363.46	346.25	376.51	402.98	381.12	378.92	555.08
ARD-VAE	<u>180.61</u>	140.79	<u>180.32</u>	154.37	283.04	267.27	279.34	281.77
LISA	**180.18**	<u>133.30</u>	**179.63**	**145.78**	**265.78**	<u>212.02</u>	**263.93**	**234.02**

Discussion: Table 13 presents the comparison on the high-resolution ImageNet-1K and LSUN-Church benchmarks. LISA consistently achieves the best generation quality across both datasets, obtaining the lowest FID for both multivariate Gaussian (MVG) and Gaussian mixture model (GMM) sampling. More importantly, LISA also achieves the best interpolation FID (iFID), indicating that the learned latent manifold is not only generation-friendly but also smooth and semantically consistent between latent codes. These results are a direct consequence of the proposed spherical latent-space regularization. By combining a reconstruction objective with the proposed uniformity regularizer, LISA explicitly encourages the encoder to distribute latent representations uniformly over the hypersphere, producing a latent geometry that is substantially easier to sample from than conventional deterministic autoencoders. Consequently, simple parametric density models such as MVG and GMM provide significantly better approximations to the latent distribution, leading to improved generation quality without requiring complex latent-space priors.

It is also worth noting that LISA is not explicitly optimized for reconstruction alone. The uniformity regularizer introduces a deliberate trade-off between reconstruction fidelity and latent-space regularity. Nevertheless, despite dedicating part of the optimization capacity towards learning a generation-friendly latent space, LISA remains highly competitive in reconstruction FID. On both ImageNet-1K and LSUN, it achieves the second-best reconstruction performance. This demonstrates that the improved generative capability is obtained without sacrificing reconstruction performance, highlighting that LISA effectively balances reconstruction with a latent representation that is both sample-efficient and interpolation-friendly.

G Effect of Number of Gaussian Components

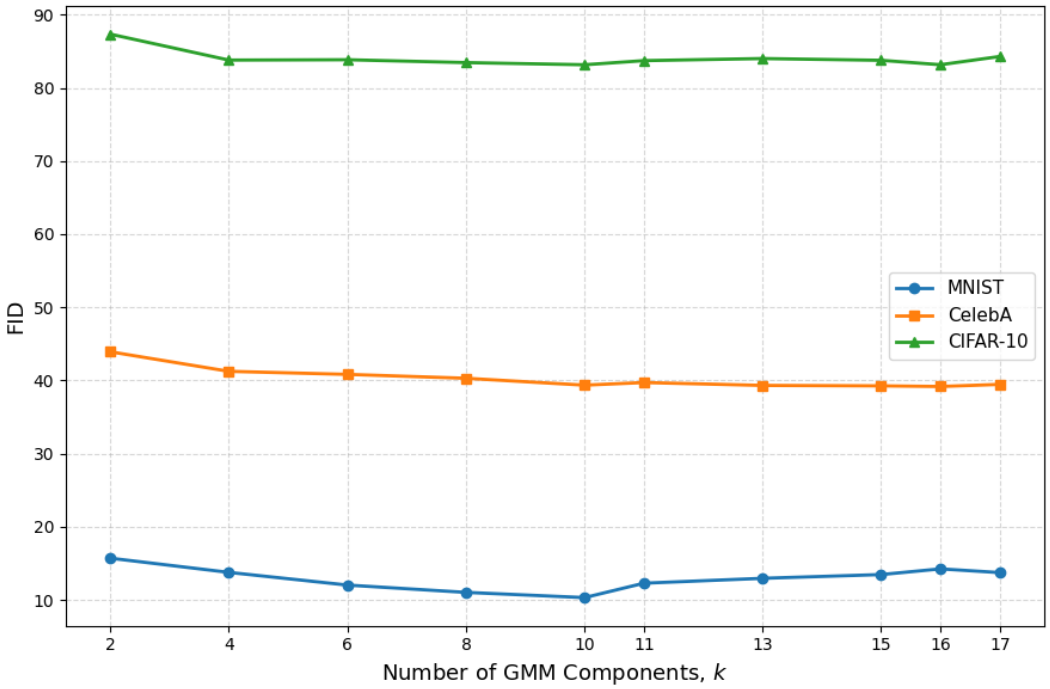


Figure 3: Effect of the number of Gaussian mixture components k on generation quality. FID is reported for MNIST, CelebA, and CIFAR-10. Lower values are better.

To investigate the sensitivity of latent-space sampling to the complexity of the density model, we vary the number of Gaussian components k used by the Gaussian Mixture Model (GMM) while keeping the trained LISA encoder and decoder fixed. Specifically, we evaluate $k \in \{2, 4, 6, 8, 10, 11, 13, 15, 16, 17\}$ and report the resulting FID scores on MNIST, CelebA, and CIFAR-10. The results are summarized in Fig. 3.

The results reveal a consistent trend across all three datasets. Increasing the number of Gaussian components from a very small value $k = 2$ substantially improves generation quality, indicating that a single or very coarse mixture model is insufficient to capture the latent distribution. However, once a moderate number of components is reached (approximately $k = 8$ to $k = 13$), the FID largely saturates, and further increasing the number of mixture components yields only marginal fluctuations. This behaviour is particularly evident on CelebA and CIFAR-10, where the FID remains nearly constant over a wide range of k, while MNIST exhibits a shallow optimum around $k = 10$.

These observations provide empirical evidence that the latent space learned by LISA is well regularized and exhibits a simple, smooth global geometry. Since the proposed spherical uniformity regularizer distributes latent representations approximately uniformly over the hypersphere, the latent density can be accurately approximated using only a relatively small number of Gaussian components. Unlike conventional deterministic autoencoders, whose irregular latent distributions often require increasingly complex density estimators,

LISA produces a latent representation whose sampling quality is largely insensitive to the exact choice of k. Consequently, selecting a moderate number of mixture components is sufficient to obtain near-optimal generation performance while avoiding unnecessary model complexity. This robustness further supports the central objective of LISA: learning a latent space that is inherently generation-friendly rather than relying on increasingly sophisticated sampling models.